# World Happiness Report: Stats 101A Final Project

Jeremy Reyes, Kanzah Jamil, Makenzie White, Sofia Jain, Sophia Angelene Santos

## Introduction

This dataset gathers responses from the Gallup World Poll (GWP) for 136 countries taken from 2020-2022. In the survey, each participant was asked to rate their "happiness score" on a scale of 0 to 10 according to Cantril's Ladder, where 0 represents hopelessness and 10 represents prosperity. Along with this, they asked additional questions to gather data on the following variables:

1. Logged GDP per capita = economic output per person (log transformation provides linearity)
2. Social support = National Average of binary responses (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?"
3. Healthy life expectancy = The World Health Organization's measurement of healthy life expectancy for the country.
4. Freedom to Make Life Choices = National averages of responses to the GWP question "Are you satisfied or dissatisfied with your freedom to choose what to do with your life?"
5. Generosity = Residual of regressing national average of the response to the GWP question "Have you donated money to a charity in the past month?" on GDP per capita.
6. Perceptions of Corruption = 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 the average of the two responses. In case the perception of government corruption is missing, we use the perception of business corruption as the overall perception.

Using this data set, we chose the research question: What predictor variables significantly influence the happiness score, Ladder, and to what extent can we quantify their impact? To answer this research question, we went through a process of determining a multiple linear regression model to fit the data and predict the ladder score for a country based on certain variables in the data set. We have 6 numerical and interpretable potential predictors, 1 numerical response variable, and a data set of 136 observations. We want to isolate and analyze the impact of each variable on the overall ladder score of a country, thus a multiple linear regression model is justified.

Before we carried out the process of creating and refining a multiple linear regression model, we scanned the data set and decided to remove observation 99 (Palestine) as it has NA values for healthy life expectancy. This is justified as we still have 136 observations after removing this observation so our model will still be robust enough to answer our research question.

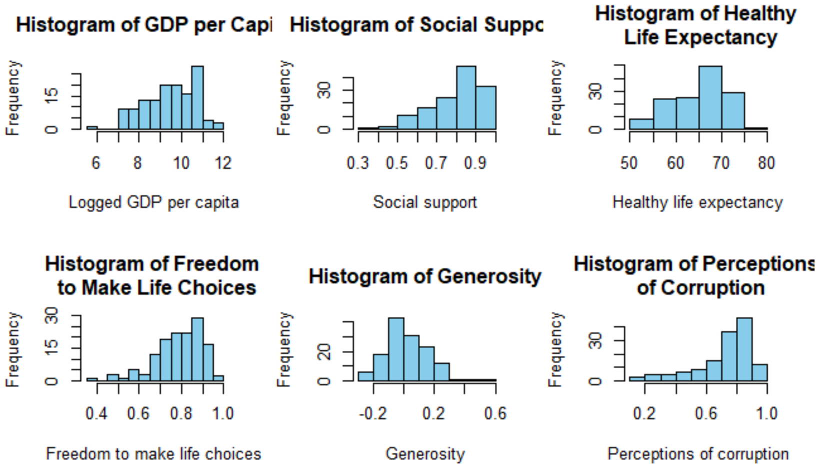
In our paper, we will first describe the summary statistics, relationships, distributions and correlations of the response and predictor variables. Then, we will develop and interpret a full untransformed model, highlighting areas in which we can improve on. Next, we utilize a box-cox transformation on our response variable to find the best λ which makes Ladder Score normal. To determine more model candidates, we find the "best" predictive model using the all possible subset, forward AIC and BIC and backward AIC and BIC approaches. Lastly, we compare our final candidates using an ANOVA test to determine the final model. We analyze the diagnostic plots, leverage points, and multicollinearity of the final model to verify it is appropriate to answer our research question. We conclude with a discussion of our final model with a real world perspective and address limitations and improvements of our analysis.

## Data Description

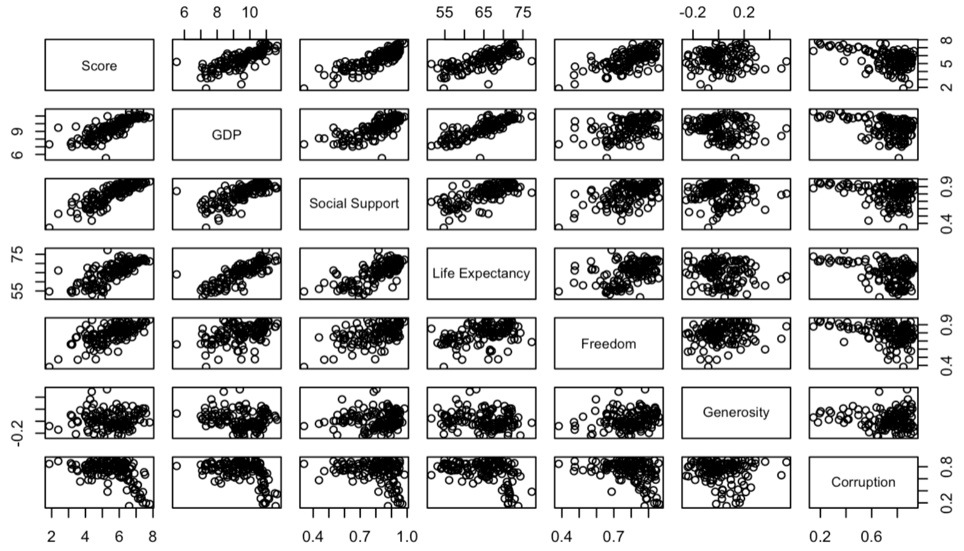
For our response variable (ladder score) and our six predictors in the data set, we calculated the summary statistics and standard deviations summarized in the table below. For anyone curious, according to the data, the country with the highest Ladder score is Finland while the country with the lowest Ladder score is Afghanistan.

