Lab 2: Does a country's economic strength dictate the happiness of its citizens and are there ways to maximize their happiness?

A causal regression analysis on the World Happiness Report

w203 Team Opal: Ryan Wong, Khakali Olenja, Chris Grimes

Contents

1 Introduction & Background

For all of the authority and power that is bestowed upon it, a nation's government ultimately exists to serve its citizens. With that authority comes the ability to make investments and implement policies that have far reaching consequences on its people. A country's economic indicators are often used as a measurement for its prosperity. But one could argue the happiness of the average citizen is the best indicator for national prosperity, and thus should be the primary outcome that governments should seek to optimize. After all, a dissatisfied nation often leads to instability and political turmoil, but prosperity brings peace, wealth, and national stability.

In order to maximize a nation's happiness, we researched what role a country's economic strength plays on an individual happiness and if there are any other influential characteristics. The objective of this publication will be to offer guidance on whether wealth translates to happiness and offer recommendations to governments and policy makers about which factors to consider in order to improve the overall happiness of their citizens.

2 Research Question and Hypotheses

Our analysis will be guided by a single research question that we hope to evaluate with this report:

> *Does a country's economic strength dictate the happiness of their citizens, and are there other factors that can be used to maximize it?*

3 Data Source

Since 2012, the United Nations has published the World Happiness Report, an annual report that measures and ranks countries based on the happiness of their citizens. This report is a highly valuable resource that provides an aggregated sense of the impact of wealth and other factors on happiness from a large number of individuals across the globe. Data in the report is derived from the Gallup World Poll which has been conducting life evaluation surveys across the world each year since 2005. The happiness score for a country is calculated based on six factors: economic production, social support, life expectancy, freedom, absence of corruption, and generosity.

We chose to use the publicly available data set "Data for Figure 2.1" from the webpage for the World Happiness Report in 2021, which is available at the following URL: https://worldhappiness.report/ed/2021/. This data set offered data on several countries across several years regarding happiness, the logarithm of national GDP per capita, and several other variables of interest in the report. These variables would help form the explanatory models described in a later section.

