Diane Jang, Min Jung Kim, Aashka Popat, Richard Xu

World Happiness Analysis

1. Introduction:

In this study, a predictive model of the state of global happiness in 2019 was constructed to study how statistically significant the six variables (economic production (GDP per capita), social support, life expectancy, freedom to make choices, absence of corruption, and generosity) are to the happiness score of each country.

The World Happiness Report, the data set of our project, is a survey of the state of global happiness. The report evaluates the current state of happiness worldwide, assigning happiness scores to 156 countries, and illustrates how the latest research on happiness elucidates the differences in happiness levels both at an individual and national level. Our research focuses on the World Happiness 2019 data set.

The World Happiness report obtains happiness scores and rankings data from the Gallup World Poll based on respondents' answers to rate their current lives on a scale of 0 to 10, with 10 representing the best possible life and 0 being the worst. These scores are adjusted using Gallup weights to ensure they are representative. All of the predictor variables except for GDP per capita represent national averages for each country.

The method we chose to model the relationship is multiple linear regression, which uses several explanatory variables to predict the response variable's outcome. Given our data has 7 variables and 156 observations, we use this method since multiple linear regression can be used to model the relationship between a response variable, the happiness score, and six predictor variables.

This paper will first describe our data, namely analyzing and interpreting our predictor variables in the context of our model. We will then run a regression to find the full model using all the variables and discuss model candidates that were deemed improvements to the full model and find which candidates were the best. Then there will be an analysis and explanation as to why the "tm4" model we chose is the best possible regression model to describe the data. Finally, we will discuss our findings in terms of the real world and find room for any improvements in the future.

2. Data Description:

Six factors (economic production (GDP per capita), social support, life expectancy, freedom to make choices, absence of corruption, and generosity) are analyzed to determine their contribution to making life evaluations higher in each country. Although the six factors do not affect the total score for each country, the factors do provide insights into why certain countries rank higher than others in happiness ranking.

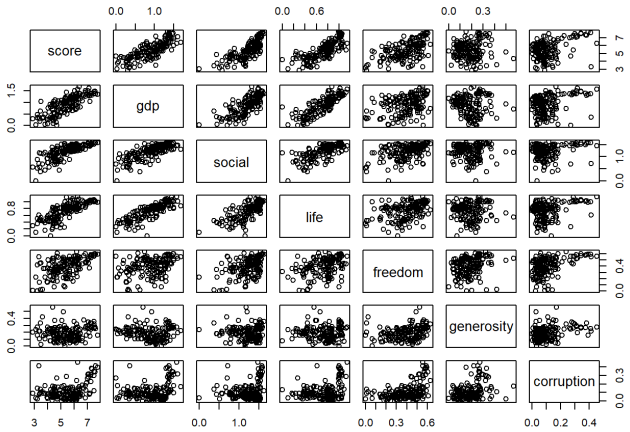
Economic production (GDP per capita) refers to the economic well-being of individuals in a country. Social support measures the level of social support that individuals have access to in their country, including access to friends, family, and community support networks. Life expectancy measures the average number of years people can expect to live in a country. Freedom to make choices measures the degree to which individuals can make choices about their lives without interference from external forces. The absence of corruption measures the extent to which corruption is prevalent in a country. Generosity measures the level of generosity that individuals display towards others.

	Mean	SD	IQR
GDP per capita	0.905	0.398	0.630
Social support	1.209	0.299	0.397
Healthy life expectancy	0.725	0.242	0.334
Freedom to make life choices	0.393	0.143	0.199
Generosity	0.185	0.095	0.140
Perceptions of corruption	0.111	0.094	0.094

Table 1. Mean, Standard Deviation, IQR of Six Explanatory Variables

	Score	GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
Score	1.00000000	0.79388287	0.77705779	0.77988315	0.5667418	0.07582369	0.3856131
GDP per capita	0.79388287	1.00000000	0.75490573	0.83546212	0.3790791	-0.07966231	0.2989198
Social support	0.77705779	0.75490573	1.00000000	0.71900946	0.4473332	-0.04812645	0.1818995
Healthy life expectancy	0.77988315	0.83546212	0.71900946	1.00000000	0.3903948	-0.02951086	0.2952828
Freedom to make life choices	0.56674183	0.37907907	0.44733316	0.39039478	1.0000000	0.26974181	0.4388433
Generosity	0.07582369	-0.07966231	-0.04812645	-0.02951086	0.2697418	1.00000000	0.3265375
Perceptions of corruption	0.38561307	0.29891985	0.18189946	0.29528281	0.4388433	0.32653754	1.0000000

Table 2. Correlation Coefficients between Happiness Score and Explanatory Variables



Graph 1. Correlation Plots between Happiness Score and Explanatory Variables

Table 1 displays the mean, standard deviation, and interquartile range of the six explanatory variables: GDP per capita, social support, life expectancy, freedom, corruption, and generosity. Table 2 displays correlation coefficients between the response variable, the Happiness score, and the explanatory variables. Graph 1 displays correlation plots between the happiness score and the explanatory variables.