| Variable | Standard Deviation | Min | 1st Quartile | Median | Mean | 3rd Quartile | Max |
|---|---|---|---|---|---|---|---|
| Ladder Score | 1.143 | 1.859 | 4.702 | 5.694 | 5.544 | 6.342 | 7.804 |
| Logged GDP Per Capita | 1.21 | 5.527 | 8.587 | 9.574 | 9.455 | 10.541 | 11.660 |
| Social Support | 0.130 | 0.341 | 0.721 | 0.826 | 0.799 | 0.896 | 0.983 |
| Healthy Life Expectancy | 5.750 | 51.530 | 60.649 | 65.837 | 64.968 | 69.412 | 77.280 |
| Freedom to Make Life Choices | 0.112 | 0.382 | 0.726 | 0.801 | 0.788 | 0.875 | 0.961 |
| Generosity | 0.141 | -0.254 | 0.071 | 0.002 | 0.024 | 0.118 | 0.531 |
| Perceptions of Corruption | 0.177 | 0.146 | 0.666 | 0.772 | 0.725 | 0.846 | 0.929 |

Additionally, to find the distributions of each variable, we plotted histograms for each of the predictors. Social Support, Perceptions of Corruption, and Freedom to Make Life Choices have a noticeable negative skew. Logged GDP per capita and Healthy Life Expectancy have a less extreme negative skew. Lastly, Generosity has a positive skewed distribution.



Next, we analyze the relationships between all of the variables using a scatter plot and correlation  matrix to analyze the correlation and relationships among the variables.



We can see there is generally a positive linear relationship between Ladder Score and Logged GDP per Capita, Social Support, Healthy Life Expectancy, Freedom to Make Life Choices, and Corruption. Ladder Score and Perception of corruption have a negative linear relationship. Ladder Score and Generosity do not appear to have a clear linear relationship. Detailed plots of each predictors and ladder score can be found in the appendix (Figure 1). Based on the correlation matrix, there are some high correlations between variables such as GDP and Life Expectancy, GPD and Social Support and Life Expectancy and Social Support that will be addressed in the following sections.

## Results and Interpretation

Our full multiple linear regression model to predict a country's ladder score before transformation is:

> Ladder Score = -2.080 + 0.203(Logged GDP per capita) + 3.926(Social Support) + 0.020(Healthy Life Expectancy) + 2.339(Freedom to Make Life Choices) + 0.143(Generosity) - 0.798(Perceptions of Corruption)

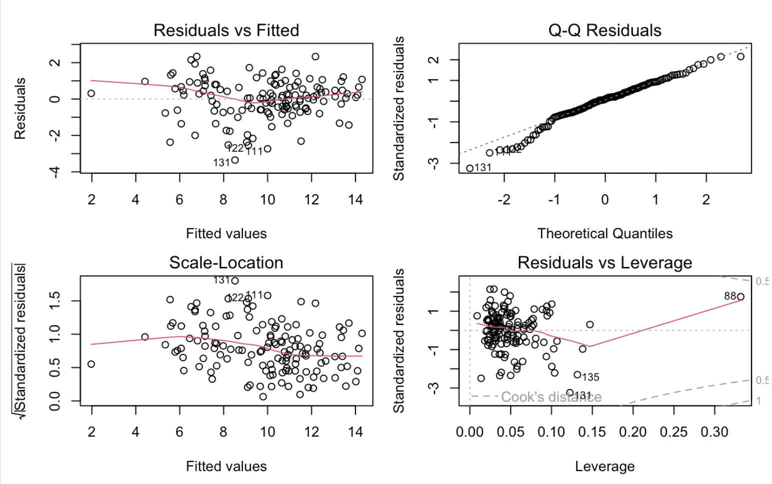
As seen from the R result, the estimates for Healthy Life Expectancy and Generosity are significant, $R^2$ is 0.8281 Based on the F-statistic, the p-value is less than 0.05 so we conclude the model is appropriate but could be improved to have more significant predictors.

To determine the best transformation, we used the Box-Cox method to transform our response variable, Ladder Score, and based on the method of the sum of squares, our best model which minimizes the RSS is when $\lambda = 1.33$ for Ladder Score (Figure 2). Due to the negative values of the Generosity predictor and to avoid getting rid of this predictor, we decided not to transform any of the predictors. Therefore, our full transformed model is:

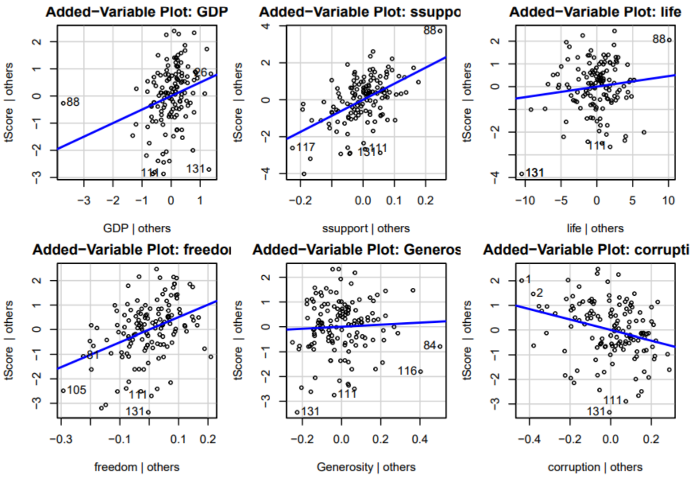
> Ladder Score^(1.33) = -7.286 + 0.499(Logged GDP per capita) + 8.698(Social Support) + 0.046(Healthy Life Expectancy) + 5.075(Freedom to Make Life Choices) +0.413(Generosity) - 2.121(Perceptions of Corruption)

We can see from the R result (Figure 3), once again, that the predictors besides Healthy Life Expectancy and Generosity are significant, the $R^2$ is 0.8308, and based on the F-statistic, the p-value is less than 0.05 so we conclude this candidate model is also appropriate.

Furthermore, the diagnostic plots and standard residual plots (Figure 4) indicate constant variance, normality of the errors with slight tailing, and some outliers and leverage points.