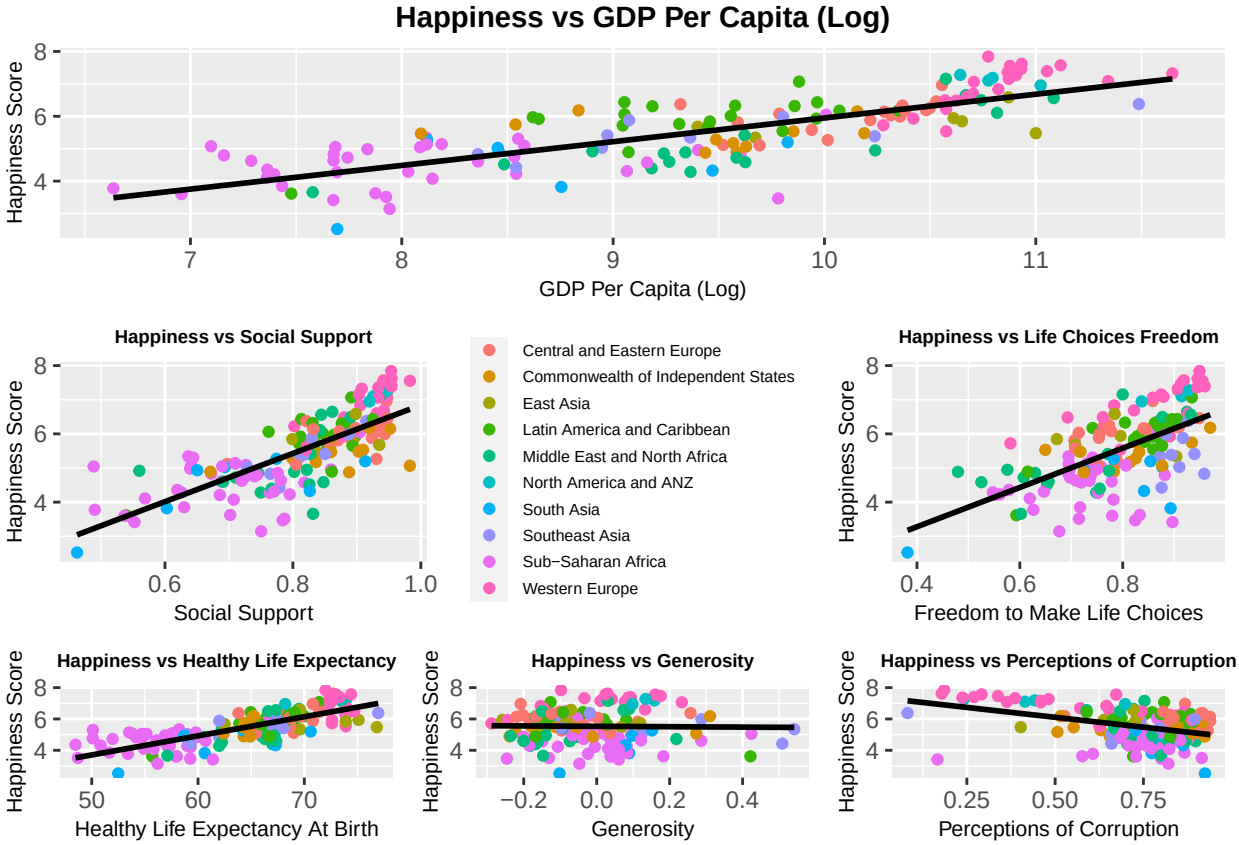
4 Variables of Interest

The World Happiness Report data holds numerical and character string data for each of our variables of interest. Each variable is given and described below:

- life_ladder: an ordinal score from 0 to 10 representing the country's general happiness level.
- log_gdp_per_capita: a numerical value of the logarithm of the country's GDP per capita. representing the average wealth of a person per country.
- life_expectancy: the life expectancy of a citizen of the country based on the World Health Organization's Global Health Observatory data repository.
- social_support: an ordinal score from 0 to 10 representing the amount of social support within a country.
- freedom_to_make_life_choices: an ordinal score from 0 to 10 representing the amount of freedom and self-determination within a country.
- generosity: an ordinal score from 0 to 10 representing the amount of generosity within a country.
- perceptions_of_corruption: an ordinal score from 0 to 10 representing the amount of corruption present within a country.
- geographic_region: categorical data that shows what geographic region of the world the country is in. The regions represented are as follows: North America and ANZ, Central and Eastern Europe, East Asia, Middle East and North Africa, South Asian, Sub-Saharan Africa, Commonwealth of Independent States, Latin America and Caribbean, Southeast Asia, and Western Europe.

It should be noted that while the variables that are in ordinal scale are represented from 0 to 10, the variables themselves are represented in continuous decimal values. This is because the values represent the mean for that variable from the entire country's respondents.

As part of our initial data exploration, a series of plots were created to visually show how each of the variables of interest were mapped to happiness. These plots are shown below.



5 Research Design and Hypotheses

With the data chosen, we needed to apply it to our research question in order to determine determine whether or not a country's wealth was influential in its citizen's happiness. We decided that the best way to achieve this was by conducting an explanatory data analysis using coefficient T-tests on a series of linear models would be created from the World Happiness Report data. Being held to a 95% confidence level, the coefficient T-tests will provide evidence towards whether or not wealth has a statistically significant impact on happiness, as well as provide further insight regarding the statistical significance of other variables in the data.

As we address our research question, we will compare our findings against the null and alternative hypotheses, listed below:

- Null Hypothesis (μ_0): Wealth does not influence happiness at all. In other words, $\mu_0 : \mu = 0$.
- Alternate Hypothesis (μ_A): Wealth indeed influences happiness. In other words, $\mu_A : \mu \neq 0$.