From Table 2 and Graph 1, we can see that GDP per capita is most strongly correlated to the happiness score with a correlation coefficient of approximately 0.794, and generosity is most weakly correlated with a correlation coefficient of 0.076. There are also some positive associations among the predictor variables, such as GDP per capita and life expectancy.

3. Results and Interpretation

a. Full Model:

```
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           0.7754     0.2182   3.553 0.000510 ***
social        1.1242     0.2369   4.745 4.83e-06 ***
life          1.0781     0.3345   3.223 0.001560 **
freedom       1.4548     0.3753   3.876 0.000159 ***
generosity    0.4898     0.4977   0.984 0.326709
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
```

$$\hat{\text{score}} = 1.7952 + 0.7754\ \text{GDP} + 1.1242\ \text{social} + 1.0781\ \text{life} + 1.4548\ \text{freedom} + 0.4898\ \text{generosity} + 0.9723\ \text{corruption}$$

Interpretation for full model:

- For every 1 unit GDP per capita increases, Happiness Score increases by 0.7754 units
- For every 1 unit perception of having Social Support increases, the Happiness Score increases by 1.1242 units
- For every 1 question answered with the question being "Are you satisfied with freedom?", Happiness Score increases by 1.4548 units
- For every 1 question answered with the question being "Have you donated to charity in the past month?", Happiness Score increases by 0.4898 units
- For every 1 unit corruption perception goes up, Happiness Score increases by 0.9723 units

The p-value of the full model is less than 2.2e-16. There is sufficient evidence to reject the null hypothesis. Therefore, the model is significant according to the p-value. The R_{adj}^2 is 77.03%, meaning 77.03% of the variability in happiness score is explained by the model. Although the p-value and the R_{adj}^2 both present a valid model, the regression coefficients of the model present that the model is not valid. The regression coefficient of generosity (p-value of 0.327) and corruption (p-value of 0.075) is insignificant. Lastly, our four diagnostic plots can be improved on the residual diagnostic plot since it has a slight positive quadratic trend (see Appendix A, Graph IV).

b. Candidate Model 1: model om5

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.82997 -0.35344  0.05803  0.35977  1.17522

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.8689     0.1973   9.471  < 2e-16 ***
gdp           0.7455     0.2161   3.450 0.000728 ***
social        1.1180     0.2368   4.722 5.33e-06 ***
life          1.0840     0.3344   3.241 0.001467 **
freedom       1.5340     0.3666   4.185 4.84e-05 ***
corruption    1.1176     0.5218   2.142 0.033839 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5335 on 150 degrees of freedom
Multiple R-squared:  0.7777, Adjusted R-squared:  0.7703
F-statistic:   105 on 5 and 150 DF,  p-value: < 2.2e-16
```

```
##   p     Radj2       AIC      AICc       BIC
## 1 1 0.6278490 -118.7760 -118.6201 -112.6763
## 2 2 0.7089994 -156.1643 -155.9028 -147.0147
## 3 3 0.7486966 -178.0669 -177.6721 -165.8675
## 4 4 0.7648643 -187.4703 -186.9140 -172.2210
## 5 5 0.7703197 -190.1688 -189.4222 -171.8697
## 6 6 0.7702711 -189.1793 -188.2129 -167.8303
```

$$\text{scôre} = 1.8689 + 0.7455\ \text{GDP} + 1.1180\ \text{social} + 1.0840\ \text{life} + 1.5340\ \text{freedom} + 1.1176\ \text{corruption}$$

Above is a reduced version of the full model without transformation. This reduced model does not include the variable generosity. This model was found by the variable selection of all possible subsets and comparing the respective values with the goodness of fit testing. A high R_{adj}^2 and lower AIC, AICc, and BIC present the best model. Looking at our values table, R_{adj}^2, AIC, AICc suggests p = 5, while BIC suggests p = 4. Therefore, after conducting the partial F-test of the four-predictor model and the five-predictor model (see Appendix B, Table I), we concluded that the five-predictor model is the best candidate for our consideration. It is easy to interpret, has good diagnostics without transformation, all variables are significant, and the R_{adj}^2, AIC, AICc, and BIC are comparable to our final model.

c. Candidate Model 2: model m4

```
Call:
lm(formula = score ~ gdp + tsocial + tlife + tfreedom + tgenerosity +
    tcorruption)

Residuals:
     Min       1Q   Median       3Q      Max
-1.54776 -0.31762  0.00765  0.33108  1.21433

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.3619     0.2116  11.163  < 2e-16 ***
gdp           0.6114     0.2192   2.790 0.005968 **
tsocial       0.6185     0.1105   5.595 1.03e-07 ***
tlife         1.0401     0.2809   3.703 0.000300 ***
tfreedom      1.5442     0.4026   3.835 0.000185 ***

tgenerosity   0.5385     0.3953   1.362 0.175217
tcorruption   0.4024     0.3651   1.102 0.272216
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.5138 on 149 degrees of freedom
Multiple R-squared:  0.7952, Adjusted R-squared:  0.7869
F-statistic:  96.4 on 6 and 149 DF,  p-value: < 2.2e-16
```

Box Cox Transformation suggests:

$$x1 \rightarrow x1^{1}$$
$$x2 \rightarrow x2^{2}$$
$$X3 \rightarrow x3^{1.48}$$
$$X4 \rightarrow x4^{1.34}$$
$$X5 \rightarrow x5^{0.50}$$
$$X6 \rightarrow x6^{0.33}$$

$$\hat{\text{score}} = 2.3619 + 0.6114\ \text{GDP} + 0.6185\ \text{tsocial} + 1.0401\ \text{tlife} + 1.5442\ \text{tfreedom} + 0.5385\ \text{tgenerosity} + 0.4024\ \text{tcorruption}$$

Above is the fully transformed model resulting from box-coxing (see Appendix B, Table IV) the predictor and response variables simultaneously. The “t” in front of the variable name is an indication of the variable having been transformed in accordance with the rounded powers suggested by the R box-cox output, which are shown above. It has slightly better residual diagnostic plots (see Appendix A, Graph V) than the full model; however, tgenerosity (p-value of 0.175) and tcorruption (p-value of 0.272) are both insignificant, so this model could be improved upon. Another issue is that interpretation is more difficult due to the transformation of variables.

d. Candidate Model 3: model tm5

```
Call:
lm(formula = score ~ gdp + tsocial + tlife + tfreedom)

Residuals:
     Min       1Q   Median       3Q      Max
-1.69191 -0.31267 -0.00718  0.36593  1.26434

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.6972     0.1242  21.721  < 2e-16 ***
gdp           0.5885     0.2160   2.725 0.007197 **
tsocial       0.5916     0.1089   5.430 2.20e-07 ***
tlife         1.0741     0.2819   3.810 0.000202 ***
tfreedom      1.9114     0.3567   5.359 3.08e-07 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.5169 on 151 degrees of freedom
Multiple R-squared:  0.7899, Adjusted R-squared:  0.7843
F-statistic: 141.9 on 4 and 151 DF,  p-value: < 2.2e-16
```

Box Cox Transformation suggests:

$$x1 \rightarrow x1^{1}$$
$$x2 \rightarrow x2^{2}$$
$$X3 \rightarrow x3^{1.48}$$
$$X4 \rightarrow x4^{1.34}$$
$$X5 \rightarrow x5^{0.50}$$
$$X6 \rightarrow x6^{0.33}$$

$$\hat{\text{score}} = 2.6972 + 0.5885\ \text{GDP} + 0.5916\ \text{social}^{2} + 1.0741\ \text{life}^{1.48} + 1.9114\ \text{freedom}^{1.34}$$

Since the previously considered model (Candidate Model 2) still had some insignificant variables, we decided to run variable selection via forward and backward stepwise regression. These stepwise regressions created two plausible models, tm4 and tm5. This is described in the findings below.

e. Best Predictive Model:

```
Call:
lm(formula = score ~ gdp + tsocial + tlife + tfreedom)

Residuals:
     Min       1Q   Median       3Q      Max
-1.69191 -0.31267 -0.00718  0.36593  1.26434

