ACT SCI 654: Regression and Time Series for Actuaries

Determinants of Country Happiness

Brandon Domash
Thomas Waltmans

University of Wisconsin-Madison
December 2018

Abstract

This study aims to find significant objective measures that contribute to a country's population overall happiness. Various objective explanatory variables are chosen that represent different departments of government, like economical measures, healthcare, technological advancement and equality in representation. Also the region of the world the country is located in is covered in the model to adjust for regional differences. After adjusting the data and the model for non-linearity and collinearity, the regression that was run found four variables that significantly explained changes in the Happiness Score. The variables GDP per capita, percentage of internet users and female representation in parliament significantly influence a country's happiness. The only region with significant different happiness scores is Latin America, which means that if a country with all other variables constant would be between five to six percentage points happier if it were located in Latin America. The other objective measures mentioned above, even though significant to the Happiness Score, had very low impact on it overall. Doubling the GDP per capita would result in a increase of less than five percent to the overall happiness. Hence, we found that it is very hard to explain the subjective measure of happiness with objective measures only and that it probably is a combination of both objective, subjective measures and the general attitude of the population.

Abstract word count: 220

Introduction

The purpose of this analysis is to investigate the different factors that make a country's citizens happy. Specifically, we want to investigate the understand which economic and social indicators have the strongest relationship with a country's happiness. In today's world, there is endless debate over what the focuses of government should be and what indicators are most important in to ensure the best outcomes for that country. For example, given limited resources, should a country's government try to stabilize their education sector or would it make more sense to improve healthcare systems in order to satisfy the needs of their citizens? We want to see which economic and social indicators directly relate to that country's happiness levels. By taking various economic and social indicators from many different areas, such as education, gender, health, inequality, trade, and work, we hope to uncover different patterns that can show us what makes a country successful in pleasing their citizens, and what might not be so important, and will find results that can tell us what the most important aspects of a government-lead society are in keeping citizens happy.

Most research about the topic includes various subjective measures to try to explain the happiness of the population of a country. The OECD Better Life Index, a well-known comparative happiness tool, tries to combine objective statistics, like GDP and Housing Expenditure, with various subjective statistics, like the level of Social Support and general Safety Sense, in order to calculate their own happiness rating for a country.[1] Various statistics that the OECD uses are from subjective polls, which can be very hard to compare internationally. For example, if you ask a European about safety they will probably measure it differently that someone from North-America. The standards are not similar or even comparable[2]. Some articles about these type of results argue that inhabitants of different countries look at the weight economic or social measures differently when it comes to happiness. For example, the Japanese generally worry about safety due to recent natural disasters like in Fukushima, while Americans would generally think of the work-life balance as their main indicator of happiness, while ironically ranking as one of the worst countries in regards to this topic[3]. This indicates that inhabitants of a certain country tend to look at their country's weak points when it comes to rating how happy they are and less at the factors the country is generally good at.

The data we use for the model comes from a variety of sources and is cross-sectional[A]. The dependent variable is each county's happiness score in 2017. The data is taken from the World Happiness Report in 2017, which "is an annual publication of the United Nations Sustainable Development Solutions Network which contains rankings of national happiness and analysis of the data from various perspectives."[4]The report ranks 156 countries by surveying approximately one thousand people in each country for three consecutive years. The survey asks the survey takers to rank their happiness on a scale of one to ten, with ten being the happiest. Each year, the thousand of surveys are combined with the prior two years' surveys and are then averaged, created a score between one and ten as well as a confidence interval. The data that we are most interested in is the single number that corresponds to a country's happiness. The explanatory variables are primarily numeric, and serve as economic and social well-being indicators for each country. The data comes from the United Nations Development Programme's Human Development Reports, which sorts development data into thirteen different dimensions. For each dimension, the United Nations reports around ten different indicators, each of which we considered in using for our model. The specifics of our data selection will be discussed in later sections. The only non-numeric data that we accumulated is each country's region. The seven regions of the world are generally agreed upon internationally, and our specific region-country data came from the World Bank.

As mentioned above, the point of this study is to find significant factors that relate to a country's overall happiness, as we want to objectify the subjective measure of happiness. Due to the nature that our dependent variable is continuous, we will look at simply using a multiple linear regression model, where the economic indicators and regional data will help explain a country's happiness. The reminder of the report will further discuss the data that we chose to use in our model and how we chose that specific data, along with a brief summary of each variable, including graphical summaries. We will then discuss our model selection in further depth, followed by an in-depth interpretation of our fitted model. Finally, we will discuss the implications of the model and our findings.

Data Characteristics

As mentioned above, the dependent variable in this research is the happiness rating of an individual country. Each country has a score between zero and ten. However, we wanted to widen the scale of this score, so we multiplied each country's score by ten, making the range from zero to one hundred. The data is approximately distributed normally, as seen in Figure 1. The mean happiness score is 53.57, which is approximately equal to the happiness of Jordan. The maximum happiness score is 75.37, belonging to Norway, while the lowest happiness score 26.93, belonging to the Central African Republic. Each of the top six countries in happiness score belong to the region 'Europe and Central Asia', whereas eight of the bottom nine countries in happiness belong to the 'Sub-Saharan Africa' region. Table 1 below groups each country by region, showing the average happiness of each region. From the table we can clearly see each region and their respective average happiness score. We see the highest regional happiness comes from North America, with a score of 71.55. However, this does not tell us much, as there are only two countries in the dataset that are classified as North America: the United States and Canada. We also see that Sub-Saharan Africa has the lowest happiness score, which is consistent with what many people would think before looking at the data. In addition, there are no outliers in the happiness score data.