6 Data Engineering

The data set from the World Happiness Report was imported from its CSV form into a usable data frame format within R. From there, a new data set named "whr_data_transformed" was created that copied the original data. Using the new data set, the following changes and transformations were made in the following order:

1. Either due to an encoding error or an unusual intentional inclusion, the data set had a strange character (ï) included in the field name for the country's name, which was renamed to be country_name within R. This change was cosmetic and did not affect the data in any way, but it did improve the usability of the data set.
2. In order to account for geographic regions for each country, the data set originally had a categorical variable named "regional_indicator" that stated the country's general geographic region. In order to include that categorical information in the regression, several new indicator fields were created in the data set as control variables for these geographic regions. If a country in the data set is part of a particular geographic region, it will be represented in the regression by the indicator field that was created for it.
3. The original data set included historical data from each country across several years, though not every country is represented consistently throughout the years. Our analysis could not utilize this historical data since it would violate the independence assumption between variables; a country's current circumstances are highly dependent on its circumstances in the previous year. Furthermore, we could not include country data from different years, as that would introduce variability due to the differences in time and external variables that are are not accounted for. As such, we chose to only use country data from 2021, as that provided both the largest sample of countries and the most recent data while also reducing the amount of unnecessary variability in our data.

The end result was a final data set of 149 sampled countries with 22 fields each. The data was entirely consistent and free of errors and NA values.

7 Model Design

In order to evaluate the influence of wealth on individuals happiness, we would be conducting a co-efficient T-Test for a series of linear models from the data. The aim of these tests was to determine whether or not happiness was a statistically significant variable in each model. Based on the variables that were available from the data set, it was decided that multiple linear models of increasing sophistication would be used in order to observe how the variability changed per model. The three models that were created for the analysis are described below:

- model_1 estimates happiness with the logarithm of GDP per capita,
- model_2 estimates happiness with the logarithm of GDP per capita, social support, life expectancy, freedom to make life choices, generosity, and perceptions of corruption,
- model_3 estimates happiness with the logarithm of GDP per capita, social support, life expectancy, freedom to make life choices, generosity, perceptions of corruption, and the geographic area of their country.

It should be noted that the coefficient T-test requires continuous, non-ordinal values, but many of the variables from the data set in question (i.e. social_support, freedom_to_make_life_choices, generosity, and perceptions_of_corruption) are non-continuous ordinal values. However, the values given are real values constrained within a particular range, making their values effectively continuous within that particular range. Therefore, the variables in question are justifiably continuous-enough to be considered valid for the coefficient T-test.

8 Model Results

Our primary research question can be distilled to what role does wealth play in an individual's happiness and are there other factors governments should consider. Below we have highlighted our findings for each model. It should be noted that at the beginning of our research there was no evidence that presupposed that GDP is relevant. We will use the P Values to determine if the results are statistically significant. The coefficient T-Test was run for each of the models that were created.

Below is the results for the coefficient T-Test on model_1.

Below is the results for the coefficient T-Test on model_2.

Below is the results for the coefficient T-Test on model_3. Do note that model_3 assumes that the normative geographic region is for a citizen living in North America or ANZ.

Below is a table that summarizes and compares the effect sizes and error values of each model using Stargazer. Note how the effect sizes of log(GDP) on happiness decreases and the R^2 values increases as additional control variables are added.

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Sun, Apr 10, 2022 - 12:44:02 pm