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.6972     0.1242  21.720  < 2e-16 ***
gdp           0.5885     0.2160   2.725 0.007197 **
tsocial       0.5915     0.1089   5.430 2.20e-07 ***
tlife         1.0741     0.2819   3.810 0.000202 ***
tfreedom      1.9114     0.3567   5.359 3.08e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5169 on 151 degrees of freedom
Multiple R-squared:  0.7899, Adjusted R-squared:  0.7843
F-statistic: 141.9 on 4 and 151 DF,  p-value: < 2.2e-16
```

$$\text{scôre} = 2.6972 + 0.5885\ \text{GDP} + 0.5916\ \text{social}^{2} + 1.0741\ \text{life}^{1.48} + 1.9114\ \text{freedom}^{1.34}$$

The main models for consideration were: the best, reduced, original model without transformation "om5", the full transformed model "m4", and the two comparable reduced, transformed models "tm4" and "tm5."

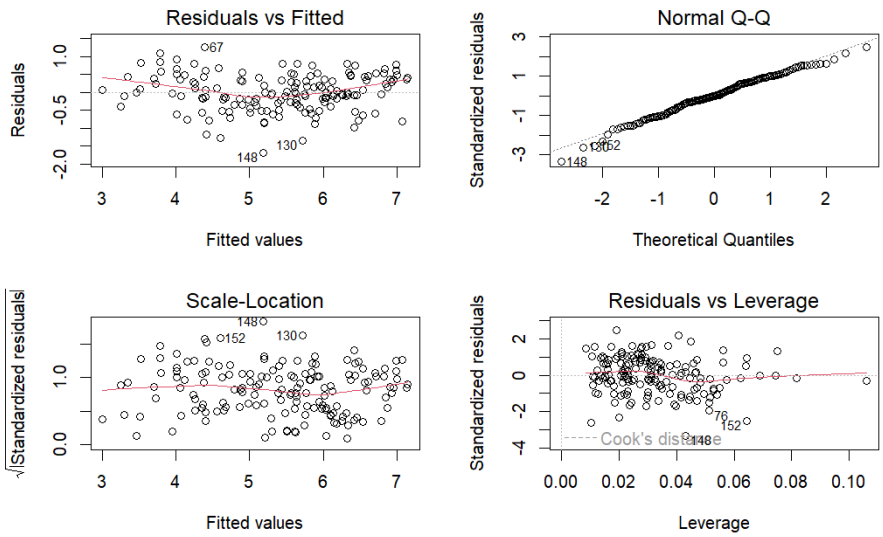
We eventually decided not to consider Candidate Model 1 (om5) as our best model since we felt that despite having solid AIC, BIC, and R_{adj}^2 it could be improved upon on the residual diagnostic plot since it had a slight positive quadratic trend (see Appendix A, Graph VI). We also dropped Candidate Model 2 (m4) because the R summary stated that the variables tgenerosity and tcorruption were insignificant, thus the reason why we considered between the two reduced and transformed models below.

- tm4: *score ~ gdp + tsocial + tlife + tfreedom*
- tm5: *score ~ gdp + tsocial + tlife + tfreedom + tgenerosity*

By checking all added-variable plots (Appendix A, Graph I), exploring all the subsets, and conducting the forward and backward stepwise selection, we concluded that the best model is tm4.

The forward stepwise selection and backward stepwise selection results came out as below:
forwardAIC = *score ~ tsocial + tlife + tfreedom + gdp + tgenerosity*
forwardBIC = *score ~ tsocial + tlife + tfreedom + gdp*
backwardAIC = *score ~ gdp + tsocial + tlife + tfreedom + tgenerosity*
backwardBIC = *score ~ gdp + tsocial + tlife + tfreedom*

While forwardAIC and backward AIC suggest *score ~ gdp + tsocial + tlife + tfreedom + tgenerosity* (tm5), forwardBIC and backwardBIC suggest *score ~ gdp + tsocial + tlife + tfreedom* (tm4). While tm5's tgenerosity variable is insignificant, tm4 has all significant predictor variables. Although tm4 has an R_{adj}^2 of 0.7843 while tm5 has an R_{adj}^2 of 0.7866, the fits are very similar. Additionally, the partial F-test (see Appendix B, Table II) of tm4 as the reduced model and tm5 as the full model had a p-value of 0.1091 which is not statistically significant, meaning that we fail to reject the null that the reduced model is better. Thus, tm4 is the best model.



Using the output above to assess our model, we can see that the residuals of the fitted values look very slightly curved but generally show no pattern, and have a mean of zero. The standardized residual plots for each variable (see Appendix A, Graph II) also show constant variance and no pattern, so we can infer linearity. The normal Q-Q plot follows a linear shape, implying the normality of the error terms. Furthermore, the square root standardized residual plot follows a horizontal line, meaning there is constant variance in the error term. Looking at the Residuals vs Leverage plot, we see that based on Cook's distance, there are no influential points. Finally, the VIF's of all four variables are less than 5 (see Appendix B, Table III), meaning multicollinearity is not an issue in our model. Therefore, the diagnostic output suggests this is a valid model.

4. Discussion

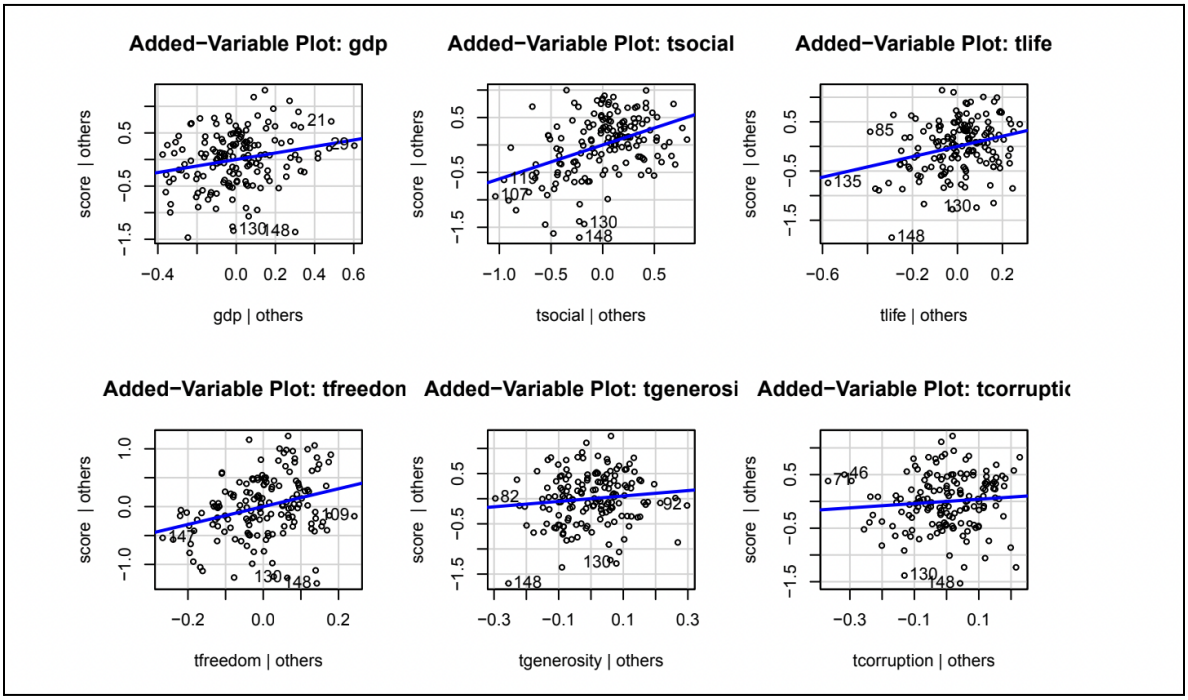
In this work, we took the data from 6 variables and 156 observations to see if we could build a multiple regression model to predict Happiness scores. Through this regression, we could find the Happiness Score of a mock country given its GDP per capita, perception of Social Support, Life Expectancy, and perception of Freedom.

Our final model is hard to understand in the real world as predictor variables that rely on "perception" can be very subjective. In addition, we transformed our model, so interpretation of variables with respect to them being transformed to a lambda power may be difficult. Happiness is also something that cannot usually be quantified, so our model doing so may seem confusing in a real-world context.

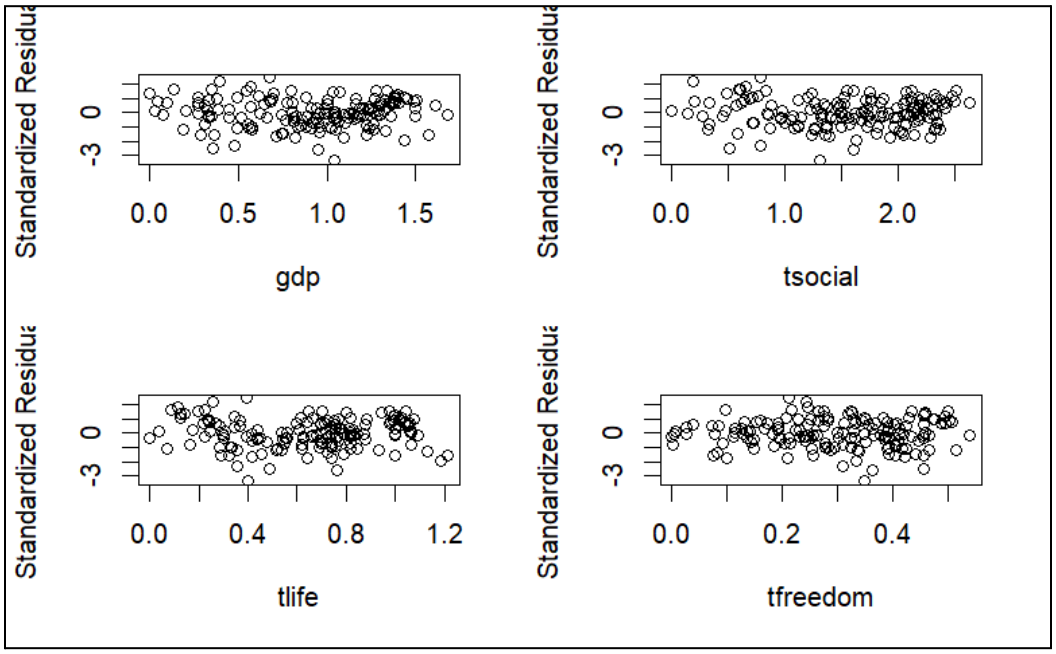
Our analysis was limited by the fact that most of our predictor variable's data were averages of binary pieces of data (ex: Being averages of "0 or 1" for "no or yes" for a question), which made interpretation difficult. If we had each country's raw collected data on these binary variables, we could make a better regression model.

Appendix A: Graphs

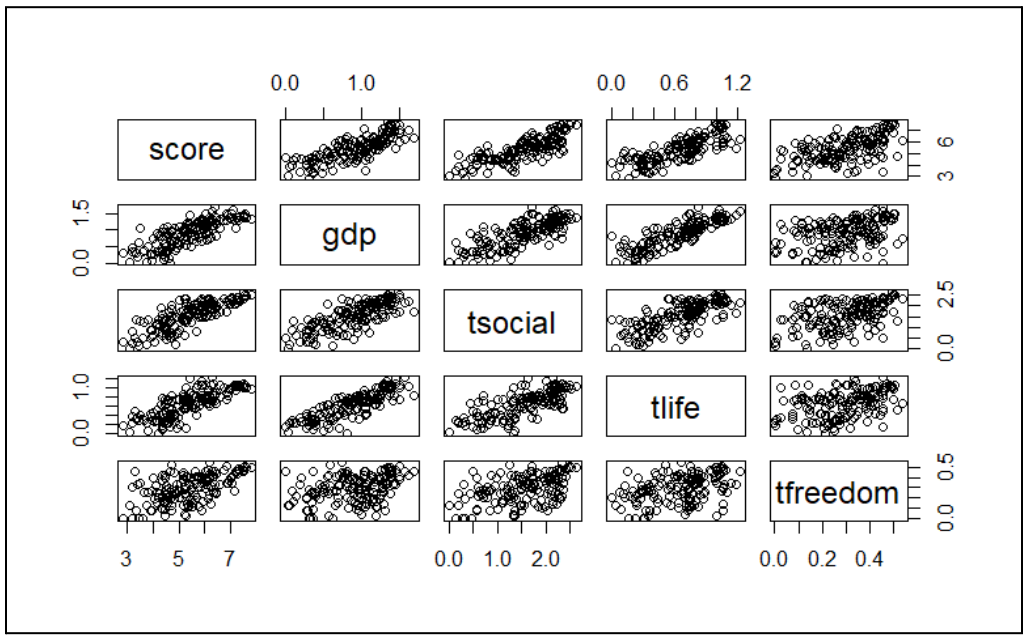
Graph I
Added Variable Plot of tm4



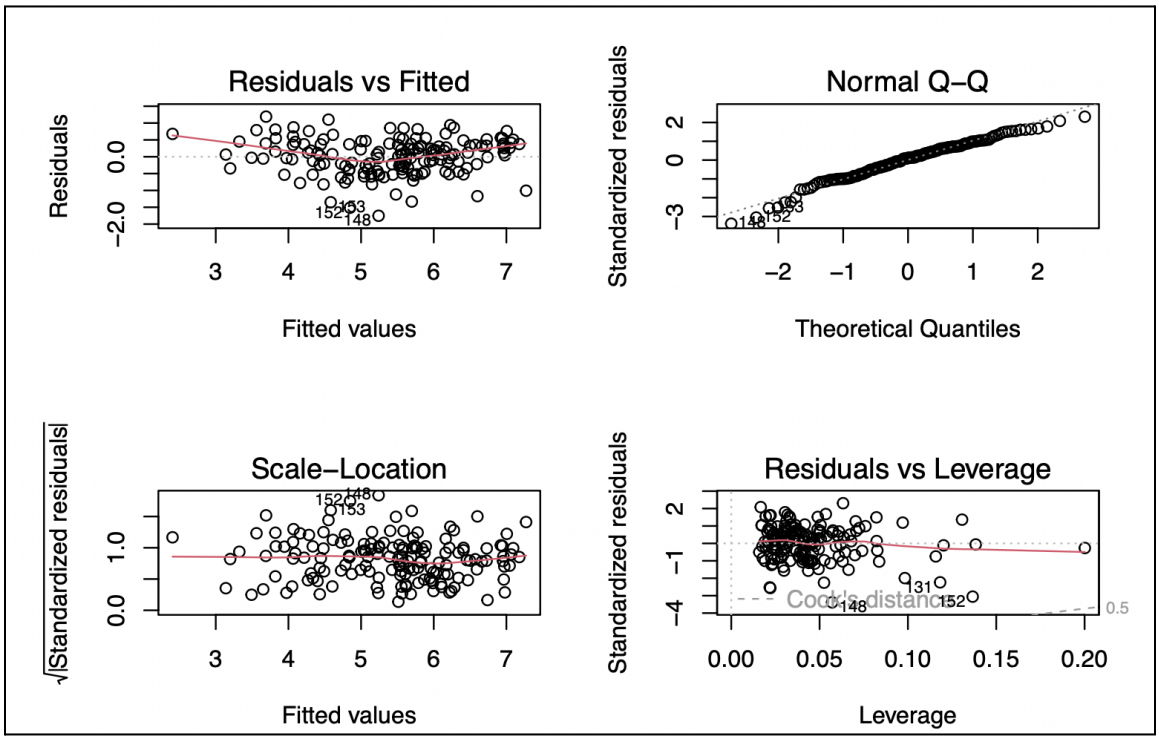
Graph II
Standardized Residuals for tm4



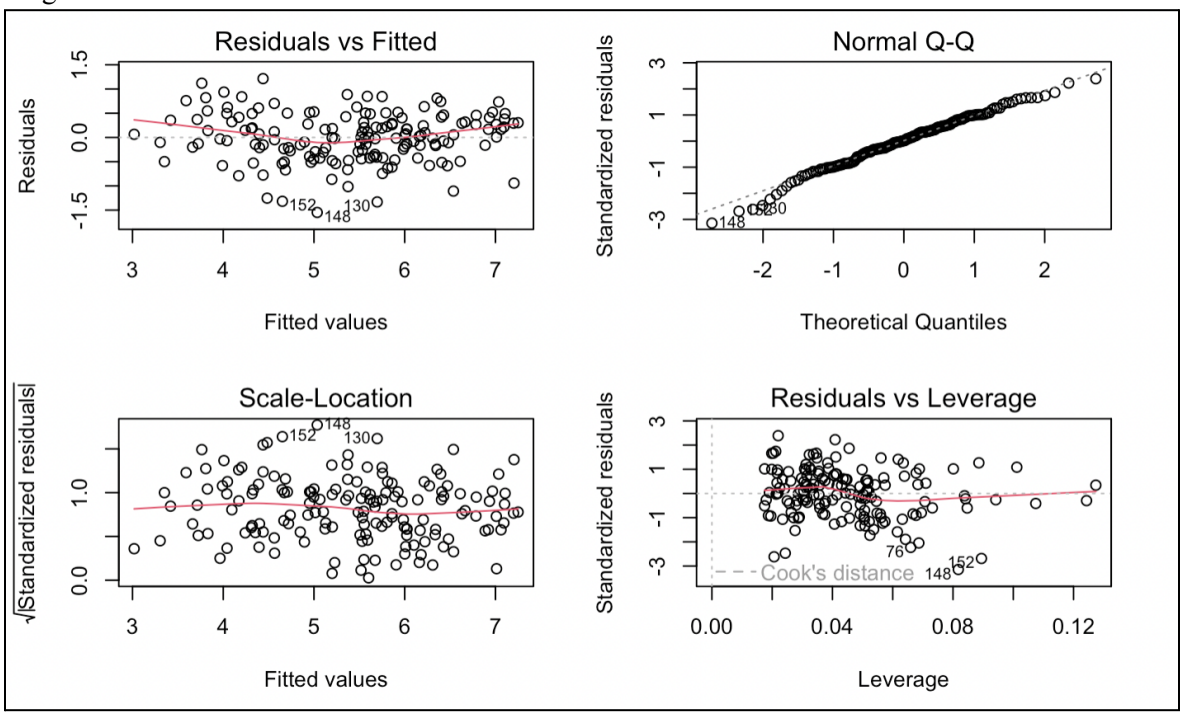
Graph III
Correlation Plot Matrix of tm4



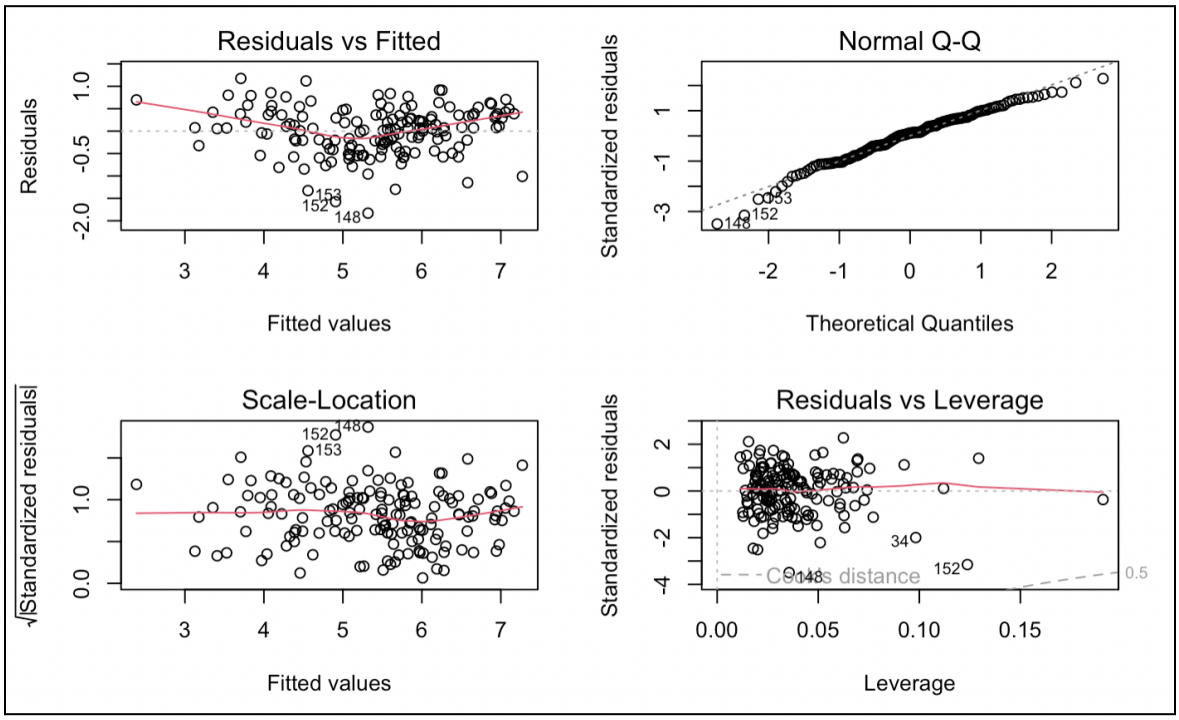
Graph IV
Diagnostic Plot of Full Model



Graph V
Diagnostic Plot of m4



Graph VI
Diagnostic Plot of om5



Appendix B: Tables

Table I
Partial F-Test of om4 vs. om5

```
Analysis of Variance Table

Model 1: score ~ gdp + social + life + freedom
Model 2: score ~ gdp + social + life + freedom + corruption
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    151 43.993
2    150 42.687  1    1.3053 4.5866 0.03384 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table II: Partial F-Test of tm4 vs tm5

```
Analysis of Variance Table

Model 1: score ~ gdp + tsocial + tlife + tfreedom
Model 2: score ~ gdp + tsocial + tlife + tfreedom + tgenerosity
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    151 40.349
2    150 39.662  1    0.6868 2.5975 0.1091
```

Table III
VIF of tm4

VIF(tm4)

gdp	tsocial	tlife	tfreedom
4.294553	2.776192	3.843576	1.271478

Table IV
Boxcox of both X and Y

```
#Boxcox both X and Y:
summary(powerTransform(cbind(score2, gdp2, social2, life2, freedom2, generosity2, corruption2)~1))
```

```
## bcPower Transformations to Multinormality
##             Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## score2         1.0232        1.00       0.5151       1.5314


## gdp2           1.0385        1.00       0.8517       1.2253
## social2        1.8659        2.00       1.4554       2.2765
## life2          1.4833        1.48       1.2142       1.7523
## freedom2       1.3446        1.34       1.0604       1.6289
## generosity2    0.5389        0.50       0.3826       0.6953
## corruption2    0.3613        0.33       0.2530       0.4695
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                          LRT df       pval
## LR test, lambda = (0 0 0 0 0 0 0) 1970.387  7 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                          LRT df       pval
## LR test, lambda = (1 1 1 1 1 1 1) 153.7953  7 < 2.22e-16
```