World Happiness Reports: Regression Analysis

Ethan Beffert, Liam Gaeuman, & Emma Johnson

Mathematics & Statistics Department, University of Minnesota-Duluth

STAT 4511 (001): Regression Analysis

Igor Melnykov

Dec. 5, 2023

Section 1: Project Overview

The purpose of this study was to analyze how a variety of variables impact the collective happiness of residents within a country. We looked at data from 137 different countries to figure out which model would be best for mapping out our response variable. The response variable is a national average ladder score of happiness (Y_i), scaled from 0 to 10. This score was based on responses to questions regarding subjective life evaluations. Initially, we examined 6 predictor variables to utilize in our model. The first variable (X_1) was GDP per capita, a logged variable of constant international dollar prices. Notably, using a logarithm on GDP helps avoid skewing in model validity testing. The second variable is social support (X_2), a national average of binary responses (0 or 1) to a polling question regarding the perception of people to depend on. Variable 3 was healthy life expectancy at birth (X_3), which was gathered from the World Health Organization Global Health Observatory data repository. The fourth variable given is the freedom to make life choices (X_4), a national average of binary responses (0 or 1) to a polling question regarding the level of satisfaction corresponding to the freedom to choose what to do with one's life. Variable 5 is generosity (X_5), which is the residual to the question of regressing the national average in response to the GWP question "Have you donated money to a charity in the past month?" on GDP per capita. Finally, variable 6 is corruption perception (X_6), the national average of the sum of two binary responses (0 or 1) to questions regarding the widespreadness of perceived corruption within a country.

Through a variety of tests, graphs, and interpretations the model that best fits our data came out to:

$$Y_i = -1.4164 + 0.2537X_1 + 4.1994X_2 + 2.3321X_4 - 0.8695X_6$$

Life expectancy and generosity were dropped from the model and no transformations or alterations were made to the model. The model was tested using model selection criteria such as, multicollinearity, interaction terms, higher-ordered terms, normality, equal assumptions of variance, possible transformations, and outliers. Our results showed to not change or transform any of our variables, as that's the best model to fit our data. Based on our results in this test we feel confident that our model could properly predict a country's happiness rating (0-10) given its logged GDP, social support, freedom, and corruption.

Section 2: Data Selection

The dataset emerged from the open data source World Happiness Reports (WHR) (Helliwell et al., 2023). They drew variables from internationally available information synthesized from the Gallup World Poll, the World Development Indicators, the World Health Organization, the World Bank, and the Worldwide Governance Indicators project. The decided outcome variable, a ladder score of world happiness ratings, reflected group interests in social phenomena. We chose this data set as one of our group members has recently taken the course Global Issues at UMD and this data is very applicable to that course. Another group member is interested in the concept of subjective well-being which this data set touches on.

The explanatory variables presented logical agreements with the happiness ratings, as influenced by the WHR statistical appendix. The narrowed-down pool of explanatory variables became logged GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity, perceptions of corruption, and country of residency. We selected these variables as they were the only ones passed in our data set. As a group, we wanted to make sure we evaluated all variables possible to give the best model we could.

Section 3: Model Synthesis

Final Model

The model we decided on to explain our response variable, the ladder happiness score, is shown as follows:

$$Y_i = -1.4164 + 0.2537X_1 + 4.1994X_2 + 2.3321X_4 - 0.8695X_6$$

Where X_1 is logged GDP per capita, X_2 is social support, X_4 is freedom, and X_6 is corruption. For every unit increase in each variable, assuming all other variables are held constant, the average change in happiness ratings is defined as:

- An increase of 0.2537 in happiness score as national logged GDP increases by one unit.
- An increase of 0.41994 in happiness score as the national average social support score increases by 0.1.
- An increase of 0.23321 in happiness score as the freedom to make life choices score increases by 0.1.
- A decrease of 0.08695 in happiness score as the national perception of corruption increases by 0.1.

*The intercept coefficient has no useful interpretation.