Figure 1

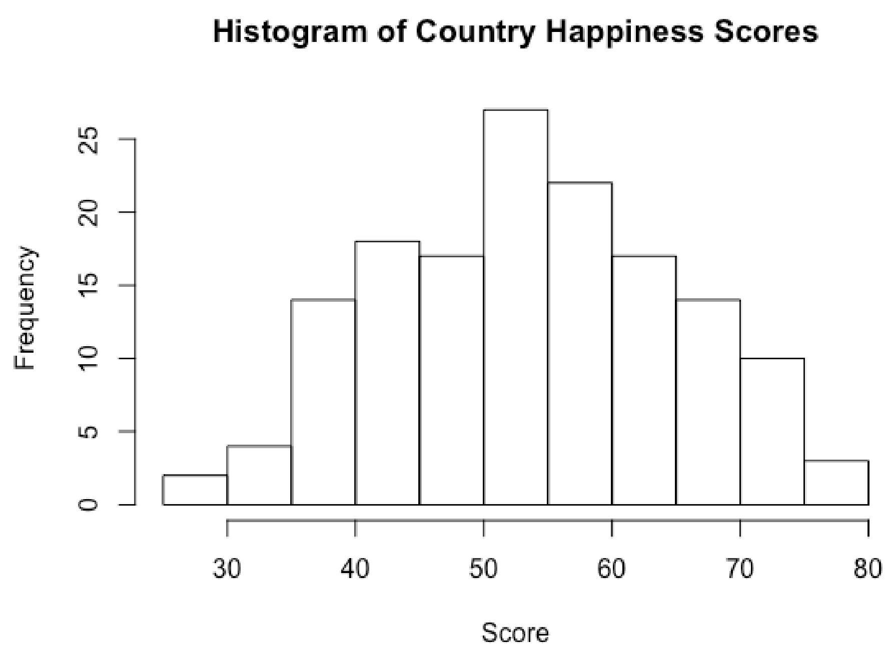


Table 1: Region Summary Statistics

Region	Number of Countries	Mean Happiness	Standard Deviation
East Asia & Pacific	**14**	**57.25**	**9.48**
Europe & Central Asia	**48**	**59.49**	**9.34**
Latin America & Caribbean	**22**	**59.58**	**7.51**
Middle East & North Africa	**18**	**54.60**	**10.37**
North America	**2**	**71.55**	**2.28**
South Asia	**7**	**46.28**	**5.00**
Sub-Saharan Africa	**37**	**40.82**	**5.86**
Total	**148**	**53.57**	**11.49**

In selecting explanatory variables, we wanted to select a variable from most dimensions of human development data in order to fully capture the different aspects of people's lives that impact their general happiness. However, for many of the economic and social indicators, there was substantial missing data. For example, income inequality is a very important economic indicator, but over twenty countries in our happiness dataset had missing values for this data, thus we chose to exclude it. Thus we looked for indicators in each dimension that had the most complete data and that we felt could have a substantial relationship with a population's happiness. We finally narrowed our selection to nine indicators of the human development each from a different dimension, and then merged this data with our happiness and region data.[B]

Each of the nine development indicator variables were continuous, and the only discrete variable is the country's region. As displayed in Table 1, there are seven world regions defined by the World Bank[5]. The nine explanatory variables that we use are:

- GDP per capita (2011 PPP USD)
- Amount of Internet Users (% of population)
- Life Expectancy at Birth (years)
- Mean Years of Schooling
- Infant Mortality Rate (per 1,000 births)
- Percentage of Population using Improved Drinking Water Sources
- The Share of Seats in Parliament held by Women (%)
- Suicide Rate (per 100,000 people)

- Unemployment Rate

As stated above, we want to look at indicators from as many different dimensions of life as possible. From these variables, we hope to capture the income composition, technological advances, health systems, the education, access to clean water, gender neutrality, human security and employment levels. The summary statistics for each variable are displayed below in Table 2.

Table 2: Summary Statistics of the Variables

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Happiness.Score	148	53.57	11.49	26.9	44.92	61.25	75.40
gdp	147	19,453	20,096	661	4,505	28,350	116,936
internet.users	147	51.405	28.231	4	25.25	75.70	98.20
life.expectancy	148	72.448	7.806	52	67.10	77.8	84
schooling.years	148	8.711	3.198	1.5	6.30	11.50	14.10
mortality.rate	148	21.835	20.654	1.6	5.05	34.375	88.50
improved.drinking	147	86.641	17.227	38.9	76.50	99.60	100.00
women.parliament	148	22.965	10.933	0.5	15.70	30.20	55.70
suicide.rate	148	16.144	9.537	2.500	9.350	19.975	58.800
unemployment	148	7.547	5.671	0.200	4.075	9.550	27.700

In this table we also see that the number of individual observations, hence the number of countries, is 148 for the majority of variables. While the world happiness report has data for 156 countries, eight countries had minimal data for any of the explanatory variables, and were thus thrown out from the data. Our set of countries still accounts for 76% the world's countries, which we felt was a large enough sample to carry out the analysis. Although some explanatory variables have one missing observations, we still chose to include them in our dataset.

While our explanatory variables capture many different dimensions of life, they are also highly correlated. This makes sense, as countries with higher amounts of GDP per capita will also likely have high levels of life expectancy, internet users, and schooling years. Only three variables lack significant collinearity as seen in Figure 2. These variables are Unemployment Rate (unemp), Suicide Rate (scd.), and Share of Women in Parliament (wmn.). This is an issue that we considered when selecting our model, and will be further discussed in that section. In addition, those three variables are the only variables that have a significant p-value, indicating that they are independent from the variables in which their p-value is greater than .05.