To further improve the transformed model, we analyzed the added variable plots and VIFs to assess covariance and multicollinearity. Based on the added variable plots, Generosity seems to have the least effect on Ladder Score, having adjusted for effects of other predictors, leading us to believe another model candidate is one with it removed. Furthermore, all predictors had a VIF under 5, but GDP's VIF of 4.27 indicates there could be improvement.



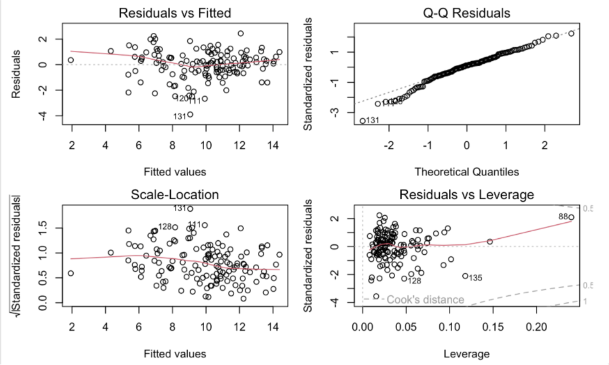
Next, we used our transformed model and the all possible subset approach to find other potential model candidates. Based on the result (Figure 5), $R_{adj}^2$ was maximized in the 5 predictor model (Generosity removed). AIC was also minimized in the 5 predictor model. AIC corrected was minimized in the size 4 predictor model (Life Expectancy and Generosity removed) . Lastly, BIC was minimized in the size 2 model. We decided to further examine the 5 and 4 predictor models, which had the following equations and R outputs (Figures 6 and 7):

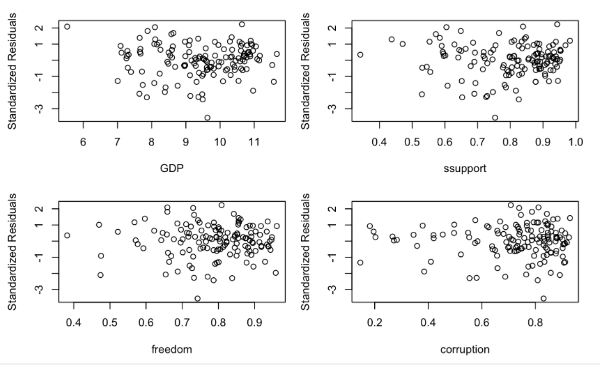
Ladder Score^(1.33) = -7.098 + 0.477(Logged GDP per capita) + 8.853(Social Support) + 0.045(Healthy Life Expectancy) + 5.159(Freedom to Make Life Choices) - 2.195(Perceptions of Corruption)

Ladder Score^(1.33) = -5.732 + 0.611(Logged GDP per capita) + 9.362(Social Support) + 5.075(Freedom to Make Life Choices) - 2.300(Perceptions of Corruption)

To determine the "best" predictive model, we analyzed the results from the Backward and Forward Step Regression approach for AIC and BIC (Figures 8-11). Based on these results, Backward and Forward AIC was minimized in the 5 predictor model and Backward and Forward BIC was minimized in the 4 predictor model. To do a direct comparison, we did an ANOVA test with model 1 being the 4 predictor model and model 2 being the 5 predictor model (Figure 12). The p-value was greater than 0.05 so we conclude the 4 predictor model is more appropriate. Furthermore, Model 4 shows to be a strong contender according to the forward and backward stepwise and all possible subset methods. All of its 4 predictors are significant while model 5 has one insignificant estimate. We determine that the model with the transformed ladder score and the predictors Logged GDP per capita, Social Support, Freedom to Make Life Choices and Perceptions of Corruption is the best predictive model.

To assess the validity of our final model, we evaluated the diagnostic plots. Based on the result, we see constant variance centered around 0 in the residual and standardized residual plots Despite a slight heavy left tail, there is sufficient normality of the errors. Furthermore, all 4 predictors had a VIF well under 5, no unexpected signs of the predictor coefficients, and minimal correlations between variables based on the scatter and correlation matrices (Figures 13 - 15) indicating no multicollinearity.





There are fewer outliers and high leverage points than the initial full non-transformed model; but to further address them, we plotted the Cook's distances and leverages (Figure 16). There were 10 outliers, 3 leverage points and 26 observations with high Cook’s distances. Venezuela and Lebanon were the only “bad” leverage points which could be due to the political and humanitarian crises in these countries during the time the data was gathered, possibly leading to higher than normal perceptions of corruption, and lower freedom to make life choices, logged GDP per capita and social support.

## Discussion

Our final model is:

Ladder Score^(1.33) =  -5.732 + 0.611(Logged GDP per capita) + 9.362(Social Support) + 5.075(Freedom to Make Life Choices) - 2.300(Perceptions of Corruption)

We can interpret the predictor variables as follows:

For every one unit increase in Logged GDP per capita, a country’s Ladder Score increases by a factor of $\log_{1.33}(0.611)$.

For every one unit increase in Social Support, a country’s Ladder Score increases by a factor of $\log_{1.33}(9.362)$.

For every one unit increase in Freedom to Make Life Choices, a country’s Ladder Score increases by a factor of $\log_{1.33}(5.075)$.

For every one unit increase in Perceptions of Corruption, a country’s Ladder Score decreases by a factor of $\log_{1.33}(2.3)$.

Our model makes sense in a real world situation as previous research has found income and societal factors such as social support, freedom to make life choices and perceptions of corruption have a strong relationship with assessing quality of life (Nilsson et al, 2024).

As previously stated, there were many negative values gathered for the Generosity variable, which kept us from attempting a box-cox transformation on the predictors. This is because we cannot simply remove a variable, transform the model, and add the variable back in. In addition, we realize that the survey method of collecting data may not always be reliable due to self-reported information. There could also be other confounding variables that could affect the score that aren’t in the data such as education level, employment status, marital status, religious beliefs etc.

In the future, we would like to find a way around the negative values in the data so that we can attempt to transform the predictors in our models. We could also draw some further conclusions about the state of certain countries according to the labels from Cantril’s Ladder. Upon further research, we notice that Ladder Scores of 0-3 indicate suffering, 3-6 indicate struggling, and 6-10 indicate thriving (Nilsson et al, 2024)

**Citations**

Helliwell, J. F., Layard, R., Sachs, J. D., Aknin, L. B., De Neve, J.-E., & Wang, S. (Eds.). (2023). World Happiness Report 2023 (11th ed.). Sustainable Development Solutions Network.

Nilsson, A.H., Eichstaedt, J.C., Lomas, T. *et al.* The Cantril Ladder elicits thoughts about power and wealth. *Sci Rep* 14, 2642 (2024). https://doi.org/10.1038/s41598-024-52939-y

## Appendix

Figure 1: Detailed plots of each predictors and ladder score

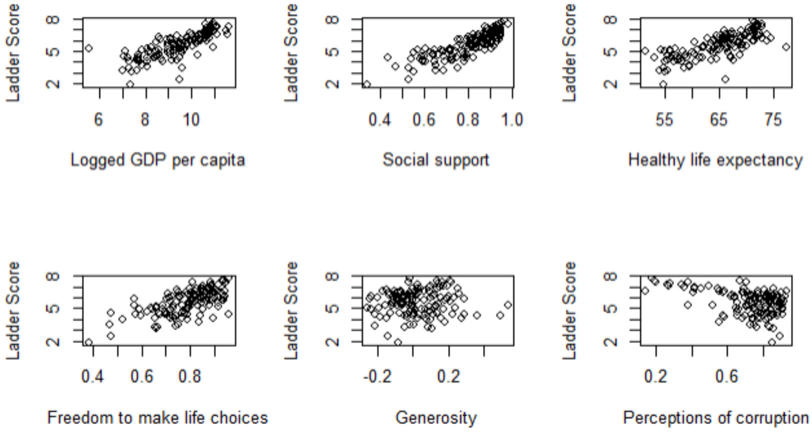
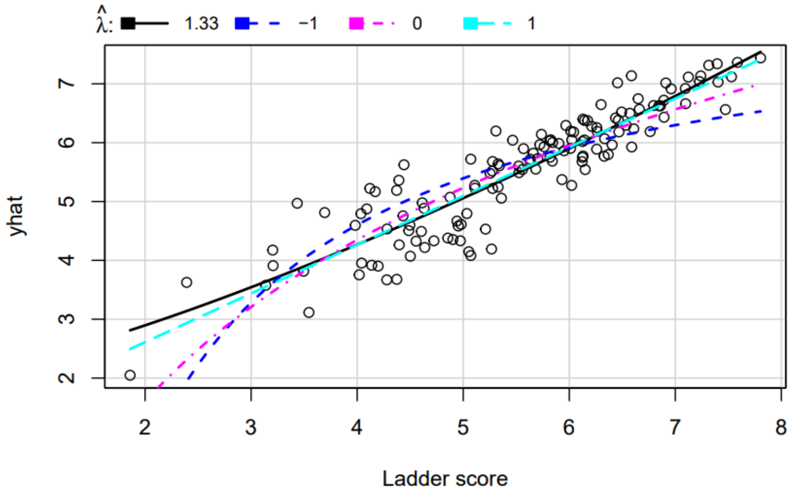
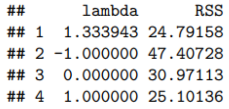


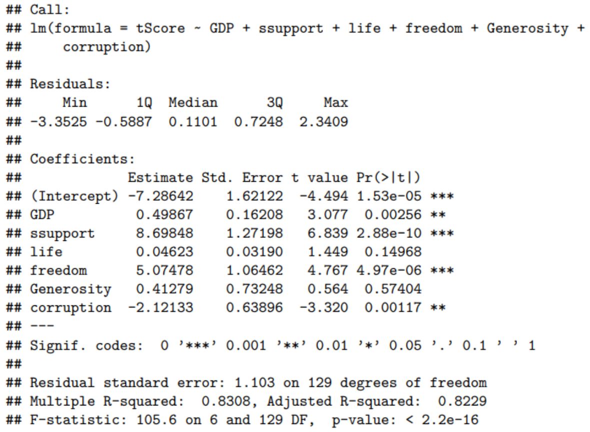
Figure 2: Inverse response plot:





```
##       lambda      RSS
## 1  1.333943 24.79158
## 2 -1.000000 47.40728
## 3  0.000000 30.97113
## 4  1.000000 25.10136
```

Figure 3: Summary of Transformed model



```
## Call:
## lm(formula = tScore ~ GDP + ssupport + life + freedom + Generosity +
##     corruption)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3525 -0.5887  0.1101  0.7248  2.3409
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.28642    1.62122  -4.494 1.53e-05 ***
## GDP          0.49867    0.16208   3.077  0.00256 **
## ssupport     8.69848    1.27198   6.839 2.88e-10 ***
## life         0.04623    0.03190   1.449  0.14968
## freedom      5.07478    1.06462   4.767 4.97e-06 ***
## Generosity   0.41279    0.73248   0.564  0.57404
## corruption  -2.12133    0.63896  -3.320  0.00117 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.103 on 129 degrees of freedom
## Multiple R-squared:  0.8308, Adjusted R-squared:  0.8229
## F-statistic: 105.6 on 6 and 129 DF,  p-value: < 2.2e-16
```

Figure 4: Standard residual plots

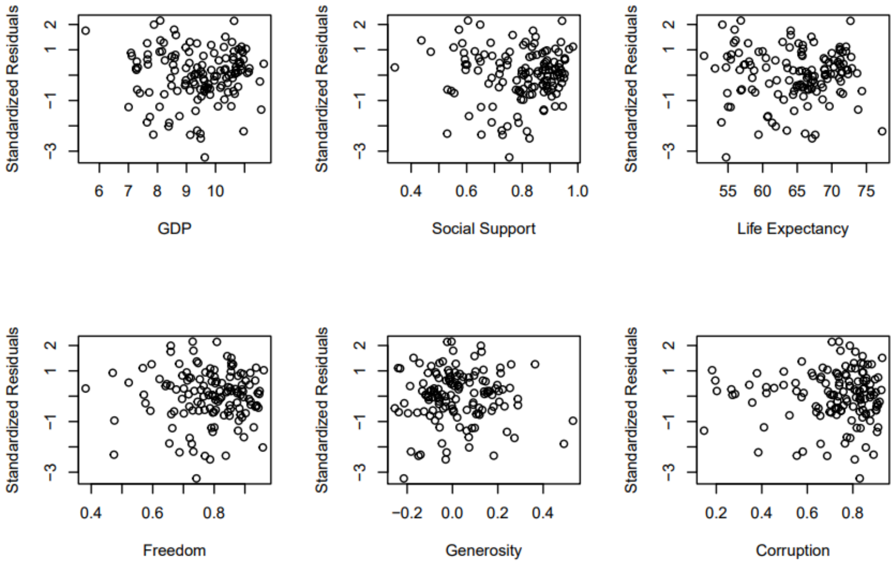
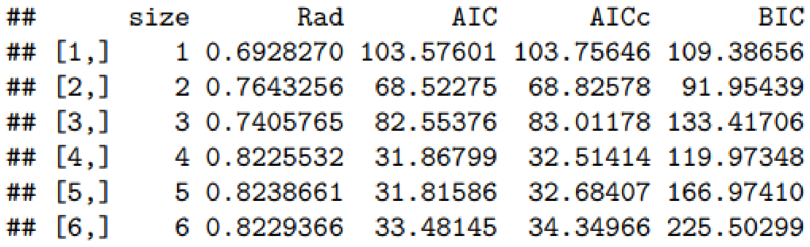
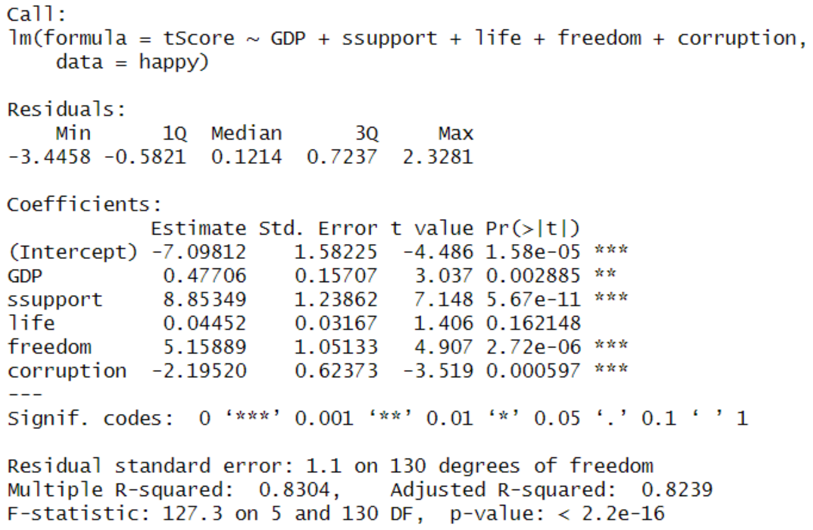


Figure 5: Results from the all possible subset approach:



```
##      size       Rad       AIC      AICc       BIC
## [1,]    1 0.6928270 103.57601 103.75646 109.38656
## [2,]    2 0.7643256  68.52275  68.82578  91.95439
## [3,]    3 0.7405765  82.55376  83.01178 133.41706
## [4,]    4 0.8225532  31.86799  32.51414 119.97348
## [5,]    5 0.8238661  31.81586  32.68407 166.97410
## [6,]    6 0.8229366  33.48145  34.34966 225.50299
```

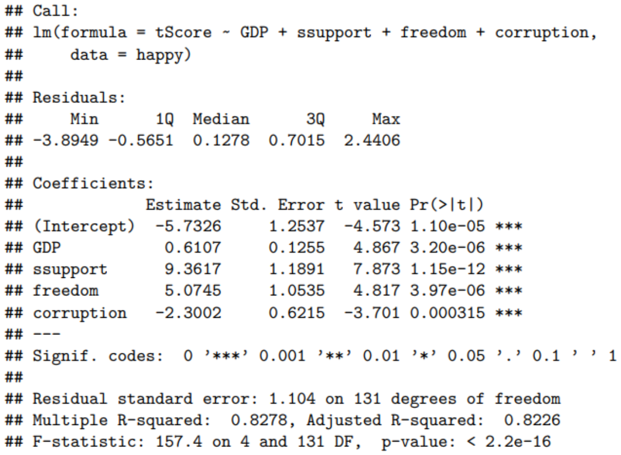
Figure 6: Summary of Model 4



```
Call:
lm(formula = tScore ~ GDP + ssupport + life + freedom + corruption,
    data = happy)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4458 -0.5821  0.1214  0.7237  2.3281 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -7.09812    1.58225  -4.486 1.58e-05 ***
GDP          0.47706    0.15707   3.037 0.002885 ** 
ssupport     8.85349    1.23862   7.148 5.67e-11 ***
life         0.04452    0.03167   1.406 0.162148    
freedom      5.15889    1.05133   4.907 2.72e-06 ***
corruption  -2.19520    0.62373  -3.519 0.000597 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.1 on 130 degrees of freedom
Multiple R-squared:  0.8304,	Adjusted R-squared:  0.8239 
F-statistic: 127.3 on 5 and 130 DF,  p-value: < 2.2e-16
```

