James Sanders

Stat 410 Linear Regression

Dr. Daniel Kowal

April 2020

Final Project Report

Project Statement:

The purpose of this analysis is to investigate the factors that lead to countries to have high average life satisfaction levels. Specifically, we will be investigating the effect that GDP per capita, social support, life expectancy, freedom, generosity, and corruption have on average life satisfaction. We will be analyzing all these variables on a country wide scale, and examining how the importance of these factors changes in societies of different wealth. We are doing this because the more we can figure out about what truly makes societies happy, the better we'll be able to improve our own society.

Data Description:

The data on happiness, social support, freedom, generosity, and corruption are all from the 2020 release of the Gallup World Poll. Thousands of individuals in all countries around the world were surveyed and asked subjective questions about their life, such as whether they have a friend they can count on whenever they need them, or how they would rank their life on a scale from 0 to 10. The GDP per capita statistics come from the World Bank's annual report on economic prospects, and the life expectancy statistics are from the World Health Organization's Global Health Observatory data repository. These data are helpful in answering the research question, because

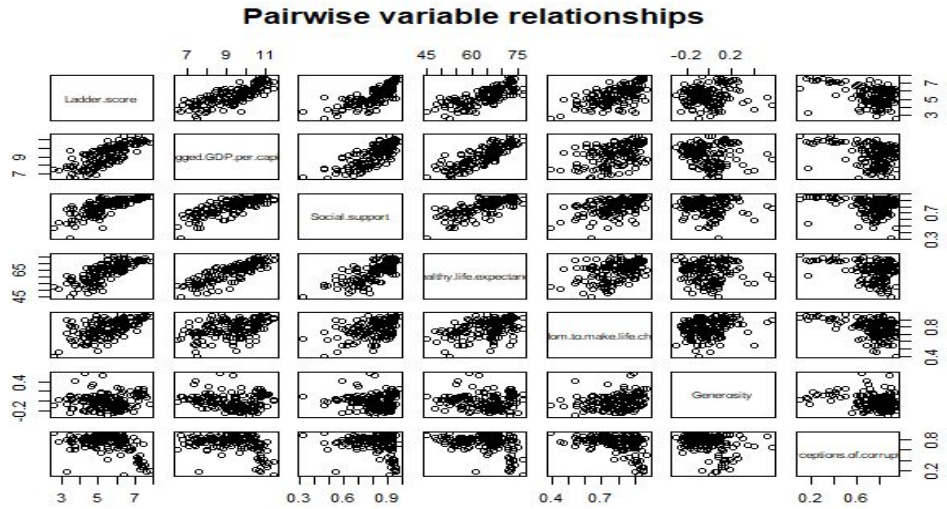
they are all parameters widely thought to be influential in personal well being, so we should be able to construct a model of estimated life satisfaction from them.

Exploratory Data Analysis:

The variables included in this analysis are summarized in the table below

Variable Name	**What it is**
Name	Name of the country
Happiness Level (or "ladder level")	Average national response to the answer to the following question in the Gallup World Poll: "Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?"
GDP Per Capita	The purchasing power parity (in 2011 international dollar prices) level of estimated 2019 GDP per capita from the World Bank for the given country
Healthy Life Expectancy	Projected average life expectancy within the given country for a healthy baby at birth from the World Health Organization
Social Support	National average response to the Gallup World Poll (GWP) binary survey 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?"
Freedom to make life choices	National average response to the GWP binary survey question "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"
Generosity	The residual of regressing national average of response to the GWP question "Have you donated money to a charity in the past month?" on GDP per capita
Corruption Perception	National Average response to the two GWP binary questions "Is corruption widespread throughout the government or not" and "Is corruption widespread within businesses or not?"

The variables for happiness, social support, freedom to make life choices, generosity, and corruption perception all come from the same survey. The plot below shows the pairwise relationships between each pair of variables (excluding “name”).

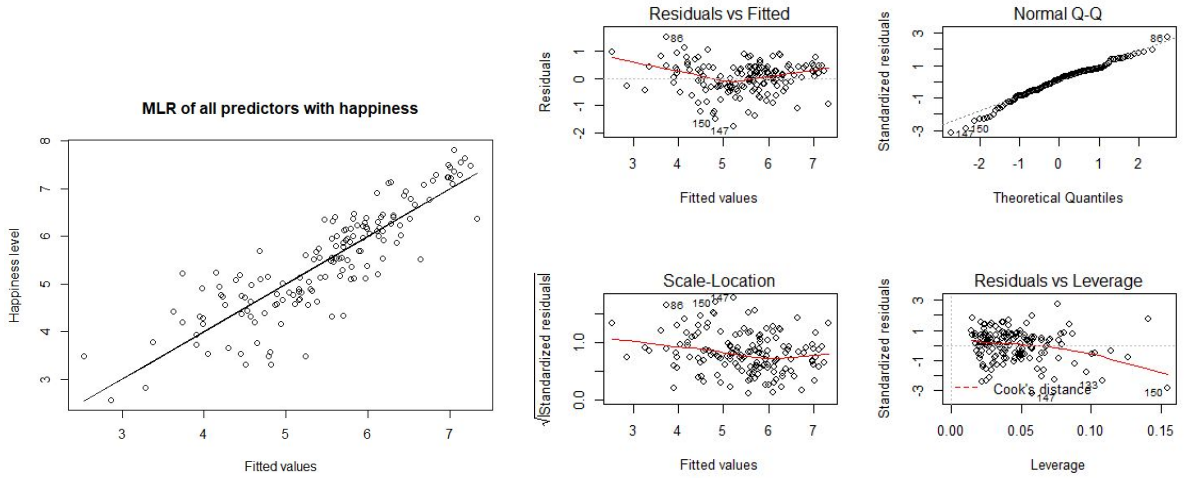


As can be seen, there is a clear positive association between happiness, gdp/capita, social support, and life expectancy. All these variables also have a positive (yet weaker) association with the freedom to make life choices. All the variables are continuous, so they all make nice cloudy scatterplots.

Data Analysis:

We began by fitting a multiple linear regression to predict happiness as a function of all six predictors. Every predictor except for generosity was found to be significant under the 0.05 level. To address the issue of multicollinearity, the variance inflation factors were computed for each of the predictors. The highest VIF was GDP with a value of 4.56, meaning that GDP is highly associated with the other predictors. However, this is below our cutoff value of 5, so it was determined that multicollinearity was not a significant issue. This MLR gave an adjusted R^2 value of 0.738,

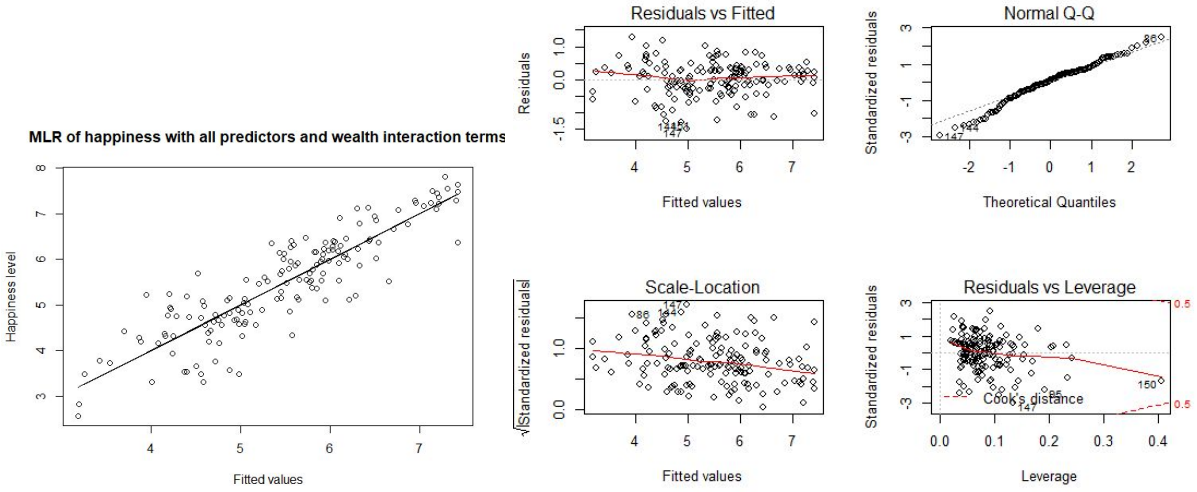
suggesting that while the six predictors do a good job of explaining most of the data, there is a significant amount of variation still not explained by the model. The plot of the MLR and diagnostics are below.



The graph of fitted values vs. residuals appears to show a slight upward parabola shape, suggesting that a strictly linear model might not truly be capturing what is really going on. The Normal Q-Q plot shows all values generally along the straight line, suggesting that the assumption of normal errors was valid. The Scale Location plot appears to be scattered without any discernible pattern, suggesting that the assumption of constant expected variance was valid. There are no leverage points outside of Cook's distance of 0.5, which suggests that there are no bad outlier points.

Next, in order to study how these effects differed between wealthy countries, an indicator variable was added in order to designate whether the country was wealthy or not. "Wealth", was defined as having a GDP/Capita over $12,375, which is the cutoff level between an upper middle class country and a high income country as defined by the World Bank (World Bank). Next, an MLR that included all six predictors and their interaction with "Wealth" was constructed. This had a large effect on the significance results. GDP was no longer significant, suggesting that if a country is

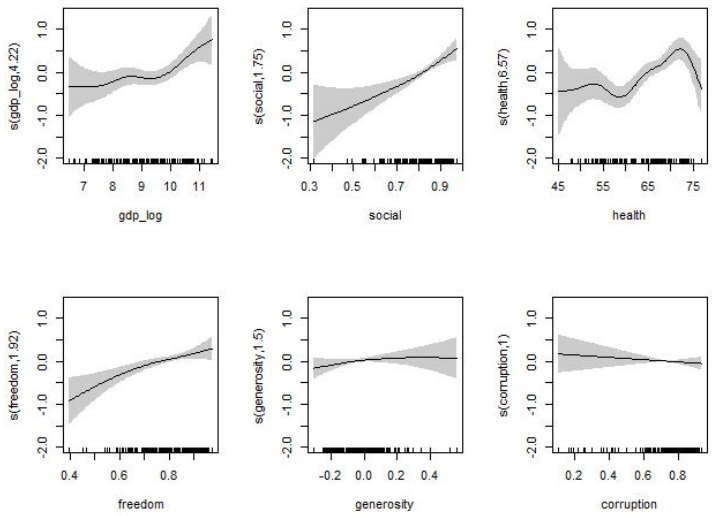
already wealthy or already not Wealth, slight changes in GDP don't do much to change happiness levels. The other predictors (excluding Generosity) were still significant, which matches the results found from the first MLR. However, the interaction terms between Wealth and Social and between wealth and corruption were found to be significant. This suggests that these are even more important in wealthy countries than they are generally. The most fascinating result, though, is that the Corruption coefficient is positive and significant, while the Corruption Wealth interaction term is strongly negative and highly significant. This seems to suggest that while corruption has a huge negative impact on happiness within wealthy countries, it has a positive impact on healthy countries. While the result in wealthy countries makes intuitive sense, I have been unable to come up with an explanation for the effect in non-wealthy countries, and this would be an interesting area for more research. A partial F test found that there was a significant difference in prediction power between this MLR with the Wealth Indicator variable, and the MLR without it. This suggests that there we are justified in including this indicator in our model.



As can be seen in the diagnostics plot above, all our assumptions about the model seem to be met, as there does not seem to be a pattern between the fitted values and residuals, the Normal

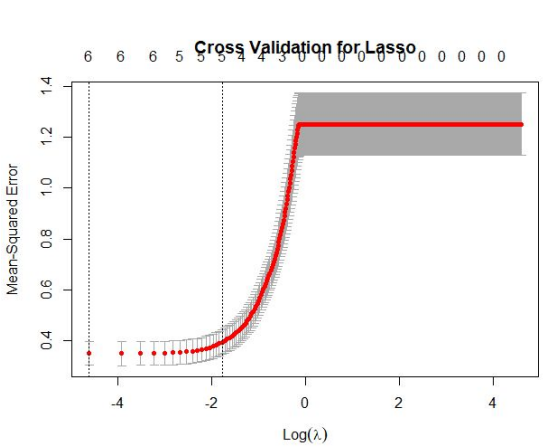
Q-Q plot appears to fall along a straight line, the Scale-Location plot seems well scattered, and there are no leverage points outside of Cooks' distance of 0.5.

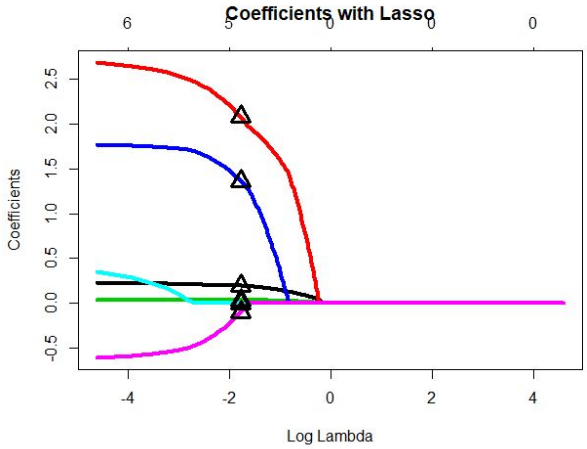
Next, in an attempt to create a better predictive model for happiness with the given predictors, an additive model was constructed using splines. All predictors were found to be significant to the additive model except for Generosity and Corruption. Within the additive model, GDP appears to not have a strong effect at all until it reaches a GDP/capita of $e^(9.5) or $1,336, which is roughly the cutoff for high income countries used by the World Bank. After this, GDP has a strong increasing effect on happiness. Interestingly, the effect from health is increasing until the life expectancy reaches about 72 years, at which point the effect is decreasing. This result does not make immediate intuitive sense, and deserves additional study. The effects of all components of the additive model can be seen below.



Next, a Lasso model was fit on the data in order to select the variables most important for the MLR. The lambda value that was chosen was one standard deviation from the lambda that minimized the MSE. All of the variables were selected as significant by the Lasso model except for

Generosity, which perfectly matches the results given by the original MLR. The lambda selection plot and coefficient values plot can be seen below.





Finally, the data points were divided up into three groups: one for Low Income to Lower Middle Income economies, one for Upper Middle Income economies, and one for High Income economies. The cutoffs for these groups are given by the World Bank's thresholds (World Bank). There are 78 high income, 37 upper middle income, and 38 low to lower middle income countries. For each group, we conducted a basic MLR to see which predictors would still be significant within just the group, and we also constructed a Lasso model to attempt to discern which variables were the most important predictors. All of the same graphs were constructed and all of the same assumptions were tested for the MLR and Lasso on each individual group as on the full data set. However, for the sake of brevity, they will be reported less fully within this paper.

The results for Low to Lower Middle Income countries were by far the most interesting. None of the predictor variables were found to be remotely close to significant except for Generosity. This is fascinating because it seems to suggest that in these countries, the only reliable predictor of happiness is how free people feel like they are. The linear model had terrible predictive power, with an R^2 value of just 0.2332. This suggests that there is still much unknown and much to be studied

about happiness in these countries. The Lasso model found that only Freedom and Social were meaningful variables.

For the Upper Middle Income countries, the MLR was slightly more helpful. Both Freedom and Social registered as significant predictors, and the model was a slightly better with an R^2 value of 0.3953. Again the Lasso model found that only Freedom and Social were meaningful variables.

For High Income countries, the MLR found all predictors except for Generosity and Corruption to be significant, and the Lasso model found all predictors except for Generosity to be relevant.

Summary and Discussion:

From this analysis, we can draw several meaningful conclusions. First, GDP seems to be hugely significant when comparing the happiness level between rich and poor countries, but it was rarely significant at all within groups of countries of similar income. This suggests that there is are thresholds of wealth levels, which countries pass which improve their happiness substantially, but that slight changes in wealth don't make a big difference after that. This fits well within the intuition that wealth significantly increases happiness when raising people out of poverty, but the difference in happiness between a middle class and an upper middle class economy doesn't change people's lives in a meaningful way. Generosity was found to rarely ever be significant. Social connection, freedom, and health were found to have positive effects and be generally significant across wealth levels. Corruption was found to be hugely significant within wealthy countries. However, this could be because many countries with corrupt, extractive economies can increase their GDP quickly without improving the quality of life for many of their citizens, therefore acting as outliers within groups of wealthy countries.

Overall, though, the model seemed to have a much higher predictive power for wealthier countries than poorer countries, suggesting that there is still much room for significant research about the variables that make people happier within poorer countries. Another interesting area for research would be the effect of corruption on happiness levels within poorer countries, as this effect was mentioned briefly above as being puzzling.

Some significant limitations to the data include that less data was collected and is available for especially poor or small countries, or for countries ravaged by war. This report only had full data on 153 countries out of roughly 190 countries in the world. This means that we have limited understanding of many countries around the world, especially less developed countries. Another limitation is the aggregation of data points into the country level instead of the individual level. This means that we are unable to tell as much about what is truly making individual people happy instead of just what makes communities happy.

References:

Helliwell, John F., Richard Layard, Jeffrey Sachs, and Jan-Emmanuel De Neve, eds. 2020. World Happiness Report 2020. New York: Sustainable Development Solutions Network

"World Bank Country and Lending Groups." World Bank Country and Lending Groups, The World Bank, 2020, datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups.