Happiness Index: a Linear Model

Emma Morrison, 406011062

2024-03-16

1. Introduction and Methods

The purpose of this regression is to determine the leading predictors that influence our happiness. As the capitalist world booms, our happiness is encoded to be determined by our place in the economic structure, the life we were born into, and the ability of our social class to change. This data set however, examines several factors affecting our ability to be happy–questioning the notion that happiness is dependent on money. These assumptions are present in the first report conducted by The *World Happiness Report*, presented by Gallup, the Oxford Wellbeing Research Center, the UN Sustainable Development Solutions Network, and the WHR's Editorial Board.

However, this linear regression and the original report both aim to contradict this notion. In this study, several predictors are examined, including GDP (gross domestic product), healthy life expectancy (based on data from the World Health Organization and World Development Indicators), social support, life choices, generosity, and governmental corruption (all four based on the average of a binary answer, collected from around the world). The happiness "score" is based on self-reported data from a "ladder" ranking, with a range 0-10. This report will be informed on the combined data from the 2018 and 2019 reports, which was collected from Kaggle. The combined datasets include 312 observations on 7 distinct variables, collected from 156 countries.

2. Statistical Analysis Methods

The statistical analysis for this study was conducted in several steps. Firstly, the data was cleaned, and the joint 2018/2019 data set was created. This was in order to increase total working data set. Each set was testedseparately first (with linear assumptions and a multiple linear regression model), in order to determine if it would be necessary to combine the data. In the end, combining the two years provided a larger set to perform more complex analysis with, which helps to draw a bigger conclusion over time.

Several methods were employed to examine the initial predictors, which included investigating the assumptions for multiple linear regression (independence, linearity, normality, and homoscedasticity), removing influential outliers using leverages, residuals, and Cook's distance, and forming a final data set to perform a model on.

Following these tests, an exhaustive method was used to determine the best model. This method was chosen as opposed to a backwards or forwards regression model, as there were only seven predictors, and the length of the test was necessary to find the best fit model. Following these steps, I arrived at the best fit linear model, the results of which are outlined below.

3. Results and Discussion

Firstly, there was an investigation of the correlation between each variable. The correlation plot is shown below (Figure 1). There is high correlation between many of the variables and "score" which led score to be chosen as the dependent variable. From the corrrelation plot, there is little to no correlation between "generosity" and "corruption" and the other variables. These led to a test of the linear model next.

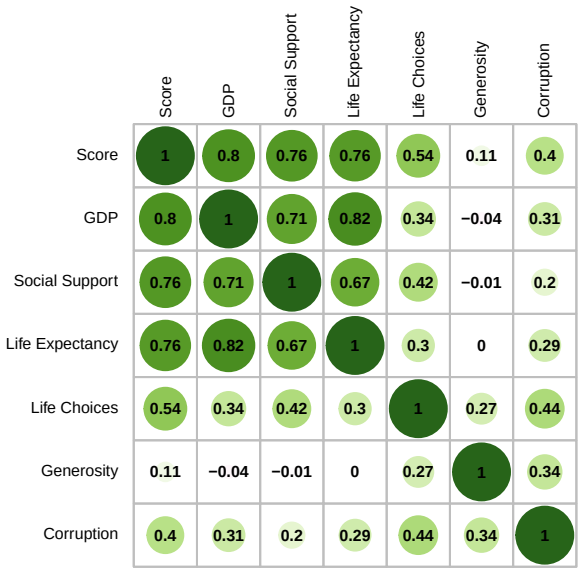

Figure 1: Correlation Plot

Below is the summary of an initial linear model including all six predictors on the dependent variable, happiness score. The initial ANOVA analysis showed several key factors. There are clearly high p-values in the predictors "generosity" and "corruption," which, like our correlation plot, indicate these predictors may not be significant. However, there is not enough evidence to throw these predictors away yet.

term	df	sumsq	meansq	statistic	p.value
gdp	1	246.038794	246.0387939	897.252304	0.0000000
social_support	1	28.963588	28.9635879	105.624180	0.0000000
life_expectancy	1	6.079748	6.0797480	22.171576	0.0000038
life_choices	1	18.804018	18.8040182	68.574343	0.0000000
generosity	1	1.466483	1.4664833	5.347960	0.0214120
corruption	1	1.399071	1.3990714	5.102123	0.0246026
Residuals	305	83.635151	0.2742136	NA	NA

When a summary of the initial linear model was run, the $R^2_{adj} = 0.7793$, which tells us the linear model only predicts 77.93% of the variation in the happiness score. Thus, the first step was to find influential points. In running diagnostic plots of the linear model, the Q-Q plot stuck out, as it was very linear, and showed a promising linear correlation between the predictors. However, the Residuals vs. Leverage plot did show a few potential outliers based on standardized residuals and Cook's Distance (Figure 2).

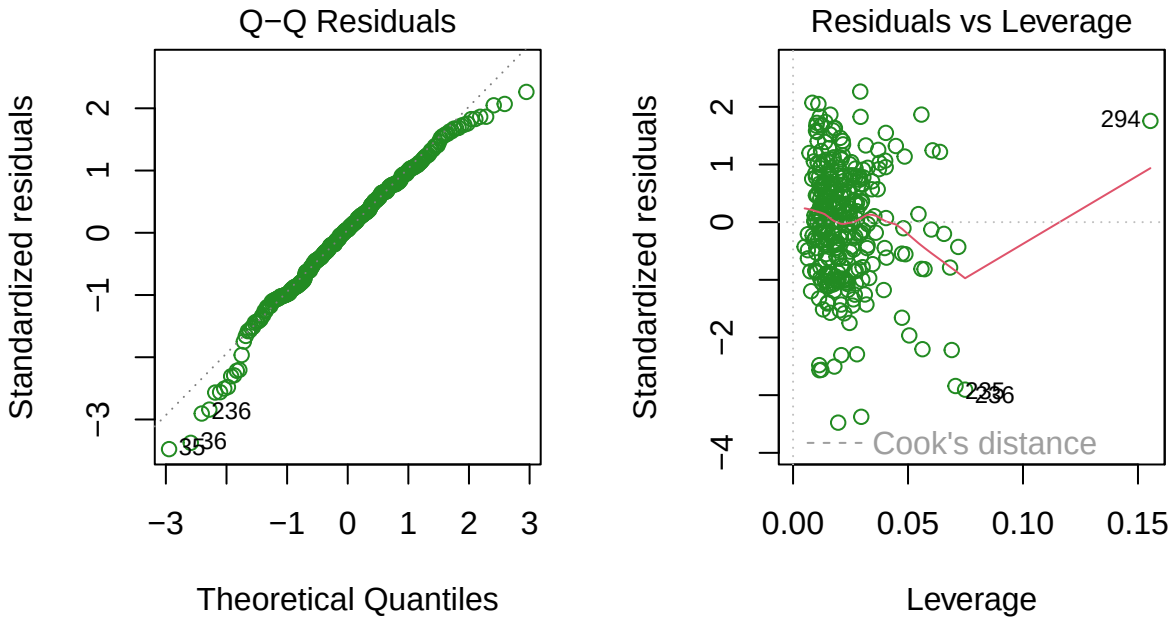

Figure 2: Diagnostic Plots for Original Linear Model

In performing extensive calculations of Cook's Distance, standardized residuals, and leverage values of potentially influential points, many outlier points were found (Figure 3). These values were removed from the data set, resulting in a higher R^2_{adj} value of 0.8414.

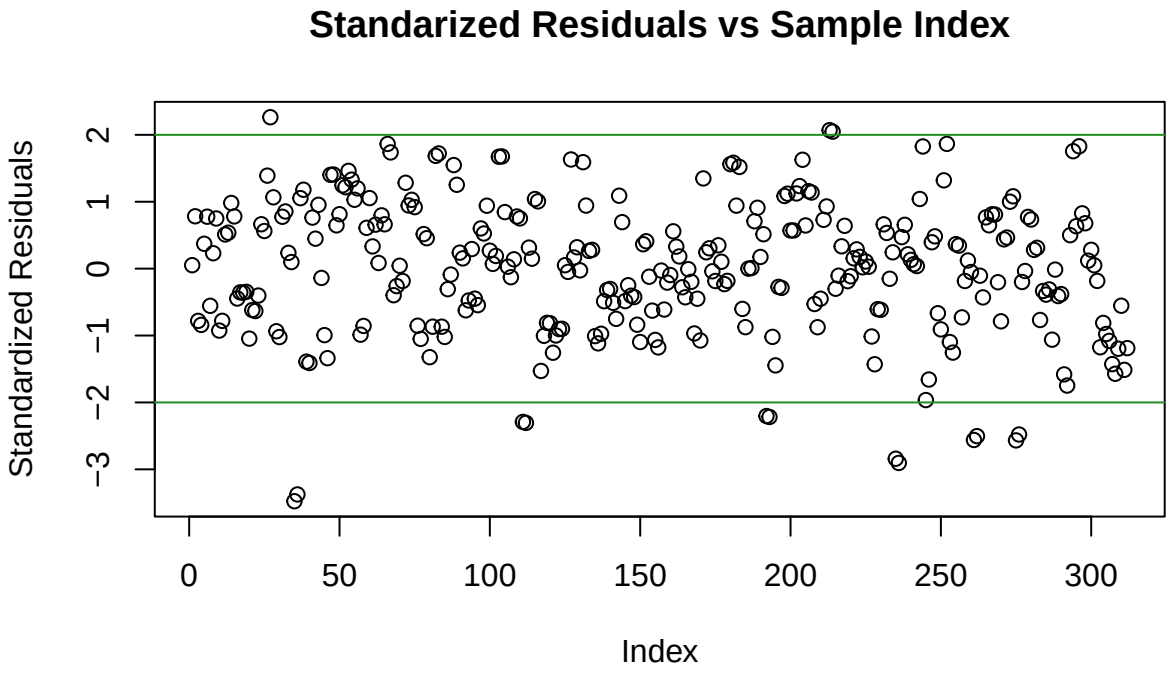

Figure 3: Outlier Plot

After removing outlier points, a reduced model was the next step. Added-variable plots were created for each of the variables, and the two predictors generosity and correlation had somewhat horizontal lines (Figure 4). The lines are very close to horizontal, emphasizing once again that these predictors do not add much to

the model. However, this does not mean that they are not influential, so determining the calculations for a reduced model was needed.

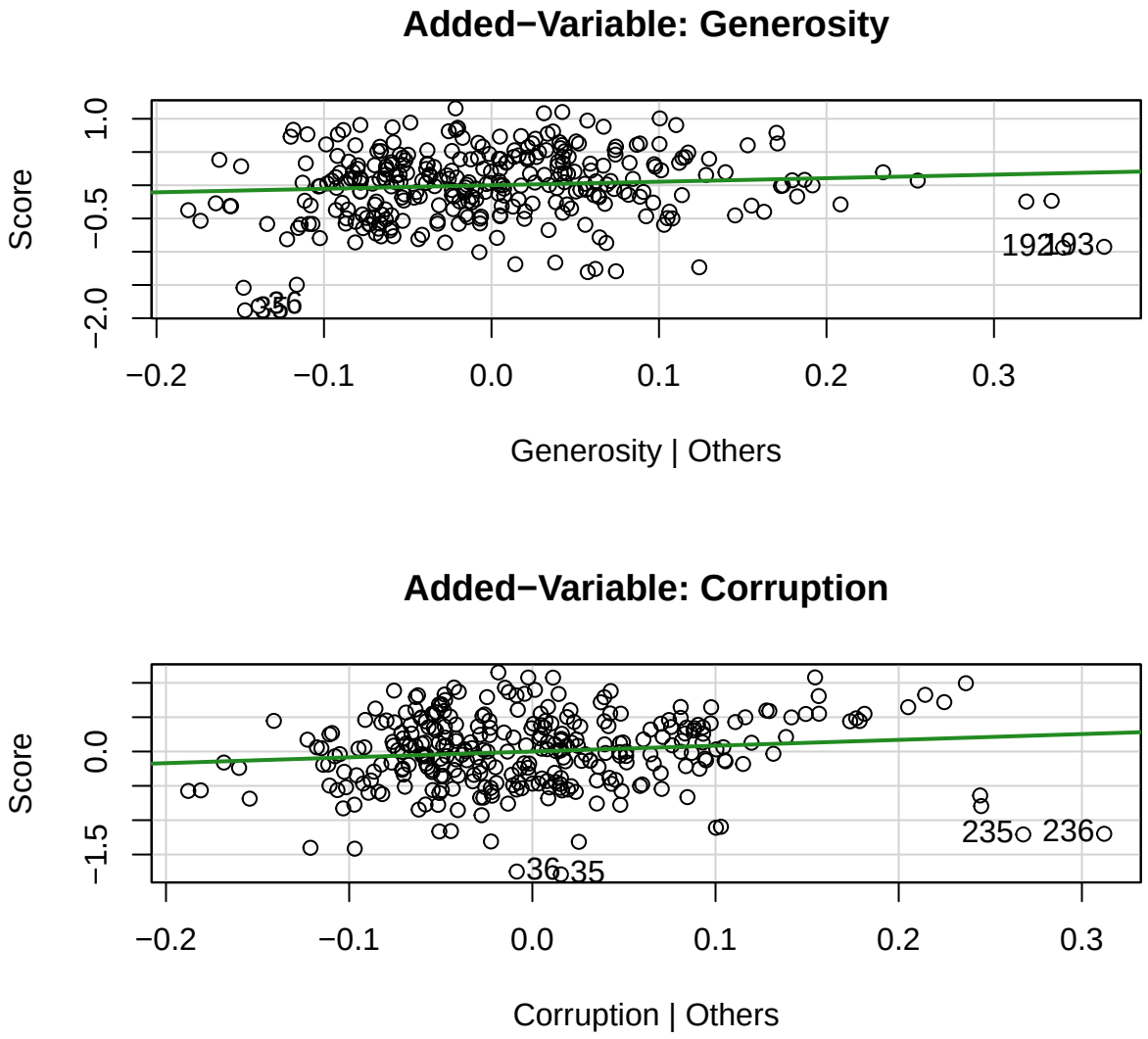

Figure 4: Added Variable Plot

A calculation was run measuring R^2_{adj}, BIC, and CP (like AIC, but differing by a constant). The results from this exhaustive calculation are shown below. This method determined that for R^2_{adj} and CP, the six variable model was the best. BIC further penalizes more complex models, which might explain why it favors the five variable model. Despite this, the six variable model appears to be the best model.

predictors	adj_r2	cp	bic
1	0.6630993	333.52498	-312.7438
2	0.7705899	134.18150	-422.1888
3	0.8104806	61.05452	-474.2400
4	0.8269791	31.49015	-496.6120
5	0.8390683	10.22234	-513.4495
6	0.8413699	7.00000	-513.0566

4. Final Linear Model

Based upon my results from variable testing, removing outliers, and diagnostic plots of the final linear model, there does not appear to be a reason to transform the data. Upon trying, I found much lower R^2_{adj} values, and non-linear Q-Q plots.

This leads me to my conclusion that the best linear model for the Happiness Index is:

Score = 1.8328 + 0.8923*GDP + 1.0238*Social Support + 0.9276 *Life Expectancy + 1.6011 *Life Choices + 0.7021*Generosity + 1.2938*Corruption + ϵ

Based on my final linear model:

term	estimate	std.error	statistic	p.value
(Intercept)	1.8328282	0.1213551	15.103015	0.0000000
gdp	0.8923459	0.1233352	7.235127	0.0000000
social_support	1.0238429	0.1297860	7.888702	0.0000000
life_expectancy	0.9276183	0.1789732	5.183000	0.0000004
life_choices	1.6010539	0.1975924	8.102812	0.0000000
generosity	0.7020789	0.3072229	2.285243	0.0230204
corruption	1.2937737	0.3336158	3.878035	0.0001303

P-value: $<$ 2.2e-16, so we conclude that each of these predictors is significant. The final R^2_{adj} for the model is 0.8414, compared to an R^2 value of 0.8446.

References

Sustainable Development Solutions Network and Larion, A. (2018). World Happiness Report, Version 2, 2018 Data. Retrieved March, 12, 2024 from https://www.kaggle.com/datasets/unsdsn/world-happiness?select=2018.csv.

Helliwell, J., Layard, R., & Sachs, J. (2018). World Happiness Report 2018, New York: Sustainable Development Solutions Network.