Figure 2: Correlation Matrix of Explanatory Variables

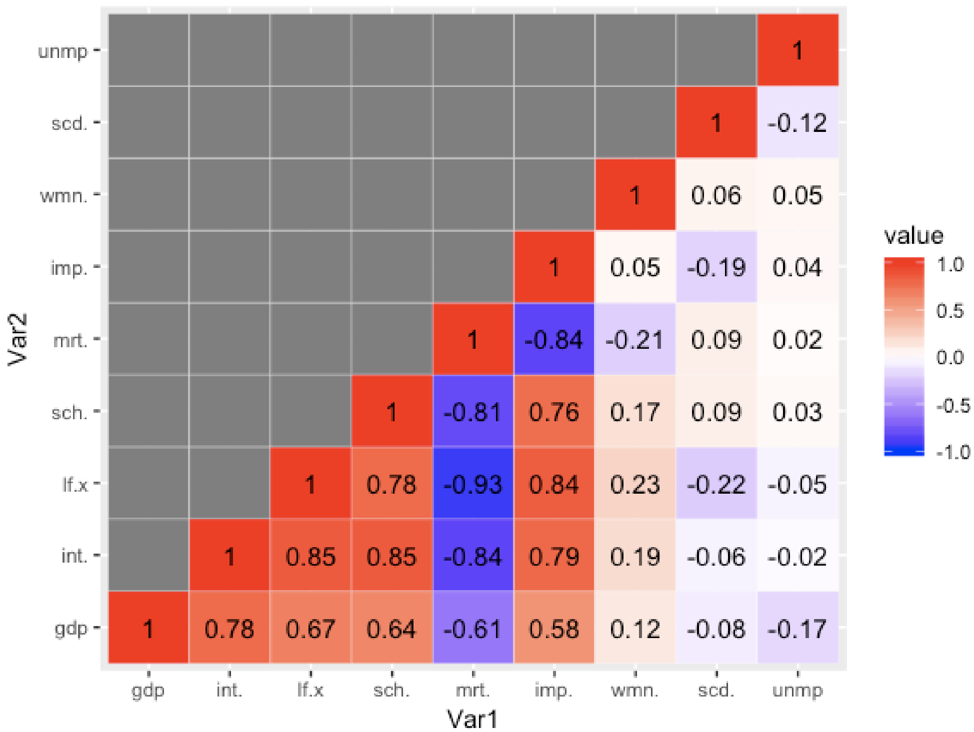
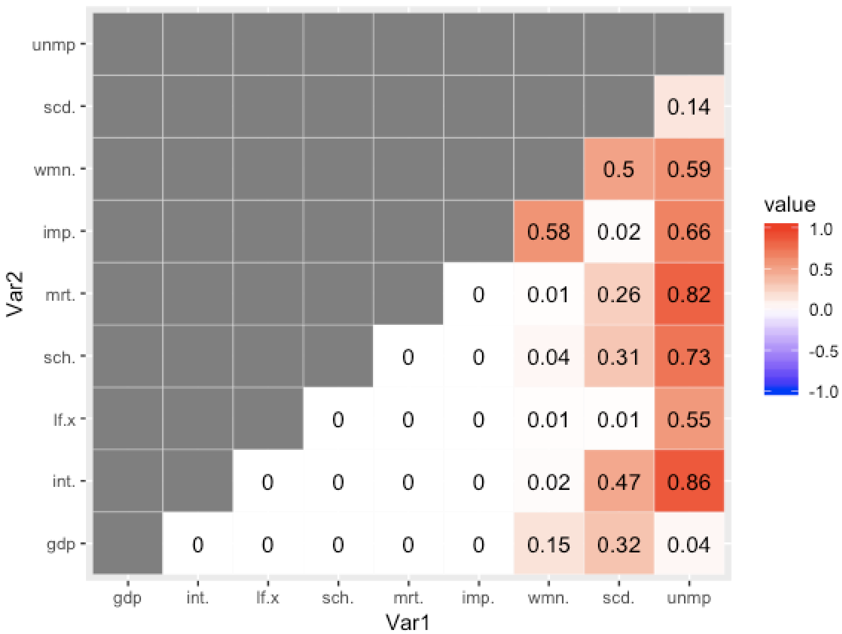


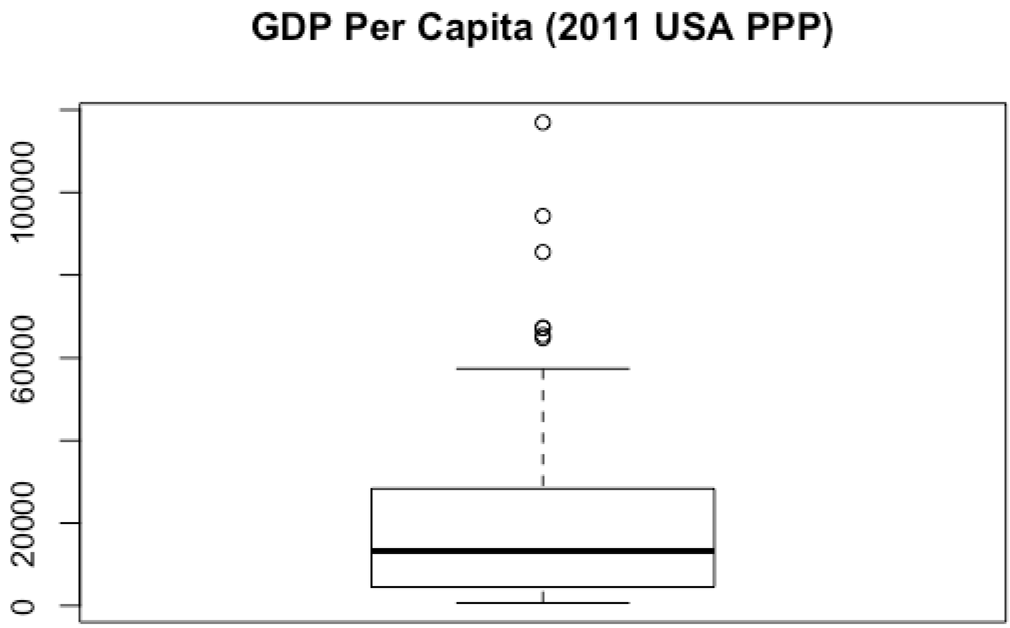
Figure 3: Correlation Matrix P-Values



unmp
scd.
wmn.
imp.
mrt.
sch.
lf.x
int.
gdp
Var2
0.14
0.5
0.59
0.58
0.02
0.66
0
0.01
0.26
0.82
0
0
0.04
0.31
0.73
0
0
0
0.01
0.01
0.55
0
0
0
0
0.02
0.47
0.86
0
0
0
0
0
0.15
0.32
0.04
gdp
int.
lf.x
sch.
mrt.
imp.
wmn.
scd.
unmp
Var1
value
1.0
0.5
0.0
-0.5
-1.0

Some of the explanatory variables contain outliers. Interestingly, GDP per Capita contains seven outliers, most of which are small countries, including Qatar, Luxembourg, Kuwait, United Arab Emirates and Singapore. The number of outliers greatly skews the data, as we will see in the model selection section. Infant mortality rate, improved drinking source percentage, percentage of women in parliament, suicide rate, and unemployment rate also all include outliers, which is something that we definitely kept in mind before constructing our model. We can also infer from this that there are many countries that are either very far ahead or very far behind most of the world in these social and economic indicators. For example, Figure 4 clearly shows that GDP per capita is fairly normally distributed aside from five countries, who are very far ahead in that statistic. This is the case for many of the explanatory variables in our data set.

Figure 4



Model Selection and Interpretation

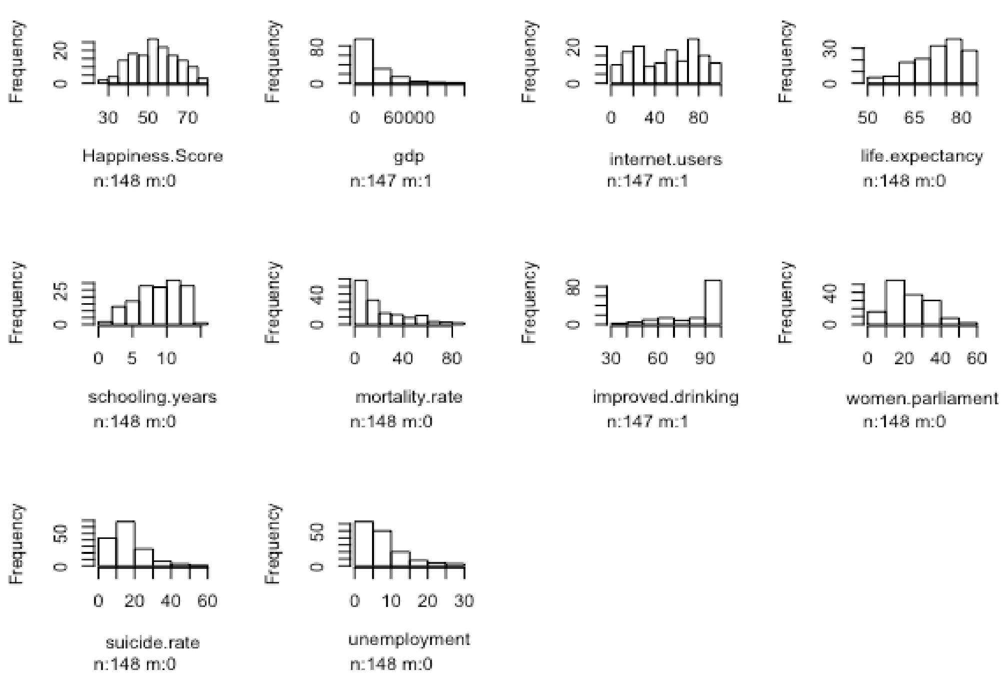
Now that the characteristics of the data are reflected upon, a model can be built to predict the dependent variable using the explanatory variables. Because the nature of our data is cross-sectional, and our explanatory variable is continuous, we will use a multiple linear regression model to explain happiness using the explanatory variables detailed above[C]. The initial model will be as follow:

$$\begin{aligned}\text{Happiness Score} \quad = \quad & \beta_0 + \beta_1 \text{ GDP} + \beta_2 \text{ Internet Users} + \beta_3 \text{ Life Expectancy} \\ & + \beta_4 \text{ Schooling Years} + \beta_5 \text{ Mortality Rate} + \beta_6 \text{ Improved Drinking} \\ & + \beta_7 \text{ Women in Parliament} + \beta_8 \text{ Suicide Rate} + \beta_9 \text{ Unemployment Rate} \\ & + \beta_{10} \text{ North America} + \beta_{11} \text{ Latin America} + \beta_{12} \text{ Europe} + \beta_{13} \text{ Middle East} \\ & + \beta_{14} \text{ Sub-Saharan Africa} + \varepsilon\end{aligned}$$

where ‘East Asia and the Pacific’ is used as a base region.

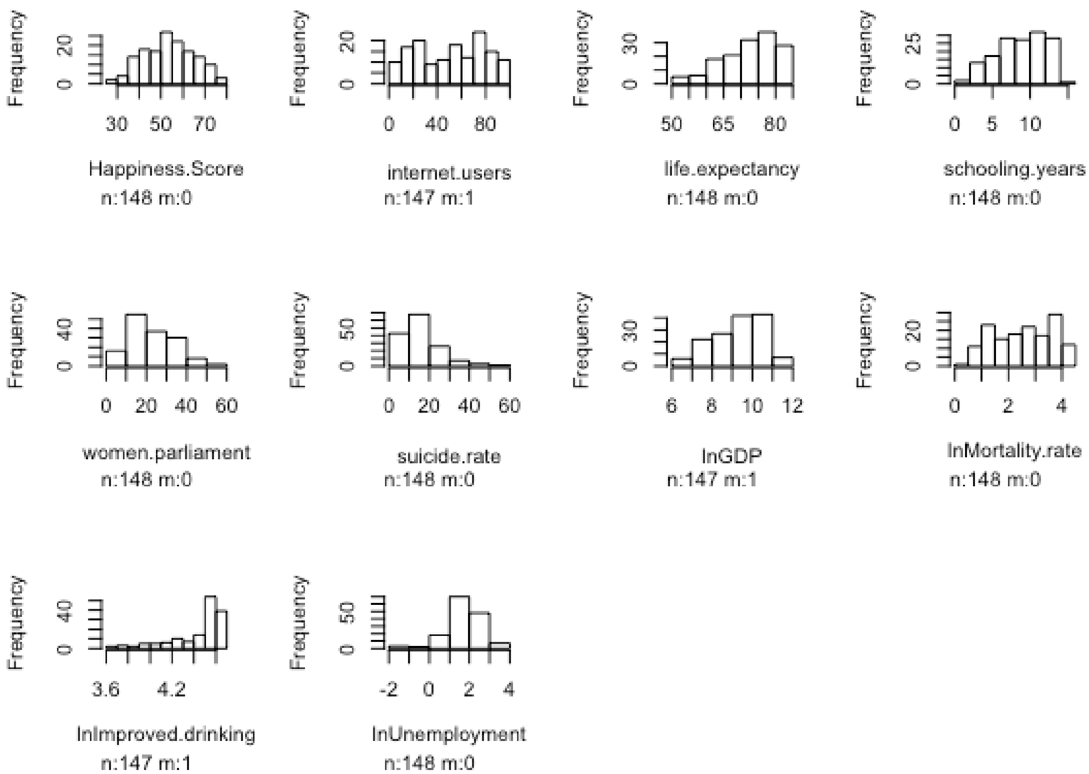
As discussed in the data section, we saw that many variables had numerous outliers. Thus, the first thing we looked for is whether the variables express non-linearity, such as having a skewed distribution. This could negatively influence the regression results as the assumption of normality has been violated, thus those variables should therefore be adjusted.

Figure 5: Distributions of the Variables



From figure 5 it is clear to see that the variables GDP, mortality rate, improved drinking and unemployment have skewed distributions. To adjust for the skewness, the logarithmic levels of these variables are used instead if their normal levels.

Figure 6: Distributions of the Variables with Logarithmic Adjustments



As seen in Figure 6, the skewness in GDP, mortality rate, and unemployment has been fairly adjusted. The logarithmic level of improved drinking is still skewed, but clearly less than its initial data.

After looking at the distribution of the explanatory variables, the correlations of the transformed variables are once again examined.

Table 3: Correlation of the Variables

	Happiness. Score	internet. users	life. expectancy	schooling. years	women. parliament	suicide. rate	ln GDP	lnMortality. rate	lnImproved. drinking	ln Unemployment
Happiness.Score	1	0.810	0.801	0.720	0.287	-0.126	0.831	-0.767	0.705	-0.030
internet.users	0.810	1	0.848	0.851	0.193	-0.061	0.903	-0.892	0.754	0.041
life.expectancy	0.801	0.848	1	0.779	0.230	-0.218	0.850	-0.904	0.805	0.049
schooling.years	0.720	0.851	0.779	1	0.170	0.085	0.822	-0.846	0.729	0.136
women.parliament	0.287	0.193	0.230	0.170	1	0.057	0.138	-0.230	0.027	0.054
suicide.rate	-0.126	-0.061	-0.218	0.085	0.057	1	-0.082	-0.015	-0.194	-0.061
lnGDP	0.831	0.903	0.850	0.822	0.138	-0.082	1	-0.858	0.806	0.057
lnMortality.rate	-0.767	-0.892	-0.904	-0.846	-0.230	-0.015	-0.858	1	-0.733	-0.041
lnImproved.drinking	0.705	0.754	0.805	0.729	0.027	-0.194	0.806	-0.733	1	0.131
lnUnemployment	-0.030	0.041	0.049	0.136	0.054	-0.061	0.057	-0.041	0.131	1