\

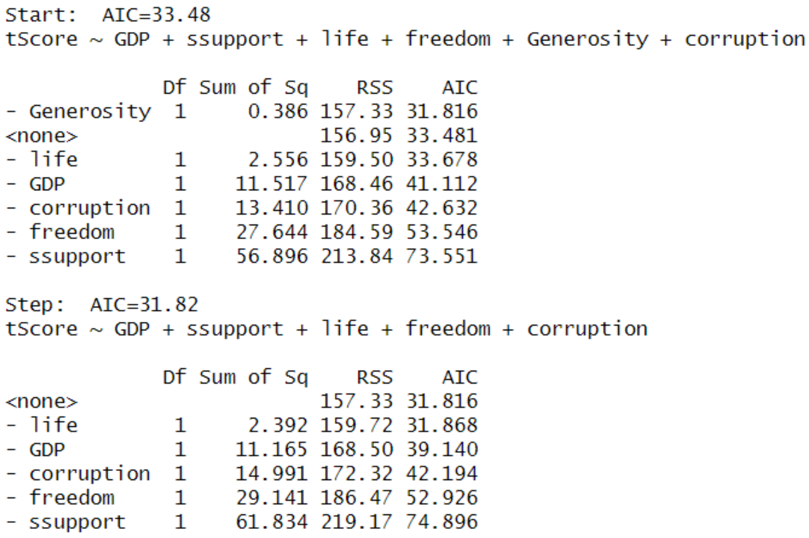
Figure 7: Summary of Model 5



```
## Call:
## lm(formula = tScore ~ GDP + ssupport + freedom + corruption,
##     data = happy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8949 -0.5651  0.1278  0.7015  2.4406
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5.7326     1.2537  -4.573 1.10e-05 ***
## GDP           0.6107     0.1255   4.867 3.20e-06 ***
## ssupport      9.3617     1.1891   7.873 1.15e-12 ***
## freedom       5.0745     1.0535   4.817 3.97e-06 ***
## corruption   -2.3002     0.6215  -3.701 0.000315 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.104 on 131 degrees of freedom
## Multiple R-squared:  0.8278, Adjusted R-squared:  0.8226
## F-statistic: 157.4 on 4 and 131 DF,  p-value: < 2.2e-16
```

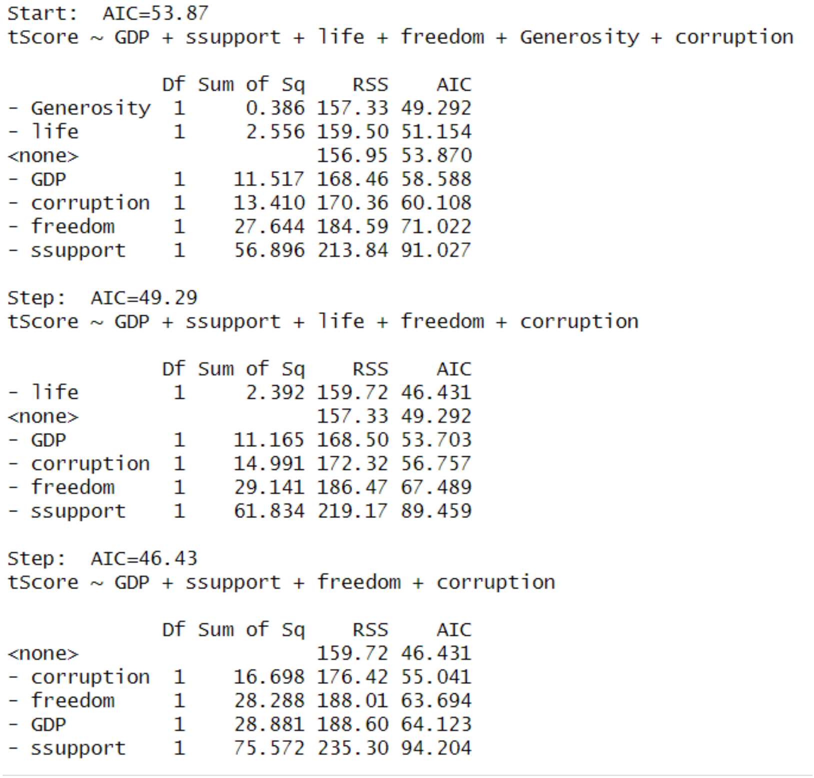
Backward and Forward Step Regression approach for AIC and BIC:

Figure 8: Backward AIC



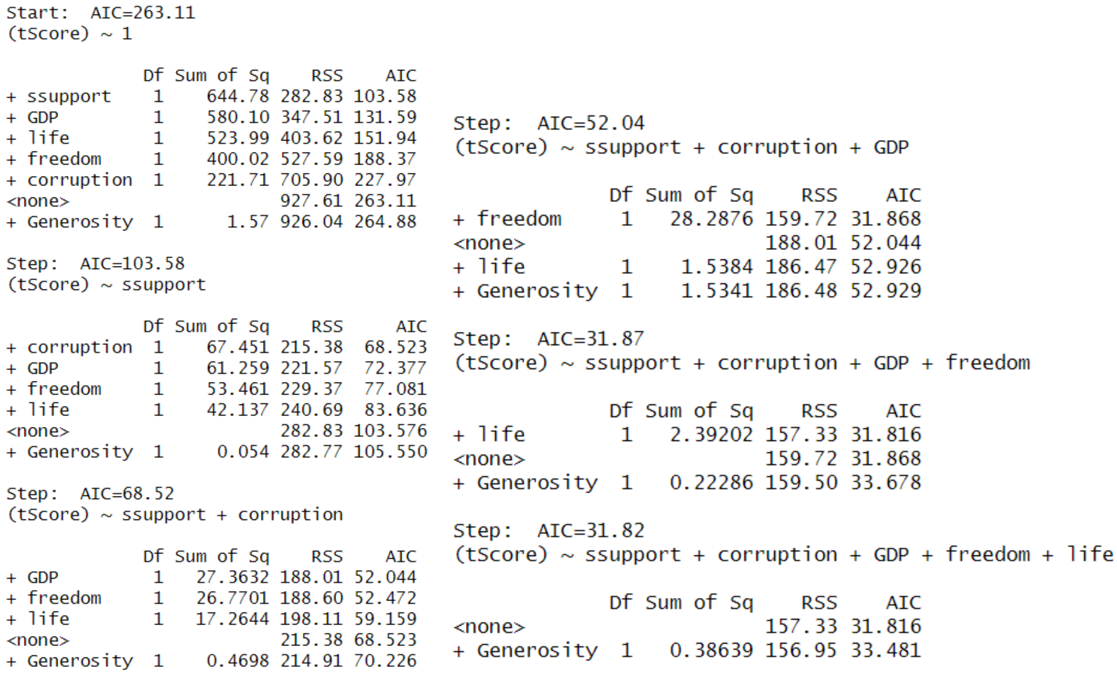
```
Start:  AIC=33.48
tScore ~ GDP + ssupport + life + freedom + Generosity + corruption

             Df Sum of Sq    RSS    AIC
- Generosity  1     0.386 157.33 31.816
<none>                    156.95 33.481
- life        1     2.556 159.50 33.678
- GDP         1    11.517 168.46 41.112
- corruption  1    13.410 170.36 42.632
- freedom     1    27.644 184.59 53.546
- ssupport    1    56.896 213.84 73.551

Step:  AIC=31.82
tScore ~ GDP + ssupport + life + freedom + corruption

             Df Sum of Sq    RSS    AIC
<none>                    157.33 31.816
- life        1     2.392 159.72 31.868
- GDP         1    11.165 168.50 39.140
- corruption  1    14.991 172.32 42.194
- freedom     1    29.141 186.47 52.926
- ssupport    1    61.834 219.17 74.896
```