Table 1: Estimates of the Effect of log(GDP per Capita) on Happiness

	Dependent variable:		
	Happiness		
	Model 1	Model 2	Model 3
	(1)	(2)	(3)
log_gdp_per_capita	0.732***	0.280***	0.268***
	(0.047)	(0.087)	(0.087)
social_support		2.476***	1.949***
		(0.668)	(0.654)
healthy_life_expectancy_at_birth		0.030**	0.014
		(0.013)	(0.015)
freedom_to_make_life_choices		2.011***	2.267***
		(0.495)	(0.501)
generosity		0.365	0.497
		(0.321)	(0.318)
perceptions_of_corruption		−0.604**	−0.328
		(0.290)	(0.310)
central_and_eastern_europe			−0.350
			(0.305)
east_asia			−0.569*
			(0.334)
middle_east_and_north_africa			−0.629**
			(0.308)
south_asia			−1.079***
			(0.357)
sub_saharan_africa			−0.656*
			(0.347)
commonwealth_of_independent_states			−0.710**
			(0.317)
latin_america_and_caribbean			−0.221
			(0.308)
southeast_asia			−0.981***
			(0.325)
western_europe			−0.014
			(0.274)
Constant	−1.372***	−2.239***	−0.550
	(0.446)	(0.630)	(1.048)
Observations	149	149	149
R^2	0.624	0.756	0.809
Adjusted R^2	0.621	0.746	0.787
Residual Std. Error	0.661 (df = 147)	0.542 (df = 142)	0.495 (df = 133)
F Statistic	243.647*** (df = 1; 147)	73.264*** (df = 6; 142)	37.489*** (df = 15; 133)

Note: *p<0.1; **p<0.05; ***p<0.01

In Model 1, it is evident that the logarithm of GDP is highly statistically significant as we have a p-value that is practically zero (p-value = $2.2 \cdot 10^{-16}$). This is an indication that we should strongly reject the null hypothesis, which would state that GDP has no impact on Happiness. As such, this model suggests that for every 10% increase in wealth, happiness increases by 0.73 points.

In Model 2, it is evident that three of the five variables are statistically significant including the logarithm of GDP, which has a p-value of 0.0039. Social support and Freedom to make life choices were similarly statistically significant, with p-values of 0.0022 and 0.0001 respectively. Because of the low p-value for the logarithm of GDP variable, we have enough evidence to reject the null hypothesis. This model suggests that for every 10% increase in wealth, happiness increases by 0.28 points.

In Model 3, it is evident that nine of the fifteen variables are statistically significant including the logarithm of GDP, which has a p-value of 0.011. Social support, Freedom to make life choices, and the regional indicators for East Asia, Middle East and North Africa, South Asia, Sub-Saharan Africa, Commonwealth of Independent States, and South East Asia were similarly statistically significant, with p-values ranging from 0.042 to practically zero. Because of the low p-value for the logarithm of GDP variable, we have enough evidence to reject the null hypothesis. This model suggests that for every 10% increase in wealth, happiness increases by 0.27 points.

9 Limitations

9.1 Statistical limitations of our model

In order for the model and regression to be considered valid, we must first meet the large-sample model requirements: the data is Independent and Identically Distributed, and that a unique Best Linear Predictor exists.

The first assumption in a large-sample assumption is that the data is IID. Given that the data set is investigating countries across the globe, it is expected that countries will be interacting with each other through commerce, politics, and cultural exchange. This may lead to countries with close geographic or historical ties influencing each other in ways that may be reflected in the data: a region's wealth, culture, and prosperity will ultimately influence the wealth, social characteristics, and happiness of the countries therein. However, that it is impossible to have countries that do not influence or trade with each other, as even pariah states such as North Korea still influences and is influenced by its neighbors. We must admit that having a truly IID data set would be impossible for this particular data set, and will simply advise that we adjust our measures of uncertainty to expect slightly larger standard error estimates due to the clustered nature of the country data. As such, we can assume that the countries and variables in question are independent and identically distributed, thus satisfying the IID assumption.

The second assumption is that a unique BLP exists. This data has a finite range and already appears to converge at a distinct mean within a sufficiently-large sample size of n = 150; it should be noted that the total population size of all countries in the world is 195 according to the website www.worldometers.info. Furthermore, its distribution appears to be fairly normal and avoids heavy skews along both sides. As such, it appears that a unique BLP indeed exists.

Therefore, all of the large-sample assumptions are satisfied with the caveat to expect slightly larger standard error estimates due to the weak IID assumption.

9.2 Structural limitations of our model

We understand that our models met some, but not all, of the assumption associated with parametric regression models. As previously noted, we discussed that we are confident that our models met the assumptions required such as independence, noncollinearity and best linear predictor. This gave our team confidence that the variables used across all three models had unbiased coefficients.

While all three models unequivocally noted the statistical significance of GDP per capita and its impact on happiness, the models did not assess multicollinearity among our predicting variables. This presents a limitation because we cannot deterministically state which variables are highly correlated. For example, GDP per capita and social security may be highly correlated, as countries that are wealthier can afford to fund social security programs for its citizens. The inability to identify what variables are highly correlated could introduce noise into our models and increase the variability of our estimators.

An example of this with respect to our models started to become more obvious as we analyzed each model. In Model 1, we only compared happiness against GDP, the results were highlighted that GDP was statistically significant. However, as we introduced more variables into Models 2 and Model 3 the statistical significance of GDP decreased over time while freedome of life choices increased. For the most optimal output, we would have to better understand the multicollinearity within our models to truly be confident in our output.

Furthermore, there is the possibility that omitted variables would impact our analysis by accounting for variability that was not measured in our data set. Below are a few examples of potential omitted variables and what direction of bias their omission would cause:

- The impact of the COVID-19 pandemic likely impacted happiness levels across the globe due to being an outstanding global event that upset world stability. Its omission would lead to a bias away from zero, since the influence of COVID-19 would have had a negative impact on happiness. Perhaps the impact of the COVID-19 pandemic on happiness could be estimated by seeing how different the happiness level was for countries before and during the pandemic.
- Wealth disparity is a possible influence on happiness. While a nation's wealth may be an influence on its happiness, that happiness may not be shared evenly if the wealth is not shared evenly either. Its omission would lead to a bias away from zero, since a country with high levels of wealth disparity would have a negative impact on overall happiness. This data may be gathered from groups that study wealth inequality, such as the World Population Review's article on wealth inequality by country: https://worldpopulationreview.com/country-rankings/wealth-inequality-by-country.
- Social closeness is a characteristic in which certain societies value having greater interpersonal social ties to people in one's life, and that social connection leads to increased happiness; compare that to societies with low social closeness that normalize social isolation and the unhappiness that follows. Its omission would lead to a bias towards zero, since social closeness would be a positive impact on overall happiness. There may be academic studies that research and analyze the amount of social closeness there is in countries that may prove to have valuable data, but this may also be determined through additional questions in the World Happiness Report survey as well.
- Education level is associated with increased awareness and appreciation of one's life, world, and area of study, and thus may be associated positively with happiness. Its ommission would lead to a bias towards zero, since the influence of education would have had a positive impact on happiness. Perhaps the average education level per country can be found with certain academic or research groups, but it may also be determined through additional questions in the World Happiness Report survey as well.

In an ideal world, we would be able to gather additional data in order to account for the variables mentioned above. However, even with their omission, we feel that the omitted variables do not confound our results enough to invalidate our findings. As such, we stand by our analysis and findings as valid regardless of its structural limitations.

10 Conclusions & Recommendations

From our analysis of the World Happiness Report data, there appears to be sufficient evidence to suggest that wealth does indeed have a statistically significant impact on happiness. According to our most detailed model, for every 10% increase in GDP per capita, an individual's happiness score increases by an average of 0.27. This may have serious social and political ramifications, as this shows a clear and positive association between wealth and happiness, thus providing a counterpoint to oft-heard statements claiming otherwise.

Our results may be of key interest to governments and non-profit organizations focused on improving their nation's wealth or happiness. This study provides empircal evidence that may motivate these groups to make drastic changes to their societies and/or policies in order to maximize their nation's wealth and happiness, as it is suggested that the acquisition of one is associated with the acquisition of the other. However, it should be noted that wealth was not the only contributor towards a country's happiness, and that happiness was instead influenced by a number of important variables. For example, the variables for freedom_to_make_life_choices and social_support were shown to be both statistically significant and more influential on happiness than wealth was, and thus may actually be more desirable to maximize in the pursuit of maximizing happiness. Further research is necessary to understand the nuances and mechanics of how factors such as wealth, freedom, and social support influence happiness. We hope that a general understanding of wealth and its impact on happiness can benefit society by providing evidence for politicians and societies to leverage in pursuit of increasing the happiness of their nation's people.