Once again we see that many variables are highly correlated with each other, which makes sense, since a country that is thriving economically would probably also have more internet users, better healthcare, and more widely available education. This indicates that the initial model will contain high amounts of collinearity, which will increase standard errors and make it more difficult to find significant relationships between the explanatory variables and happiness scores. We then calculated Variance Inflation Factor (VIF)[D] scores for each explanatory variable to quantitatively measure collinearity. While many variables have high VIF scores, when looking at the collinearity of the variable lnMortality.rate, we find that the VIF is over thirteen. When the VIF is over ten, it indicates severe collinearity. Because of the high VIF and the presence of other healthcare related variables, the mortality rate is left out of the final model. After removing this variable the VIF of the other variables decreased significantly, leaving no variables with a VIF greater than ten. Thus we decided to leave the remaining variables in the final model. The final model looks like this:

$$\begin{aligned}\text{Happiness Score} \quad = \quad & \beta_0 + \beta_1 \ln(\text{GDP}) + \beta_2 \text{ Internet Users} + \beta_3 \text{ Life Expectancy} \\ & + \beta_4 \text{ Schooling Years} + \beta_5 \ln(\text{Improved Drinking}) \\ & + \beta_6 \text{ Women in Parliament} + \beta_7 \text{ Suicide Rate} + \beta_8 \ln(\text{Unemployment Rate}) \\ & + \beta_9 \text{ North America} + \beta_{10} \text{ Latin America} + \beta_{11} \text{ Europe} + \beta_{12} \text{ Middle East} \\ & + \beta_{13} \text{ Sub-Saharan Africa} + \varepsilon\end{aligned}$$

Results

Now that the final model has been decided upon, it is time to run a regression on it and see if it has statistical significance and if our explanatory variables are good estimators of the dependent variable.

Table 4: Regression Summary

	Estimate	Std. Error	t value	Pr(> | t |)	
(Intercept)	4.151	(18.524)	0.224	0.823	
lnGDP	4.638	(1.032)	4.495	1.52e-05	***
internet.users	0.096	(0.045)	0.130	0.036	*
life.expectancy	0.130	(0.154)	0.844	0.401	
schooling.years	0.142	(0.346)	0.411	0.682	
lnImproved.drinking	-2.172	(4.319)	-0.503	0.616	
women.parliament	0.137	(0.047)	2.918	0.005	**
suicide.rate	-0.060	(0.057)	-1.041	0.300	
lnUnemployment	-1.116	(0.582)	-1.917	0.058	.
RegionEurope	-0.664	1.852	-0.359	0.721	
RegionLatin America	5.731	1.972	2.906	0.005	**
RegionMiddle East	-1.145	2.117	-0.541	0.590	
RegionNorth America	6.421	(4.220)	1.522	0.131	
RegionSouth Asia	0.560	(2.751)	0.204	0.839	
RegionSub-Saharan Africa	-1.457	(2.495)	-0.584	0.561	
– – –					
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					
Residual standard error: 5.406 on 130 degrees of freedom					
(3 observations deleted due to missingness)					
Multiple R^2	0.797				
Adjusted R^2	0.775				
Residual Std. Error	5.406 (df = 130)				
F Statistic	36.36 on 14 and 130 DF				
p-value	< 2.2e-16				

Happiness_hat = 4.15 + 4.64* ln(GDP) + .10 Internet Users + .13 * Life Expectancy + .142 * Schooling Years - 2.17 * ln(Improved Drinking) + .14 * Women in Parliament - .06 * Suicide Rate - 1.11* ln(Unemployment Rate) + 6.42 * North America + 5.73 * Latin America - .66 * Europe - 1.15 * Middle East - 1.46 * Sub-Saharan Africa

From the regression results only four significant explanatory variables are found at a 5% significance level. These are the GDP per capita, the percentage of internet users, the share of seats in parliament held by women and the region Latin America and the Caribbean. For GDP per capita, this means that a 100% increase would result in a 4.6 points increase in the Happiness Score (out of 100 total points), holding each other variable constant. For internet users and seats in parliament held by women, a ten percentage points increase would correspond to a happiness score increase of approximately 1.0 and 1.4 respectively, holding the other regressor constant. Finally, a country in Latin America with everything else held constant would be 5.7 points happier than a country in East Asia and the Pacific. Moreover, the adjusted R-square of the model is relatively close to one, which indicates that our model is a good fit of the happiness score.

The results of this model are not that surprising. It is likely that collinearity was a large factor so many of the variables were unable to be rejected. Even though no variable in our final model had a VIF score over 10, there were still a group of highly correlated explanatory variables, which makes it harder to distinguish one variable's effect on happiness from another. For example, mean years of schooling and life expectancy have a .78 (Table 3) correlation coefficient, meaning they have a strong positive relationship. Thus, how can we tell if it is mean years of schooling or life expectancy that is actually impacting a country's happiness. This is the case for many variables, which is why there are large standard errors in our regression results and many variables that are deemed insignificant. Running an F-Test with all of the insignificant explanatory variables tells us that these variables are jointly significant, thus we are fine leaving them in the model as they still help to explain a country's happiness score.

When analyzing the residual plot (Figure 7), we see that the residuals are fairly evenly spread out, and thus do not demonstrate a non-linear relationship in our model. The fairly equal distribution also indicates the model is homoscedastic, which is one of the assumptions of linear regression.

Figure 7: Residuals vs Fitted Plot

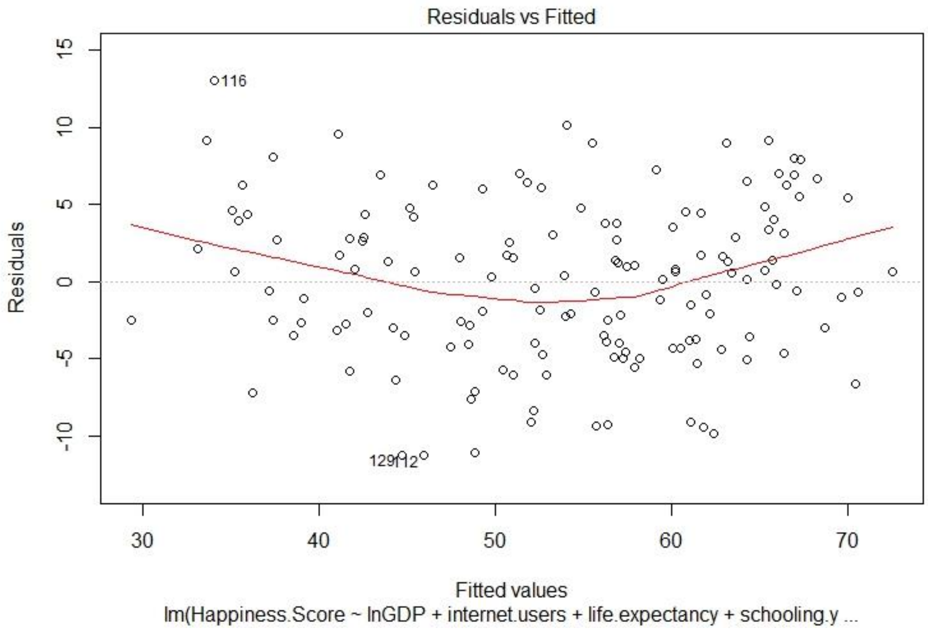


Figure 8: Normal Q-Q Plot

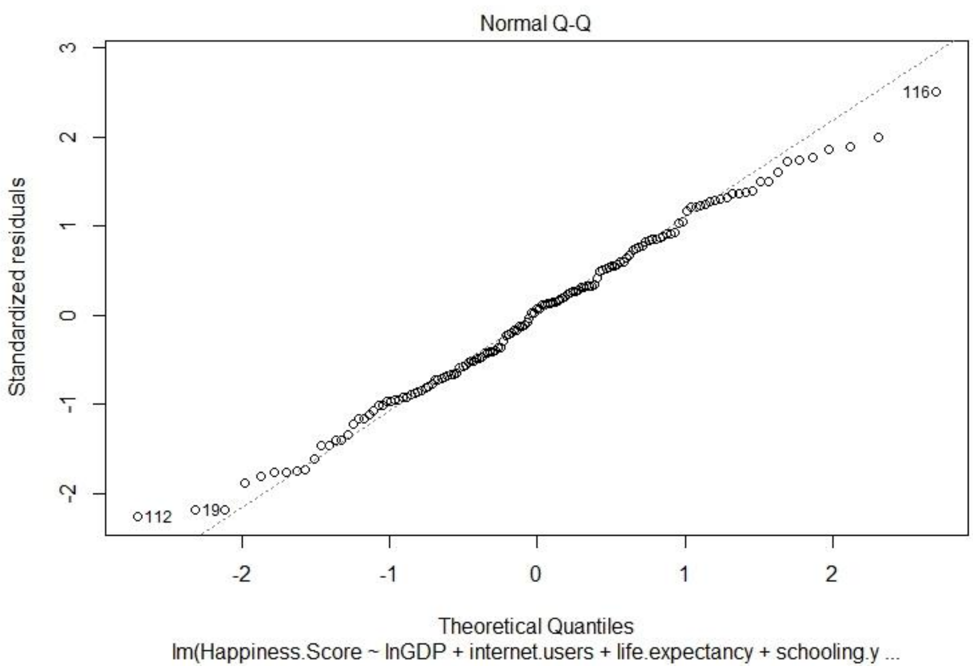
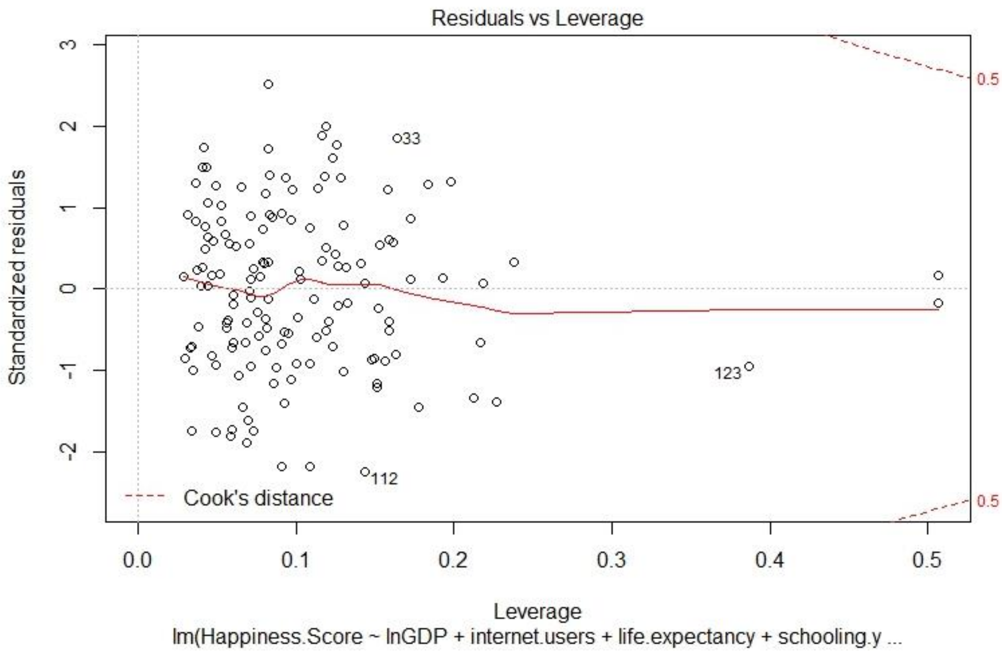


Figure 8 displays that the residuals come from a normal distributed, which is also one of the assumptions of linear regression. They line up reasonably in a straight line, indicating that the data could have plausibly come from a normal distribution.

Figure 9: Residuals vs Leverage Plot



Finally, in our last plot, the Residuals vs Leverage Plot, there are no extreme data points which are very influential to the regression line, since there are none outside the Cook's distance. If there were highly influential points, the regression results would be altered based on their inclusion, as these points would have high amounts of leverage on the regression equation. In summation of these these plots, the regression analysis shows no errors or signs of a bad model, and we can feel confident that our linear regression model fits the data well and that the assumptions of linear regression are satisfied.

Summary and Concluding Remarks

To summarize, the main objective measures found that the significant influences of happiness are the GDP per capita, the percentage of internet users and women in parliament and the Latin America dummy variable. All of these variables positively influence the Happiness Score, but only very marginally. If the GDP per capita of a would be doubled, which in theory is possible but in practice could take decades, the Happiness Score would increase by less than five points. For the percentage of internet users and women in parliament, each ten percentage point increase would result in only a single percentage point increase in the Happiness Score. Overall, one could question whether a government should go out of their way and focus on these factors to increase overall happiness, if it only makes such a marginal difference. Besides, even the small changes mentioned would take years before being achieved. Furthermore, for obvious reasons a country cannot focus on the region they are located in and cannot choose to move to Latin America. One explanation for their significantly higher happiness is that they have a more positive outlook on life which would make it more of a subjective measure than anything else.

A potential problem in the results would be that there is still very high collinearity between variables. An explanation for this is that when a country does better in a certain objective measure, it would automatically come with a positive development in another area. For example, if more people have internet connections, the country is more integrated in modern society, which would then result in them increasing their GDP per capita more easily, increasing the level and availability of education. In research of multiple aspects of a country as a whole, it is nearly impossible to completely discard the collinearity. However, this does result in less reliable models with variables that are harder to find significance.

Also, the dataset used does not include every single country in the world. There are still over forty countries on which there was no data on the Happiness Score. In general these are very small countries which have very small populations to get data from in the first place. However, including these countries might make a difference in our overall model and could result in different conclusions.

Some objective variables that in theory are probably very important to a countries happiness are very hard to include. One of such variables is the type of healthcare system in place in a country. A significant amount of the top ranking happiest countries are known for having universal free healthcare. One could include a categorical variable for each healthcare system, but it is very hard to specify and compare the different systems in place all over the world. Even not every universal free healthcare system is exactly the same, hence it would be nearly impossible, as there are many different complexities that make each country's healthcare systems unique. Another variable that we discussed including was whether or not a country had been at war. However, once again this is hard to define exactly what it means to be at war. There are some countries where this is obvious, but others where this is ambiguous. For example, there had been a genocide against the Rohingya ethnic group in Myanmar, but does this get classified as war[6]? Because of these complexities, we decided not to include either variable in the model.

Finally, there are also some missing data points for some of the variables. The data for these variables in these particular countries is not known publicly, hence is not included. Since there are only a few of these missing data points, it is safe to assume they would not significantly change the results of the model.

In conclusion, it is very hard to objectify happiness, Even the significant results found make very little difference on the actual Happiness Score. There may not be one solution to the happiness of each country, but rather a combination of multiple different objective and subjective factors, with some factors very hard to include in a proper statistical analysis. However, we do can confidently say increasing the GDP per capita of a nation will influence the happiness of a country, and raising the GDP per capita will very likely also improve other economic and social indicators of a country, resulting in happier citizens.

References

1. How's life? (n.d.). Retrieved from http://www.oecdbetterlifeindex.org/
2. Author, N. (2015, June 01). The American-Western European Values Gap. Retrieved from http://www.pewglobal.org/2011/11/17/the-american-western-european-values-gap
3. Southern California Public Radio. (2014, September 26). What factors contribute to a nation's happiness? Retrieved from https://www.scpr.org/news/2014/08/11/45917/what-factors-contribute-to-a-nation-s-happiness/
4. World Happiness Report. (n.d.). Overview. Retrieved from http://worldhappiness.report/
5. Countries and Economies. (n.d.). Retrieved from https://data.worldbank.org/country
6. The Plight of the Rohingya. (n.d.). Retrieved from https://www.ushmm.org/confront-genocide/cases/burma/introduction/the-plight-of-the-rohingya

Appendix

A. Cross sectional data is data that observes many subjects at a given point of time. The subjects in this case are the countries, each of which have happiness scores and economic variables that were observed at a different time. Other types of data, such as time-series and panel data differ from this, and would require other methods besides our multiple linear regression model, such as autoregressive or moving average models for time series data. However, because the nature of our data is cross sectional, we do not need to worry about these types of models.

B. Merging the data actually proved to be quite difficult. For starters, the country names in for each dataset were slightly different. This meant cleaning the data to match the country names in each dataset to the country names in the happiness data was necessary. There was also many times in which there was data for countries not included in the happiness dataset, and no data for some countries within the dataset, creating lots of frustration. In the end, we were able to merge all of the data successfully and find datasets in which nearly all of the data overlapped, but it required lots of tedious tasks.

C. Because happiness is a continuous variable, we do not need to worry about the entire class of regression models relating to binary dependant variables, such as the Logit or Probit model. While these models predict a probability given levels of the explanatory variables, we only care about predicting an actual happiness score, and thus we do not need to take these models into consideration.

D. Variance Inflation Factor (VIF) directly proportional to the square of the correlation coefficient between one explanatory variable, x_j, and the combination of the remaining explanatory variables. Thus, to find VIF for each variable, we regressed each of our explanatory variables on all of the remaining explanatory variables to get an R^2_j in which we used to calculate a VIF_j for each variable. VIF_j is defined as $1/(1-R^2_j)$. We then did the same thing after removing the variable with a VIF greater than ten.

Statements:

- This work is original and we completed it on our own
- This report can be shown to future students as examples

Word count: 4989