Figure 9: Backward BIC



```
Start:  AIC=53.87
tScore ~ GDP + ssupport + life + freedom + Generosity + corruption

             Df Sum of Sq    RSS    AIC
- Generosity  1     0.386 157.33 49.292
- life        1     2.556 159.50 51.154
<none>                    156.95 53.870
- GDP         1    11.517 168.46 58.588
- corruption  1    13.410 170.36 60.108
- freedom     1    27.644 184.59 71.022
- ssupport    1    56.896 213.84 91.027

Step:  AIC=49.29
tScore ~ GDP + ssupport + life + freedom + corruption

             Df Sum of Sq    RSS    AIC
- life        1     2.392 159.72 46.431
<none>                    157.33 49.292
- GDP         1    11.165 168.50 53.703
- corruption  1    14.991 172.32 56.757
- freedom     1    29.141 186.47 67.489
- ssupport    1    61.834 219.17 89.459

Step:  AIC=46.43
tScore ~ GDP + ssupport + freedom + corruption

             Df Sum of Sq    RSS    AIC
<none>                    159.72 46.431
- corruption  1    16.698 176.42 55.041
- freedom     1    28.288 188.01 63.694
- GDP         1    28.881 188.60 64.123
- ssupport    1    75.572 235.30 94.204
```

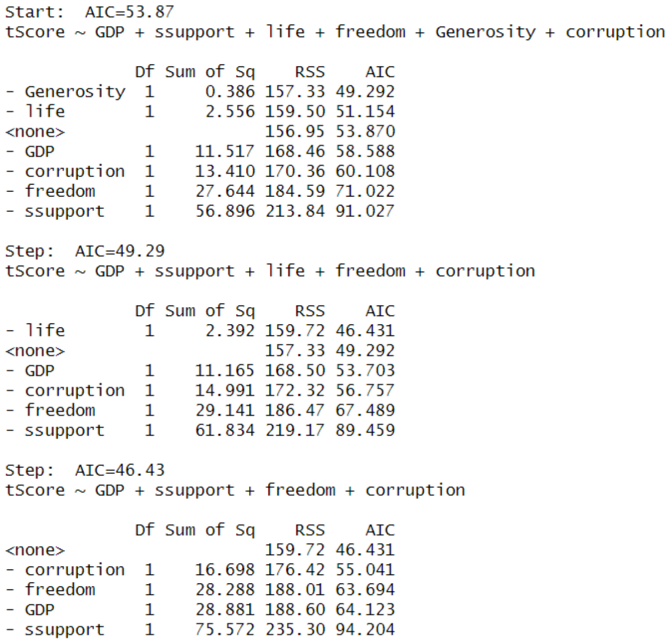
Figure 10: Forward AIC



```
Start:  AIC=263.11
(tScore) ~ 1

             Df Sum of Sq    RSS    AIC
+ ssupport    1    644.78 282.83 103.58
+ GDP         1    580.10 347.51 131.59
+ life        1    523.99 403.62 151.94
+ freedom     1    400.02 527.59 188.37
+ corruption  1    221.71 705.90 227.97
<none>                    927.61 263.11
+ Generosity  1      1.57 926.04 264.88

Step:  AIC=103.58
(tScore) ~ ssupport

             Df Sum of Sq    RSS     AIC
+ corruption  1    67.451 215.38  68.523
+ GDP         1    61.259 221.57  72.377
+ freedom     1    53.461 229.37  77.081
+ life        1    42.137 240.69  83.636
<none>                    282.83 103.576
+ Generosity  1     0.054 282.77 105.550

Step:  AIC=68.52
(tScore) ~ ssupport + corruption

             Df Sum of Sq    RSS    AIC
+ GDP         1   27.3632 188.01 52.044
+ freedom     1   26.7701 188.60 52.472
+ life        1   17.2644 198.11 59.159
<none>                    215.38 68.523
+ Generosity  1    0.4698 214.91 70.226
```