Hold-Out Data Set

To validate our model, we opted for a hold-out data approach employing a training and testing set of the collective data frame. In doing so, we randomly split the data in half, assigning one half for modeling (i.e. training set) and the other half for prediction validation (i.e. testing set). This yielded separate testing and training samples of 68 and 69 countries, respectively. When addressing specific model assumptions, we utilized the half-collective sets to ensure that the final model could be applied to the other half, since we did not have information to assess extra countries outside of the data set selected. A seed was encoded so that our results could be shared under the same random index. The linear regression model under our indexed training set, with the same variables defined previously, is shown below.

$$Y_i = -2.5495 + 0.3270X_1 + 3.7855X_2 + 3.2090X_4 - 0.8413X_6$$

Model Selection

Selecting Variables

The first step in arriving at our final linear regression model was narrowing down which variables to utilize. We ran a series of model selection criteria to choose the explanatory variables that best fit our response variable. These tests are shown below in Figures 1.1, 1.2, and 1.3. Figure 1.1 shows that a model with logged GDP, social support, freedom, and corruption yields the lowest Bayesian Information Criterion (BIC) values (around -97). Figure 1.2 supports these same variables under the index of Mallows' Cp, where p is equal to 5 with the four variables selected, and the best model under this criterion reflects a minimized Cp of 4.3. Lastly, the adjusted R squared method in Figure 1.3 supports these same four variables, suggesting an adjusted R squared value of 0.81.

Figure 1.1

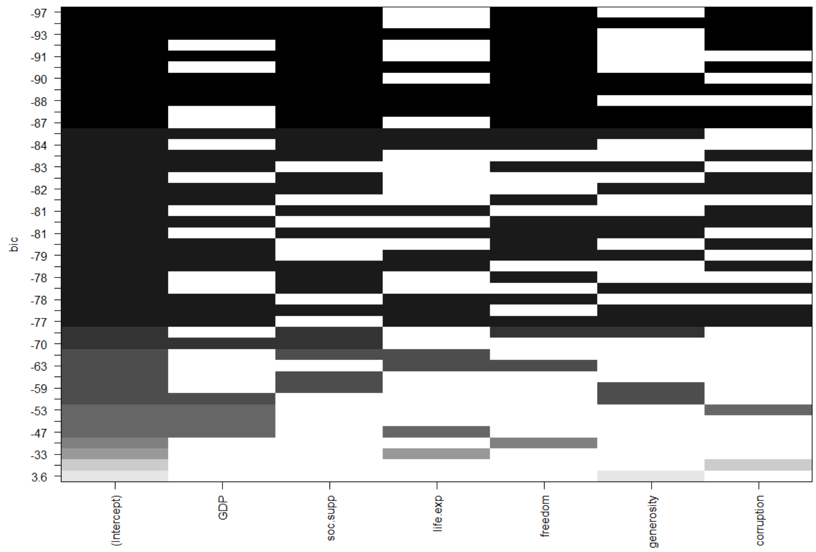


Figure 1.2

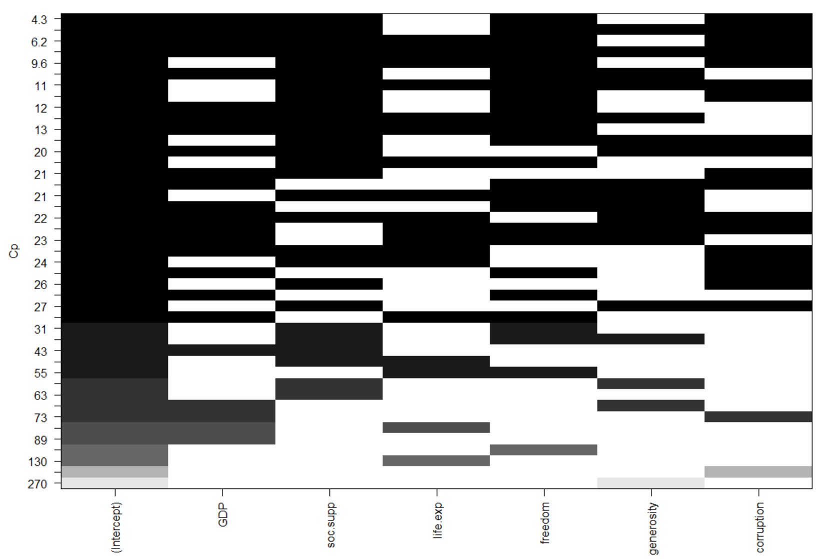
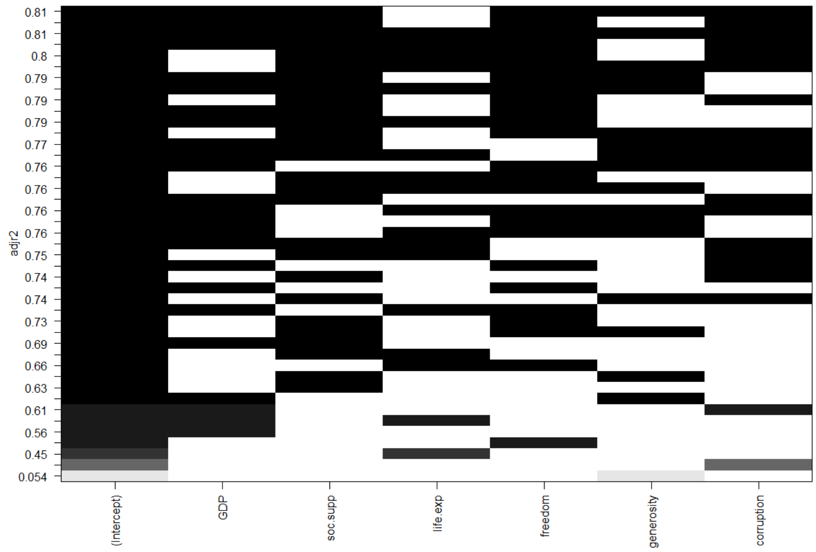


Figure 1.3



Further, we assessed potentials for multicollinearity by graphing the variables in pairs and taking a look at their correlation coefficients. Figure 2 visualizes the correlation relationships among the variables, whereas Figure 3 displays the correlation coefficients. Notably, GDP, social support, generosity, and life expectancy show the highest correlations with happiness scores. Perceptions of corruption seem to be negatively correlated with happiness. On the other hand, generosity was very weakly correlated with any variable, especially the response variable. Life expectancy was not as obvious but correlated with the majority of the other variables. This raises a concern, along with other variables displaying a higher degree of correlation, of whether or not the variable can be left out of the model due to other variables displaying more explanatory power for happiness ratings.

Figure 2

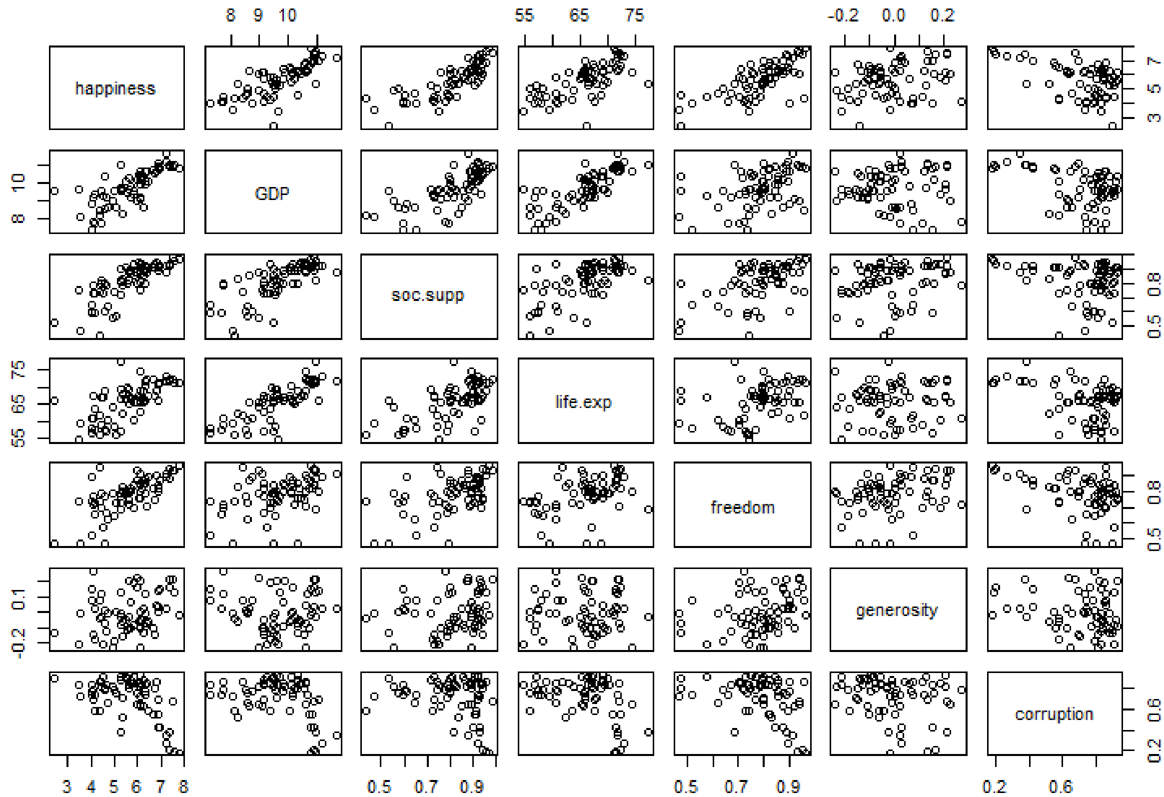
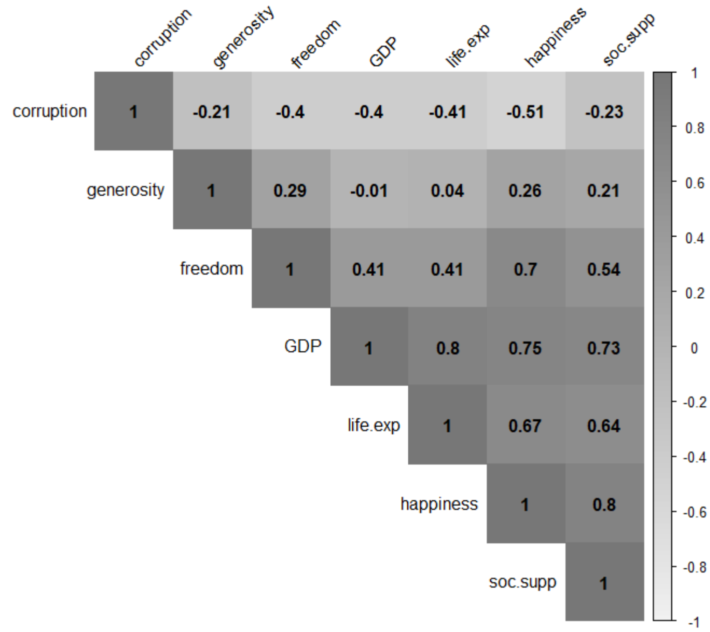

happiness
GDP
soc.supp
life.exp
freedom
generosity
corruption
8 9 10
55 65 75
-0.2 0.0 0.2
3 5 7
0.5 0.8
0.2 0.6
3 4 5 6 7 8
0.5 0.7 0.9
10
8
75
65
55
0.1
-0.2

Figure 3



Based on the model selection criteria methods of BIC, Cp, and adjusted R squared, we opted to leave life expectancy and generosity out of further model testing. The results show us the best models contain GDP, social support, freedom, and perceptions of corruption. Life expectancy is highly correlated with GDP and social support and is thus redundant and can be removed.

As another measure of multicollinearity, we observed the Variance Inflation Factors (VIF) of the chosen explanatory variables. The VIF results are shown below in Table 1.

Table 1

Variable	GDP	Social Support	Freedom	Corruption
VIF	2.481436	2.686079	1.622082	1.385124

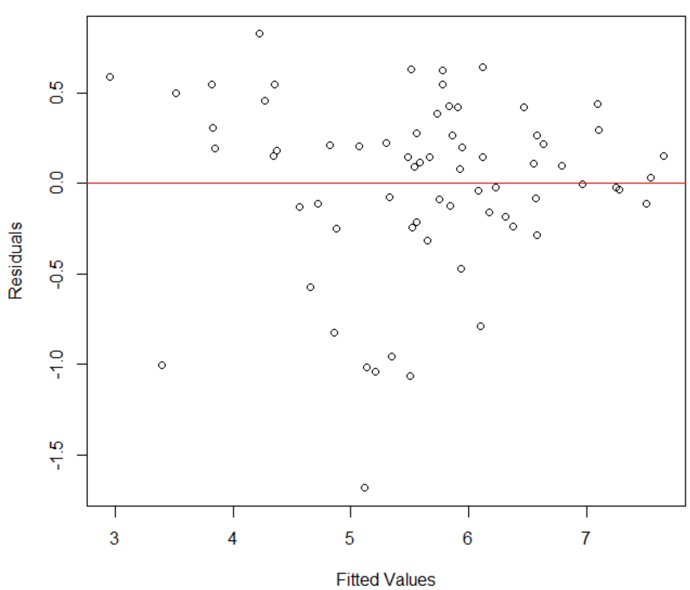
The variance inflation factors are all well under 10, indicating that there is no problematic correlation among the variables chosen. Therefore, we continued with the model utilizing these four variables to predict our response.

Analyzing Assumptions

Linear Form of the Regression Formula

As a first test of normality assumptions, we created a scatterplot of the residuals of the training model versus the fitted values. No obvious concerns were raised, as the residuals appeared arbitrarily dispersed (see Figure 4). The residual versus fitted values plot shows a high number of large negative residuals. This is not an immediate concern and we accept that the residuals display a decent linear form.

Figure 4



Normal Distribution of Errors

Further, a histogram of residuals (Figure 5.1) and a normal QQ-Plot of the residuals (Figure 5.2), raised a bit of concern. The histogram indicated a left-hand skew and the QQ-Plot has the tails of the residual points straying away from the normal QQ-line. This informs a lack of normal distribution of errors, suggesting that the proportions of the distribution are non-normal. It does not align with our assumption that the errors are independent and identically distributed on a normal distribution with a mean of zero.

Figure 5.1

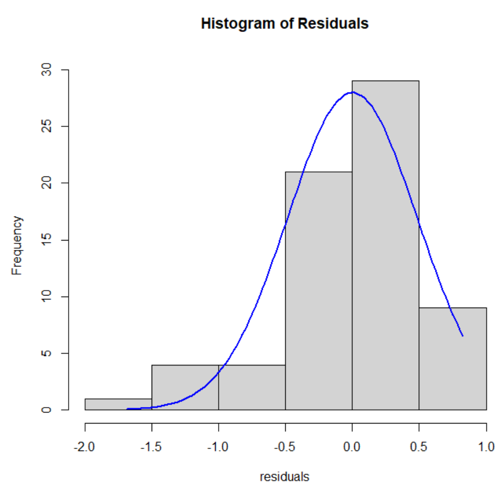
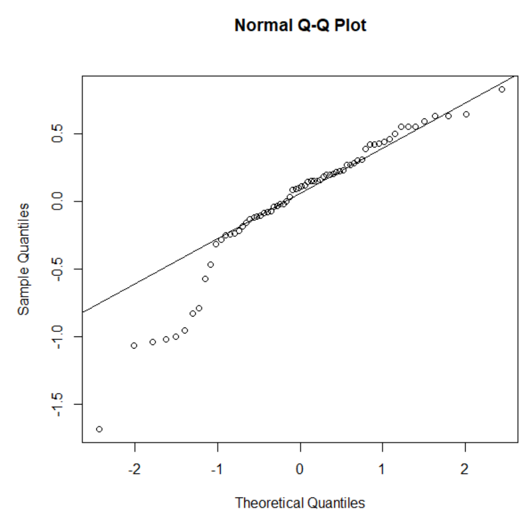


Figure 5.2

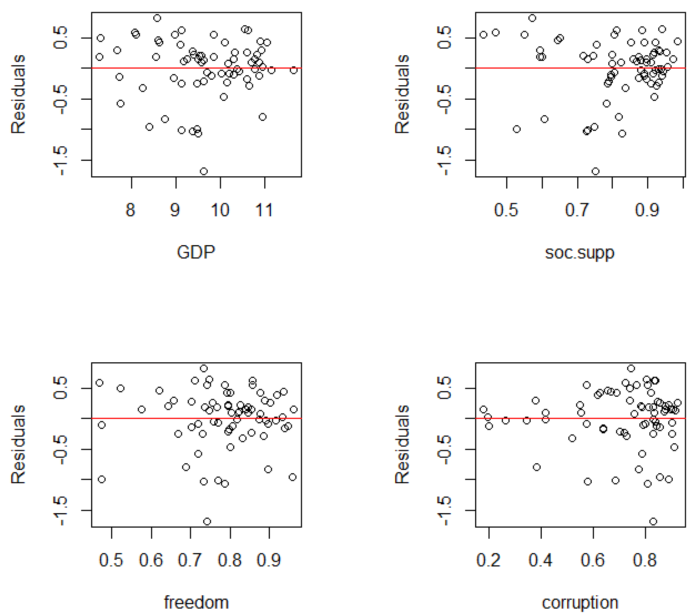


We ran a Shapiro-Wilk Normality Test, which explicitly led us to deny the assumption of a normal distribution of errors among our data: $W = 0.9122,\ p < .005$. This is due to several larger negative residuals which can be seen in the figures above, many of which will be flagged as outliers.

Constant Variance of Errors

By plotting each explanatory variable against the residuals from the training model, we were able to address whether the residuals appeared constant or problematic. Figure 6 displays these plots, showing no obvious signs of concern. The residuals appear to spread out and lack a megaphone pattern or other obvious issues. In line with the residual plot of the fitted values versus the residuals of the training model, the range of the residuals reaches more negative values than positive ones. This is notable, yet, not a significant worry until addressing further testing.

Figure 6



A formal test addressed the constant variance assumption. The null hypothesis was that the variance of errors is constant, whereas the alternative hypothesis suggested that the variance of errors is not constant. Specifically, we implemented a Breusch-Pagan Test, which tests the null hypothesis that the natural log of the variance of residuals does not depend on the explanatory variables in a linear manner. The test failed to reject the null hypothesis, suggesting that the variance of errors is constant: $X^2_{BP}(4) = 8.1783,\ p > .05$. Although not significant at an alpha level of 0.1, we are continuing with the assumption of the constant variance of errors.

Independence of Errors

A time-series plot was unnecessary in addressing the independence of errors in our data since time is not a framework for our variables. The residual plots (Figures 4, 6) touch on the independence of errors assumption, although there are no obvious issues in the display of residuals amongst these plots.

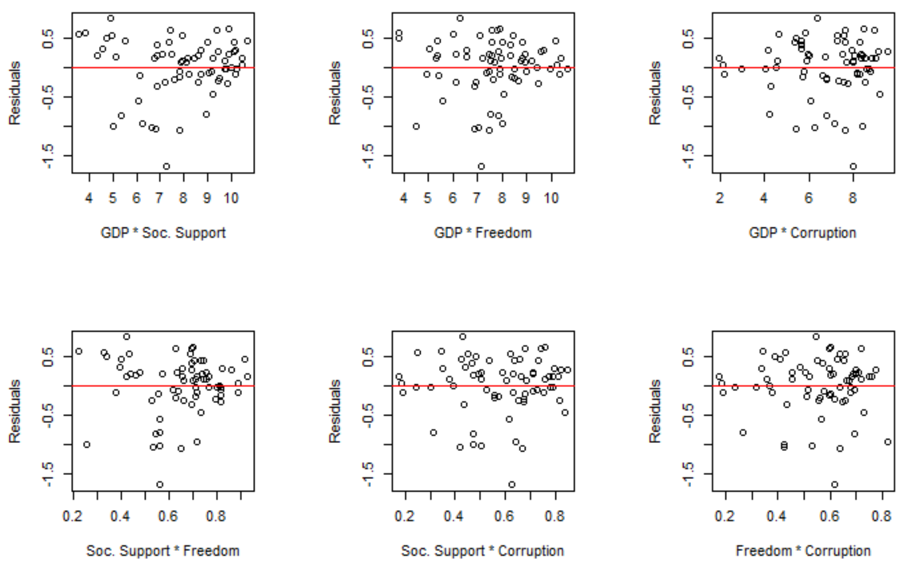
Transformations

To improve the normality assumptions, we attempted to transform variables in the model to fit a normal distribution better. In these efforts, we addressed the shape of the explanatory variables against the residuals (see Figure 6 above) to inform on potential curvature and patterns within the errors. The logged GDP and social support variables displayed slight curvature. Transformations were applied to square both of these variables independently, which was unhelpful in correcting the assumption checks. Similarly, square roots didn't change results, either. Further, we attempted to take the logarithm of social support and corruption, respectively, yielding no improvements, as well.

Interactions

We plotted the residuals against each potential interaction term to decide whether interaction terms would be advantageous in our model. Figure 7 displays these plots. As shown, no curvature or blatant pattern appears in any of these plots, so we left interaction terms out of further model considerations. To reiterate, neither the transformation methods nor the interaction potential helped reduce the errors, so we opted for outlier tests to inform us more about the apparent skew of the data residuals in the negative direction.

Figure 7



Outliers and Influential Observations

Outliers in the Response Variable

We identified potential outliers in the response variable by comparing the magnitude of the studentized deleted residuals to the mean square error from the regression function omitting the observed response values. Under this method, we computed $t(62) = 3.55275$ with Bonferroni adjustment of an alpha level equal to 0.05. The country with a studentized residual magnitude greater than this t-value was Botswana. In fact, Botswana is the point in Figure 5.2 that settles in the lower left, deviating the furthest away from the QQ-Line. Eliminating this point is expected to reduce the negative skew of the residual distribution.

Leverage

We computed which leverage values were higher than 0.1471, an index reflecting our number of explanatory variables and sample size. We considered Benin, Lebanon, Turkey, Madagascar, and Comoros as outliers under the leverage criterion. It appears that these variables exhibited abnormal explanatory observations, out of line with the current fit of the model.

Influential Observations

To assess individual influences on fitted values for the model, we ran the difference in fits (DFFITS) diagnostic where we considered magnitudes above 1 as influential, due to the size of our data set. We found that Lebanon had the largest DFFITS influence.

Next, we took a look at which cases had a large influence on the entirety of fitted happiness scores. Under Cook's Distance method, we found no countries displaying a large influence on the collective predicted response.

Model Validation

Before taking the final steps toward unifying our final model, we must address some obvious concerns. Our training model failed some important normality assumptions, as noted early on when analyzing the normal distribution of error potential. The histogram of residuals, normal QQ-Plot, and Shapiro-Wilk Test all supported skew in the residuals of our data, indicating further investigations. Once we ran tests for outliers and influential points, we gathered more information on the cause of this skewed distribution. It is overwhelmingly likely that the data did not follow a normal distribution because of these outliers.

Predictions for Testing Set

Using the model fitted on the training set, we predicted the happiness score for the remaining 69 observations in our testing data set. The mean square prediction (MSPR) on the testing set was 0.2437, compared to the mean square error (MSE) from the training set which was 0.25. The two values are comparable, however, the MSPR should be greater than the MSE. We believe ours is around the same due to an error in overfitting as our model possibly learned the training data too well.

Synthesizing Final Model

Combining the data from our split training and testing sets produced the final linear regression model:

$$Y_i = -1.4164 + 0.2537X_1 + 4.1994X_2 + 2.3321X_4 - 0.8695X_6$$

Section 4: Generalizability

As briefly mentioned in previous sections, it may be difficult to apply this model to countries not included in the data set, as they were likely omitted or disregarded due to a lack of public information on the variables included. However, this model could be used within the set of countries to predict their happiness scores given changes in a specific variable.

For instance, we could look at how high corruption would affect the happiness of a country. We attempted to show this by inputting the mean of logged GDP, social support, and freedom, respectively, into a hypothetical hypothetical country with a corruption score of 2 standard deviations above the mean for that variable. Comparing this to the predicted happiness score of a country exhibiting the mean across all explanatory variables shows directly how corruption affects happiness under the model. We saw that the happiness score of this hypothetical country was 5.07 which is lower than the mean happiness score of 5.54.

We ran a similar test by taking the mean scores for logged GDP, freedom, and corruption while inputting a social support score that is 2 standard deviations above that mean. This yielded a hypothetical country happiness score of 6.32, which is higher than the average score of 5.54.

By looking at these two predictions, it is clear that social support affects the happiness of a country more than the perceptions of corruption do. This is one example of how predictions utilizing our model could statistically analyze how different variables affect happiness and to what extent they do.

Section 5: Addressing Expectations

We found that our model was generally a good fit, although we ran into various issues when addressing normality assumptions. Of course, a linear regression model such as ours must align with these assumptions, otherwise, it lacks validity and reliability. Addressing these concerns became the core of transforming our model. Another difficulty came about when deciding how to transform our model. This included multiple attempts to transform variables with logarithms and roots, as well as assessment of higher-order terms. We learned that there often is no obvious solution to these problems and that it is important to try multiple different methods to improve the model before settling on one.

We also learned how to deal with the type of data and variables utilized in our model. When interpreting coefficients, we realized that the nature of the scores mattered in accordance with the interpretation of the variable. When a variable is measured on a small scale, such as ours with averages of binary responses, interpretation of coefficients also follows a small scale. This stressed the importance not only of competence in statistical analysis but also in an understanding of the implications of variables gathered.

The process underlying regression analysis and linear modeling requires particular attention to detail. Gathering our knowledge through this course and applying it in our model synthesis highlighted this.

References

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.