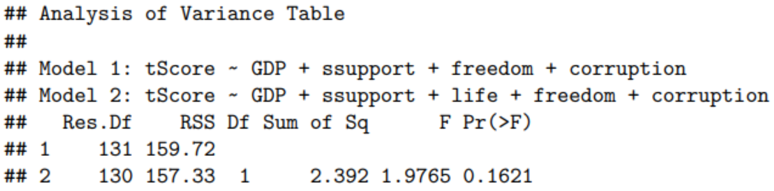
```
Step:  AIC=52.04
(tScore) ~ ssupport + corruption + GDP

             Df Sum of Sq    RSS    AIC
+ freedom     1   28.2876 159.72 31.868
<none>                    188.01 52.044
+ life        1    1.5384 186.47 52.926
+ Generosity  1    1.5341 186.48 52.929

Step:  AIC=31.87
(tScore) ~ ssupport + corruption + GDP + freedom

             Df Sum of Sq    RSS    AIC
+ life        1   2.39202 157.33 31.816
<none>                    159.72 31.868
+ Generosity  1   0.22286 159.50 33.678

Step:  AIC=31.82
(tScore) ~ ssupport + corruption + GDP + freedom + life

             Df Sum of Sq    RSS    AIC
<none>                    157.33 31.816
+ Generosity  1   0.38639 156.95 33.481
```

Figure 11: Forward BIC:



```
Start:  AIC=53.87
tScore ~ GDP + ssupport + life + freedom + Generosity + corruption

             Df Sum of Sq    RSS    AIC
- Generosity  1     0.386 157.33 49.292
- life        1     2.556 159.50 51.154
<none>                    156.95 53.870
- GDP         1    11.517 168.46 58.588
- corruption  1    13.410 170.36 60.108
- freedom     1    27.644 184.59 71.022
- ssupport    1    56.896 213.84 91.027

Step:  AIC=49.29
tScore ~ GDP + ssupport + life + freedom + corruption

             Df Sum of Sq    RSS    AIC
- life        1     2.392 159.72 46.431
<none>                    157.33 49.292
- GDP         1    11.165 168.50 53.703
- corruption  1    14.991 172.32 56.757
- freedom     1    29.141 186.47 67.489
- ssupport    1    61.834 219.17 89.459

Step:  AIC=46.43
tScore ~ GDP + ssupport + freedom + corruption

             Df Sum of Sq    RSS    AIC
<none>                    159.72 46.431
- corruption  1    16.698 176.42 55.041
- freedom     1    28.288 188.01 63.694
- GDP         1    28.881 188.60 64.123
- ssupport    1    75.572 235.30 94.204
```

Figure 12: ANOVA for Model 4 vs Model 5



```
## Analysis of Variance Table
##
## Model 1: tScore ~ GDP + ssupport + freedom + corruption
## Model 2: tScore ~ GDP + ssupport + life + freedom + corruption
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    131 159.72
## 2    130 157.33  1     2.392 1.9765 0.1621
```

Figure 13: Scatter plot matrix

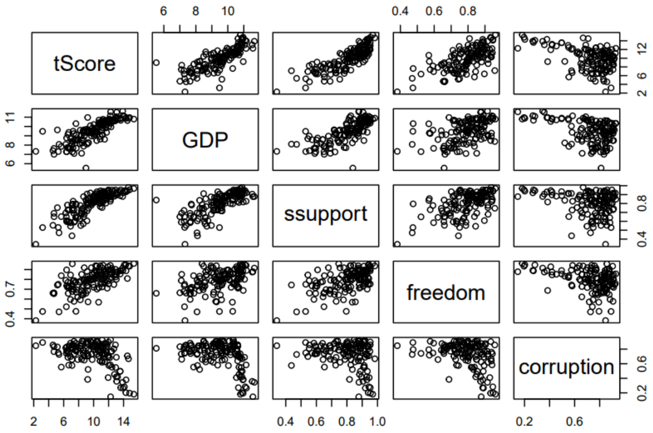
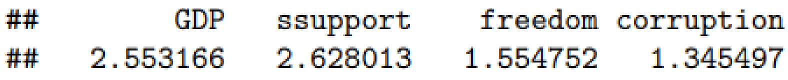


Figure 14: VIF of the variables in the final model

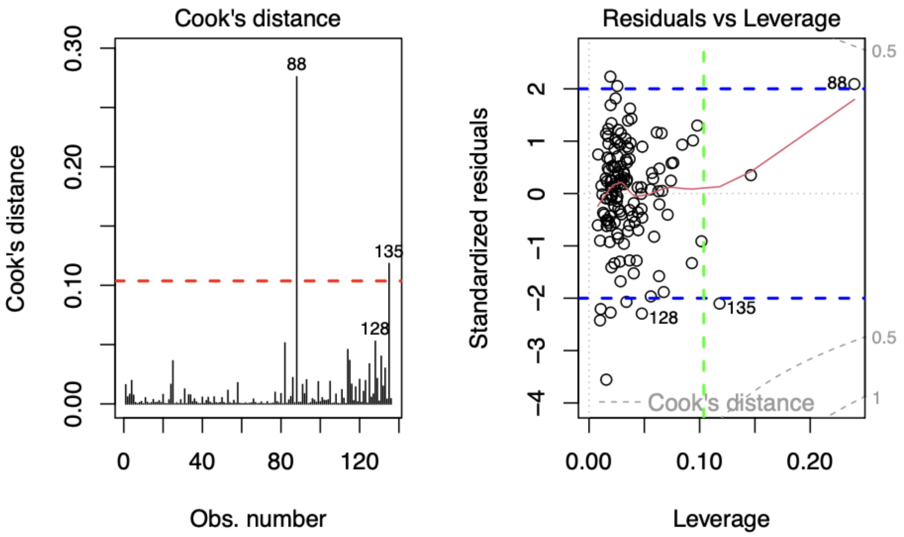


```
##       GDP  ssupport   freedom corruption
##  2.553166  2.628013  1.554752   1.345497
```

Figure 15: Correlation of all variables in the final model

```
##             tScore     GDP ssupport freedom corruption
## tScore      1.0000  0.7908   0.8337  0.6567    -0.4889
## GDP         0.7908  1.0000   0.7418  0.4494    -0.4354
## ssupport    0.8337  0.7418   1.0000  0.5465    -0.2755
## freedom     0.6567  0.4494   0.5465  1.0000    -0.3814
## corruption -0.4889 -0.4354  -0.2755 -0.3814     1.0000
```

Figure 16: Plotted Cook's distance and leverage for each observation


Cook's distance
Cook's distance
Obs. number
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage