Can Money Buy Happiness?
Understanding the Differential Between a Country's GDP Rank and Happiness Rank

Edith Edwards-Mizel, Varya Kluev, Greta Laesch
Professor Upton
STAT 346 - Regression Theory and Forecasting
May 19, 2023

Abstract

Countries around the world are annually ranked by their gross domestic product (GDP) as well as by population happiness. Though one would expect economic strength to Be In Line With happiness, often, an individual country's two rankings do not equal each other. We set out to explore what factors could contribute to what we call a country's "ranking differential," or what social, economic, or political metrics could be linked to a country either overperforming or underperforming in its expected happiness. Using primarily linear regression in conjunction with supplementary methods like principal component analysis and stepwise regression, we find that the most influential predictors affecting the wealth-happiness differential often vary by the economic status of a country; after creating subgroups based on income, these significant factors include rates of life, death, and access to clean resources.

Introduction

Each year the World Happiness report is released. Each country is given a score based on a survey given to residents of a given nation. The last report came out in 2021 and was done using surveys over the course of three years: 2019, 2020, and 2021. Although the survey, score, and ranking are not themselves a linear regression, the World Happiness Report attempts to explain each score through their own analysis, including six factors: dystopia and residual, perceptions of corruption, generosity, freedom to make life choices, healthy life expectancy, social support and GDP per capita. We wonder if that's all there is to it.

A plot and regression summary of standardized World Happiness (WH) rankings and standardized GDP per capita rankings shows a clear linear relationship between the two variables[1]. A simple linear regression of these two rankings has an R-squared value of 0.695. This is itself interesting and intuitive— greater wealth is associated with greater happiness— but we were interested in happiness beyond these rankings, in the deviations from this pattern. It is established that money can partially buy happiness, but what accounts for the additional variation? The World Happiness report has its answer— we have decided to conduct research and find our own.

More specifically, we aim to investigate the question of under-performance and over-performance of a country in its happiness ranking. A country with a GDP rank of 15 and a WH rank of 10 would be said to be "overperforming" its monetary resources. Meanwhile a nation with a GDP rank of 5 and a WH rank of 25 would be said to be "underperforming." What countries are doing more with less and how are they doing it? And what countries are squandering their monetary advantage?

The role of government is ultimately the well-being of the people. The wealth of a nation can be supported or inhibited by government policy, but this is not a goal in itself. To understand happiness in a more holistic manner is to look at government policies and social factors that go beyond the monetary. To a certain extent the question comes down to, money can buy happiness, but it can only go so far. What else makes a difference?

The Data

We are synthesizing many data sets for our analysis containing information about World Happiness rank, GDP information, and other various demographic data we are treating as explanatory variables population, rate of natural increase, total fertility rate, life expectancy for males, life expectancy for females, death rate, proportion of population with clean energy access, harmful air mean concentration, mortality homicide rate, adult obesity rate, tobacco use rate, expenditure health percentage, morality suicide rate, alcohol consumption, universal healthcare, total deaths, and corruption percentage index.

To calculate our response variable— the ranking differential— we utilized the World Happiness Report and World Population Review. The World Happiness Report calls upon Gallup World Poll data, which provides a framework for understanding subjective notions like happiness. Respondents are asked to make individual assessments of their lives and rank their happiness on a 0-10 scale. This single-item evaluation is collected from nationally representative samples over three years and is used to rank countries by happiness— the typical annual sample for each country is 1,000 people, making the total sample size 3,000. For our analysis, we are utilizing the most recent 2021 report. The World Population Review sources data from International Monetary Fund and United Nations databases to create a ranking of all countries by their calculated GDP. We are using their 2023 as opposed to their 2021 report to avoid exploring fluctuations in GDP resulting from COVID policies and unpredictability. We are using these quantitative ranks to create our quantitative response metric: subtracting a country's World Happiness Report ranking from its GDP ranking creates a score indicating whether a country is over-performing or underperforming in its happiness (the score's magnitude and sign capturing the phenomenon).

We derived the bulk of our explanatory variables from US Census Bureau data and the World Health Organization, and gathered other explanatory data directly from select sources. We selected a wide variety of both quantitative and categorical variables since our question is exploratory in nature rather than motivated by a narrow potential relationship. The US Census Bureau collects data from local, state, and federal governments as well as commercial entities through censuses and surveys. They obtain, reuse, and combine data from these parties for their own purposes, which primarily include providing metrics of assessing United States' populations and economies. We derived information about population, rate of natural population increase, total fertility rate, life expectancy for males and females, and crude death rate for our project. The World Health Organization collects and compiles data from both population and institution based sources, including civil registration systems, administrative activities of health facilities, and representative household surveys in over 101 countries, all of which measures 50+ indicators in the last few years. The information it collects allows the organization to create a comprehensive overview of worldwide health vulnerabilities and determinants of risks, and in turn to promote global wellbeing. While the WHO is a reputable organization that makes every effort to fairly and accurately compare statistics across countries and over time, the organization notes that this primary data may differ by data collection methods, population coverage and estimation methods used. We gathered data from Annex 2 for 2022 for our own work, specifically aggregate data for the proportion of clean energy access, average harmful air concentration, homicide rate, suicide rate, adult obesity rate, tabacco use rate, percentage of government expenditure on health care, and average alcohol consumption. We also included a categorical variable of universal health care.

Next, we wanted to pull metrics that were not included in the previous sources that we believed would be relevant to our study, specifically total COVID deaths and corruption perceptions index (CPI) ranking. Data for total COVID deaths came from the Johns Hopkins COVID resource center in the mortality resource data set, which we seeked out because we figured the effect of COVID would be pertinent as the study period for other data was during the pandemic. This data set had daily and cumulative death counts for each day for each country, and so we used subsetting to reduce the data to just the cumulative deaths from 01/03/2020 to 12/31/2022. For the CPI rankings we used the 2022 (most recent) report from Transparency International. Transparency International is an independent non-profit research and advocacy organization that produces comprehensive data measuring corruption within and across countries, and the effect on citizens, economies, and politics. We used the comparative data, ie. the rankings across countries as opposed to the index scores, as we figured rankings would be easier to interpret especially since that matches the formatting of our response variable. The data we collected reflects the state of social and economic factors in 2022-23, some variables being measured annually while others being compiled over the course of 2020-2023.

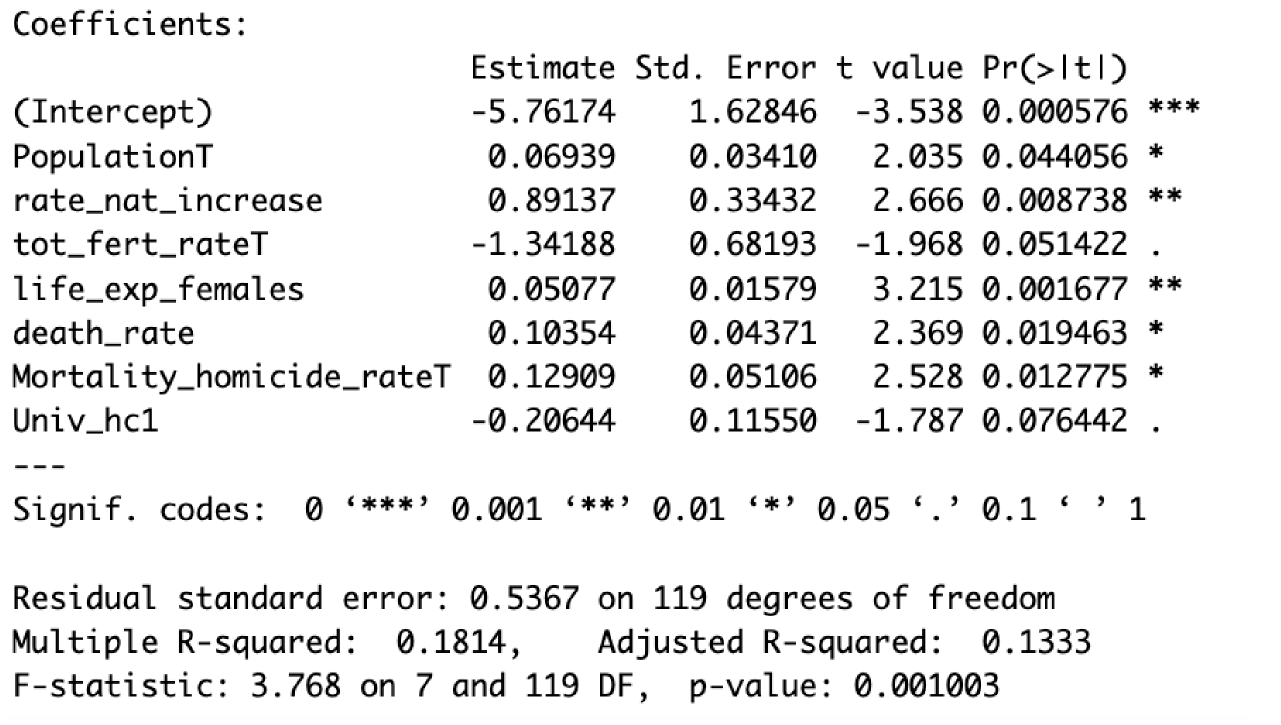
Strategy and Analysis

To understand what factors contribute to a country's large differential between GDP rank and World Happiness Rank, we used a variety of different analyses and methodologies. To begin, we combined our predictor variables of interest into one dataset. We started off with a lot of variables of interest: population, rate of natural increase, total fertility rate, life expectancy for males, life expectancy for females, death rate, proportion of population with clean energy access, harmful air mean concentration, mortality homicide rate, adult obesity rate, tobacco use rate, expenditure health percentage, morality suicide rate, alcohol consumption, universal healthcare, total deaths, and corruption percentage index (CPI). Some of these were directly oriented towards those factors that the WH report outlined. CPI corresponds to the World Happiness reports corruption value. Life expectancy, health expenditure, adult obesity, and alcohol assumption are attempts to capture a healthy life expectancy. Universal healthcare may serve as a stand in for social support. Freedom to make life choices may be observed through such explanatory variables of mortality suicide rate. Mortality homicide rate might correspond to utopia. Others were simply throwing things at the wall and seeing what stuck.

We first decided to engage in preliminary exploratory data analysis, with the goal of cleaning up pieces of messy or potentially incorrect or missing data as well as narrowing down our predictor variables to a handful of the most pertinent ones that could be explored, analyzed, and distinguished as key players. During our process of finding funky entries (of which there were not many), we were able to spot interesting data values that we could mentally take note of as being potentially statistically influential. Furthermore, we came across a few blank cells in certain variables for certain countries. We decided to keep the *NA*s in our full final dataset, since taking out every row containing an *NA* value would result in a loss of a handful of countries to analyze, perhaps ones with large differentials or interesting and insightful associated data. Moreover, a lack of information can be interpreted as information in its own right— a country lacking a certain statistic— having no measure of the proportion of people who have access to clean energy, for instance— can itself point to its social, political, or economic structure and priorities that could be important in understanding an existing difference in wealth and expected happiness.

To narrow down our variables to those most applicable, we compared pairs of variables and concurrently conducted stepwise analyses. Using the function ggpairs on our first full data set, we were able to take note of which variables looked highly correlated with each other and with the differential we set out to explore. To rely on more of a mathematical method of selection relative to interest and intuition, we proceed to run a backwards stepwise regression using AIC as a metric of evaluating model efficiency to derive a reduced model. Backwards stepwise regression, as opposed to forwards or bi-directional stepwise regression, was the most sensible option due to the nature of our exploration. We did not go in with a hypothesis of variables that we thought would be determinant and instead took an exploratory approach where we wanted to narrow down on specific, significant variables out of many that we highlighted as curious or potentially noteworthy. Allowing for all to be accounted for at the beginning of the model selection process, we run less of a risk of leaving out an important effect.

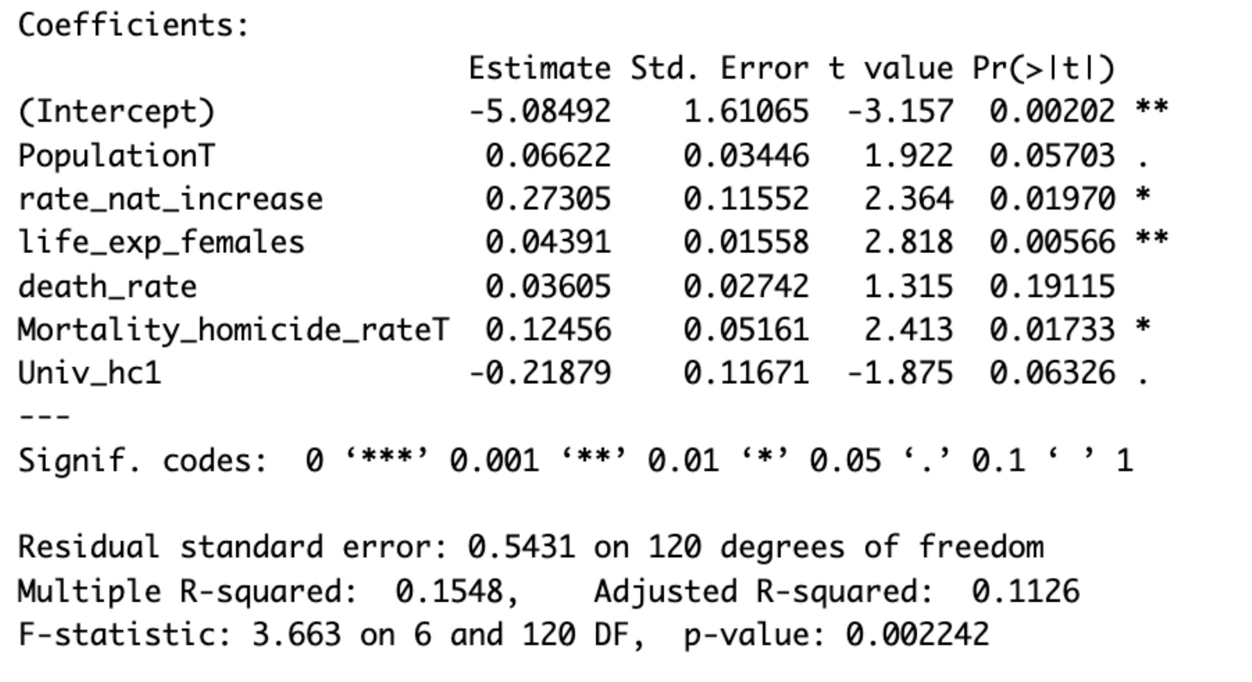
First, before doing this, we studied individual plots of each of the variables against our response variable DIF_rank. Viewing each of these plots we transformed[2] 6 variables: population, total fertility rate, harmful air mean concentration, mortality homicide rate, mortality suicide rate, and cpi score. All were transformed by log. Having performed these preliminary transformations we performed our backwards stepwise regression. The resulting model included, transformed population, rate of natural increase, total fertility rate transformed, life expectancy for females, death rate, mortality homicide rate transformed, and universal healthcare with an AIC value of -163.67.



```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)               -5.76174    1.62846  -3.538 0.000576 ***
PopulationT                0.06939    0.03410   2.035 0.044056 *  
rate_nat_increase          0.89137    0.33432   2.666 0.008738 ** 
tot_fert_rateT            -1.34188    0.68193  -1.968 0.051422 .  
life_exp_females           0.05077    0.01579   3.215 0.001677 ** 
death_rate                 0.10354    0.04371   2.369 0.019463 *  
Mortality_homicide_rateT   0.12909    0.05106   2.528 0.012775 *  
Univ_hc1                  -0.20644    0.11550  -1.787 0.076442 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5367 on 119 degrees of freedom
Multiple R-squared:  0.1814,    Adjusted R-squared:  0.1333 
F-statistic: 3.768 on 7 and 119 DF,  p-value: 0.001003
```

Checking the residuals it appeared that the residuals showed approximately constant variance[3]. Moving on to checking the issue of multi_colinearity, there was an issue with the rate of natural increase and total fertility rate transformed. They had VIF values of 54.416278 and 32.5499 respectively, and a ggpairs plot showed a correlation value of 0.948. Because of the issue of multicollinearity we decided to eliminate the total fertility rate transformed explanatory variable. This was the resulting model summary.



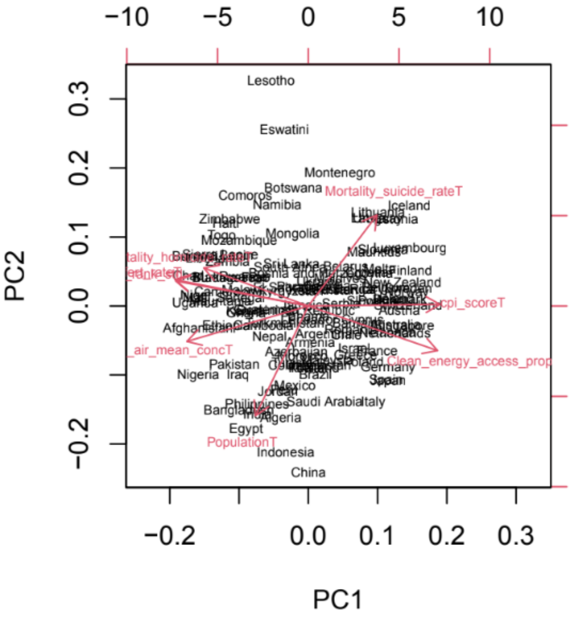
```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)   
(Intercept)               -5.08492    1.61065  -3.157  0.00202 **
PopulationT                0.06622    0.03446   1.922  0.05703 . 
rate_nat_increase          0.27305    0.11552   2.364  0.01970 * 
life_exp_females           0.04391    0.01558   2.818  0.00566 **
death_rate                 0.03605    0.02742   1.315  0.19115   
Mortality_homicide_rateT  0.12456    0.05161   2.413  0.01733 * 
Univ_hc1                  -0.21879    0.11671  -1.875  0.06326 . 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5431 on 120 degrees of freedom
Multiple R-squared:  0.1548,	Adjusted R-squared:  0.1126 
F-statistic: 3.663 on 6 and 120 DF,  p-value: 0.002242
```

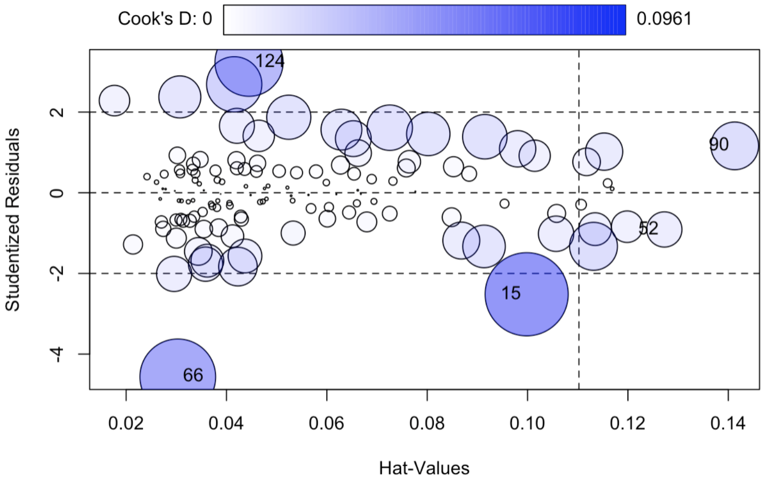
A visualization of the residuals showed constant variance in the residuals[4]. The model assumptions did not appear to be met. We once more looked at the pairwise correlations of the explanatory variables included with our response variable to check for multicollinearity as well as to better understand the relationships existing in our data. Studying our calculated VIF values again, all values were below 7.

Evaluating this overall model we find that only 15.48% of the variation in our response variable is explained by our model. Interpreting the coefficients, we can the following: on average, with all other variables remaining the same, for every 1 unit increase in log(population) we expect an increase of 0.0662 in our difference rank response variable. On average, with all other variables remaining the same, for every 1 unit increase in rate of natural increase we expect an increase of 0.273 in our difference rank response variable. On average, with all other variables remaining the same, for every 1 unit increase in life expectancy for females we expect an increase of 0.04391 in our difference rank response variable. On average, with all other variables remaining the same, for every 1 unit increase in death rate we expect an increase of 0.03605 in our difference rank response variable. On average, with all other variables remaining the same, for every 1 unit increase in log(mortality by homicide rate) we expect an increase of 0.0662 in our difference rank response variable. On average, with all other variables remaining the same, we expect countries with universal healthcare to have a -0.219 decrease in our difference rank response variable.

We will not further examine these interpretations because, as evidenced by the low r-squared value, further analysis is necessary to examine our response variable.



To conduct this further analysis, we viewed principal component analysis to look for next steps. We were curious to see whether the countries in our data set whose differential we calculated could be split up into a few different subsets, and thought that looking at principal components would help us see potential groupings and in general better understand our predictor variables. Using our final model (with categorical variables taken out for the purpose of this analysis), we looked at the biplot of our first two principal components. Countries seem to cluster by geographical region— nations in southern Africa, such as Lesotho, Eswatini, Botswana, Namibia, and Zimbabwe were grouped together. Similarly, one could interpret the countries' clusters as being grouped by economic development status. Furthermore, the biplot highlighted countries that seemed like outliers in terms of the principal components, which pointed us in another important direction for our exploration of influential cases.



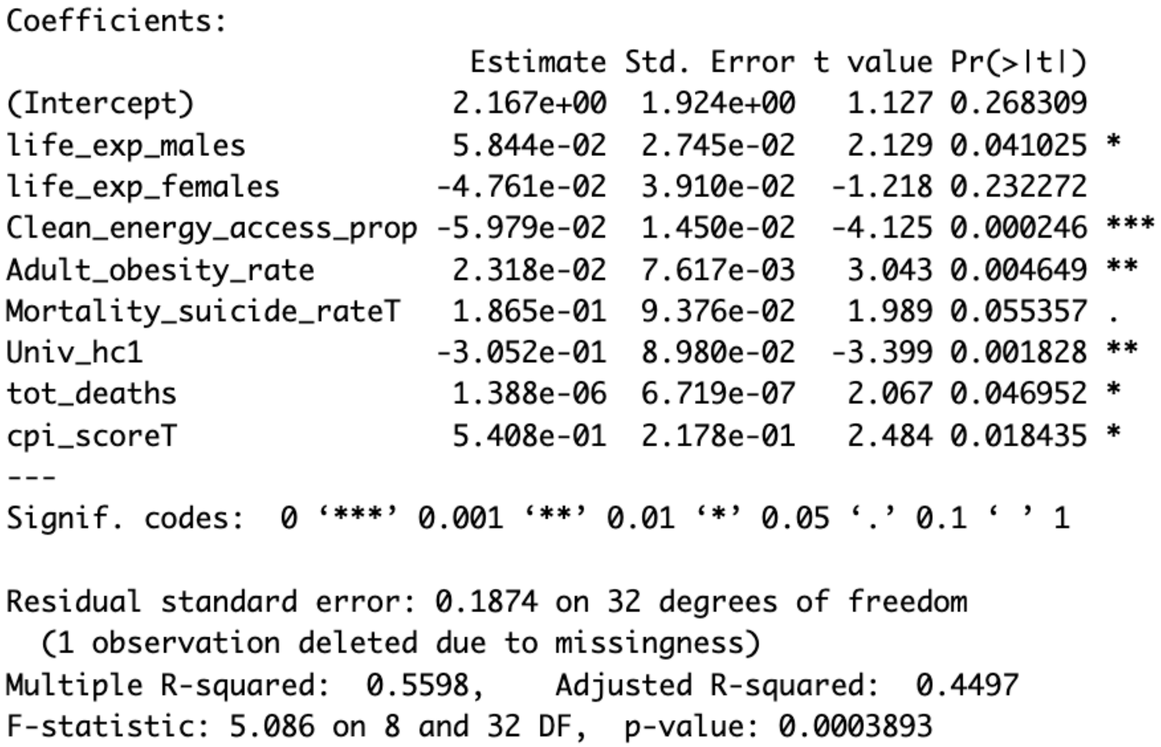
We decided to look at interesting observations that did not fit the general trends in our analyses by utilizing an influence plot measuring leverage, residuals, and Cook's Distance. We took note of the information that countries highlighted by our influence plot (namely, 66: Lebanon, 90: Niger, 15: Botswana and 124: Uzbekistan), and decided to include them in our data set for consequent analysis despite their deviations. Our priorities lie with having more complete and representative data, and no specific metrics from our exploration of interesting observations raised any critical red flags that threatened to compromise our work. However, they have been flagged for further investigation in our conclusion section.

Because of the limited success of our regression analysis, we decided to subset the countries by GDP into three categories of equal size, which we will henceforth refer to as first, second, and third, with first being the wealthiest nations as demonstrated by GDP per capita. Fitting the final overall model on each of the subsets, we observed that our model was not suitable for our first group. Using the same variables, it had an r-squared value of 0.06591, an adjusted r-squared value of -0.09422, and a p-value of 0.8663. The residuals did meet our normal assumption. Our third group likewise showed the original model variables to be unsuitable, with an r-squared value of 0.1597, adjusted r-squared value of 0.01563, and a p-value of 0.3722. The residual plot did show approximate normality. The model variables appeared to work best for the second group, or those countries in the middle for GDP. With an r-squared value of 0.5399. Additionally, the residual plot was the most normal for this subsection of the data.

This suggested that there was not one set of explanatory variables, nor regression coefficients that could capture the variation in our response variable. Therefore we decided to perform individual stepwise-regressions for each of the three subsections of countries, changing the scope of what we were attempting to answer. It appeared that at least in terms of the variables we outlined, there were not universal relationships between differential in rank and any of the variables across gdp.

First Group:

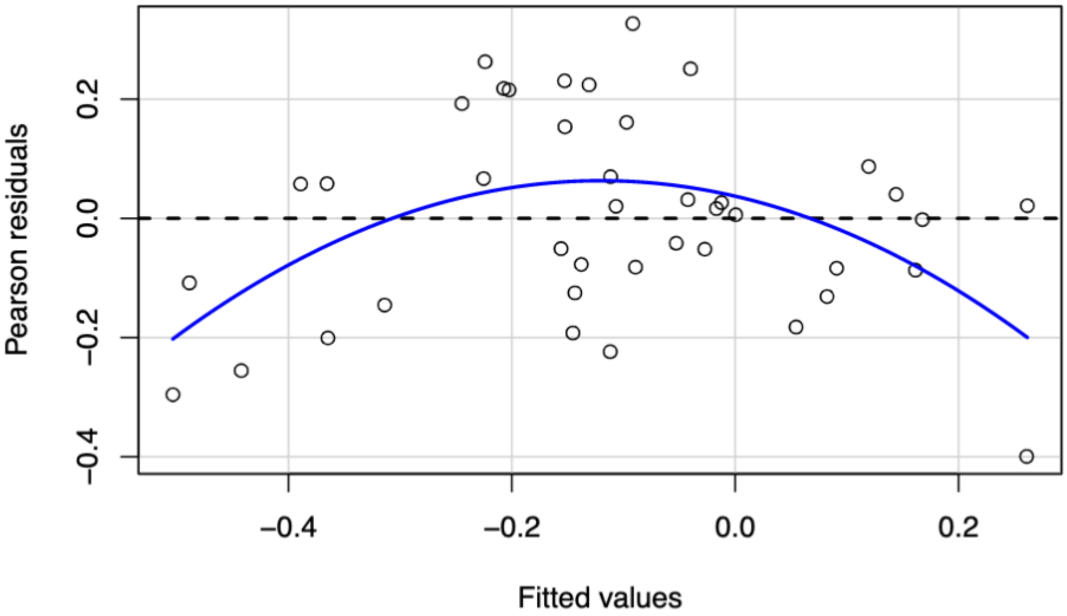
We performed a backwards stepwise regression as we had done for the model across all GDP ranks. Doing so the final model output included life expectancy males, life expectancy females, clean energy access proportion, adult obesity rate, mortality suicide rate transformed, universal health care, total deaths and CPI score transformed with an AIC value of -133.82.



```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              2.167e+00  1.924e+00   1.127 0.268309    
life_exp_males           5.844e-02  2.745e-02   2.129 0.041025 *  
life_exp_females        -4.761e-02  3.910e-02  -1.218 0.232272    
Clean_energy_access_prop -5.979e-02  1.450e-02  -4.125 0.000246 ***
Adult_obesity_rate       2.318e-02  7.617e-03   3.043 0.004649 ** 
Mortality_suicide_rateT  1.865e-01  9.376e-02   1.989 0.055357 .  
Univ_hc1                -3.052e-01  8.980e-02  -3.399 0.001828 ** 
tot_deaths               1.388e-06  6.719e-07   2.067 0.046952 *  
cpi_scoreT               5.408e-01  2.178e-01   2.484 0.018435 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1874 on 32 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.5598,    Adjusted R-squared:  0.4497 
F-statistic: 5.086 on 8 and 32 DF,  p-value: 0.0003893
```

This was the summary of the model. 55.98% on the variation in our response variable is explained by our model. All VIF values were under 10. But the residuals do not appear to have constant variation.



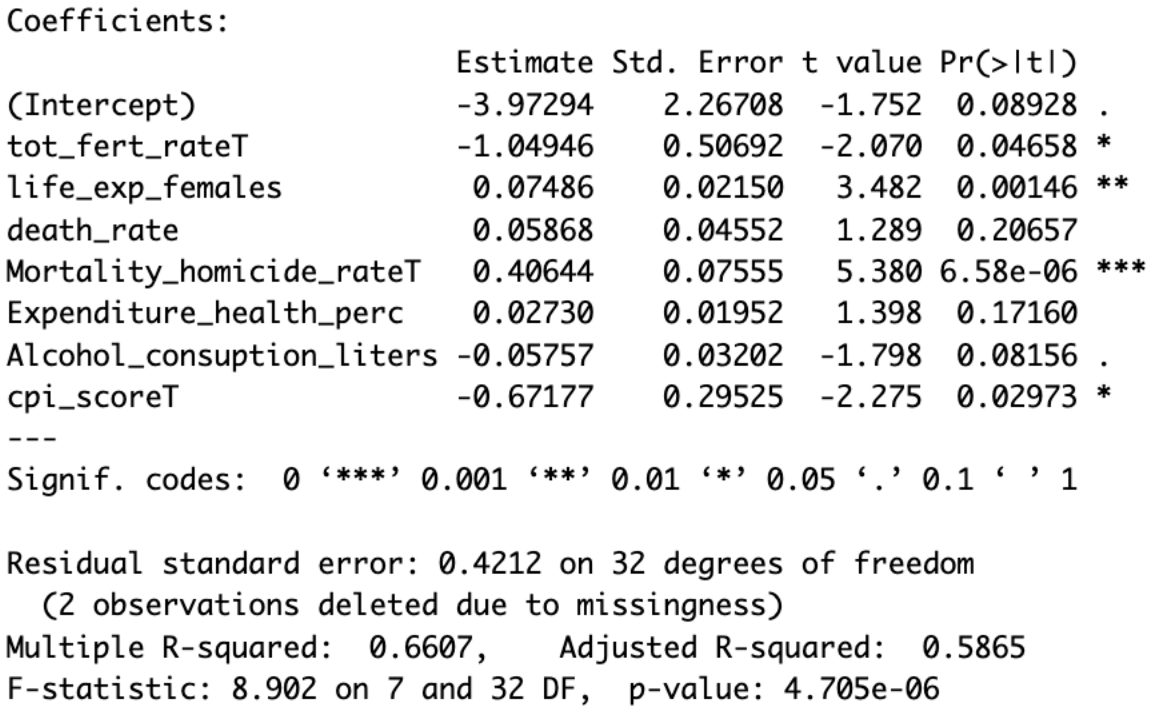
Viewing the ggpairs of this model, it did not appear that any of the individual relationships of the explanatory variables to the response variable were non-linear. Which suggests that a transformation of the y-variable should be useful. However, since this was not an issue with the other models, and for purposes of consistency as well as for reasons of interpretability we are going to proceed with caution.

The interpretations of each of the coefficients are as follows: We expect on average, with all other variables remaining the same, an increase in standardized happiness differential of 0.0584 for every 1 year increase in life expectancy of males. We expect on average, with all other variables remaining the same, a decrease in standardized happiness differential of 0.0476 for every 1 year increase in life expectancy of females. We expect on average, with all other variables remaining the same, a decrease in standardized happiness differential of 0.0598 for every percentage increase in proportion of the population with access to clean energy. We expect on average, with all other variables remaining the same, an increase in standardized happiness differential of 0.02318 for every 1 unit increase in adult obesity rate. We expect on average, with all other variables remaining the same, an increase in standardized happiness differential of 0.1865 for every 1 unit increase in log of mortality suicide rate. We expect on average, with all other variables remaining the same, a decrease in standardized happiness differential of 0.3052 for countries with universal health care. We expect on average, with all other variables remaining the same, an increase in standardized happiness differential of 0.5408 for every 1 unit increase in log CPI score.

CPI accords with our intuition and with the World Happiness reports on calculations. As a high CPI score shows higher trust in the government or lower values of perceived corruption. Life expectancy for males likewise accords with the World Happiness Report’s understanding of healthy life expectancy. All others are contrary to expectation, but we will attempt to explain what they are capturing. Life expectancy for females follows more closely with access to health care as women are more likely to seek health care. Therefore similar to our universal health care variable it may be capturing wealthier nations without a corresponding benefit to happiness. Total deaths may likewise accord with GDP increases as industrialized nations tend to be older, and capture the GDP rank without having a significant effect on Happiness. We can find no similar explanations for mortality suicide rate or adult obesity rate, having positive values.

Second Group:

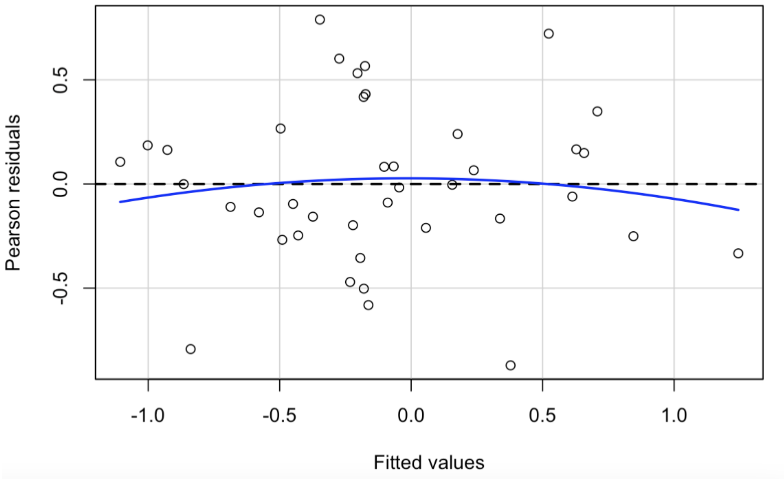
We performed a backwards stepwise regression as we had done for the model across all GDP ranks. Doing so the final model output included mortality suicide rate transformed, life expectancy females, death rate, total fertility rate transformed, clean energy access proportion, CPI score transformed, and Morality homicide rate transformed with an AIC value of -54.41.



```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)                -3.97294    2.26708  -1.752  0.08928 .  
tot_fert_rateT             -1.04946    0.50692  -2.070  0.04658 *  
life_exp_females            0.07486    0.02150   3.482  0.00146 ** 
death_rate                  0.05868    0.04552   1.289  0.20657    
Mortality_homicide_rateT    0.40644    0.07555   5.380 6.58e-06 ***
Expenditure_health_perc     0.02730    0.01952   1.398  0.17160    
Alcohol_consuption_liters  -0.05757    0.03202  -1.798  0.08156 .  
cpi_scoreT                 -0.67177    0.29525  -2.275  0.02973 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4212 on 32 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.6607,	Adjusted R-squared:  0.5865 
F-statistic: 8.902 on 7 and 32 DF,  p-value: 4.705e-06
```

This was the summary of the model. With an r-squared value of 0.6607 we understand that our model explained 66.07% of the variation in the response variable. All VIF values were under 4. And the errors appeared to be normally distributed.

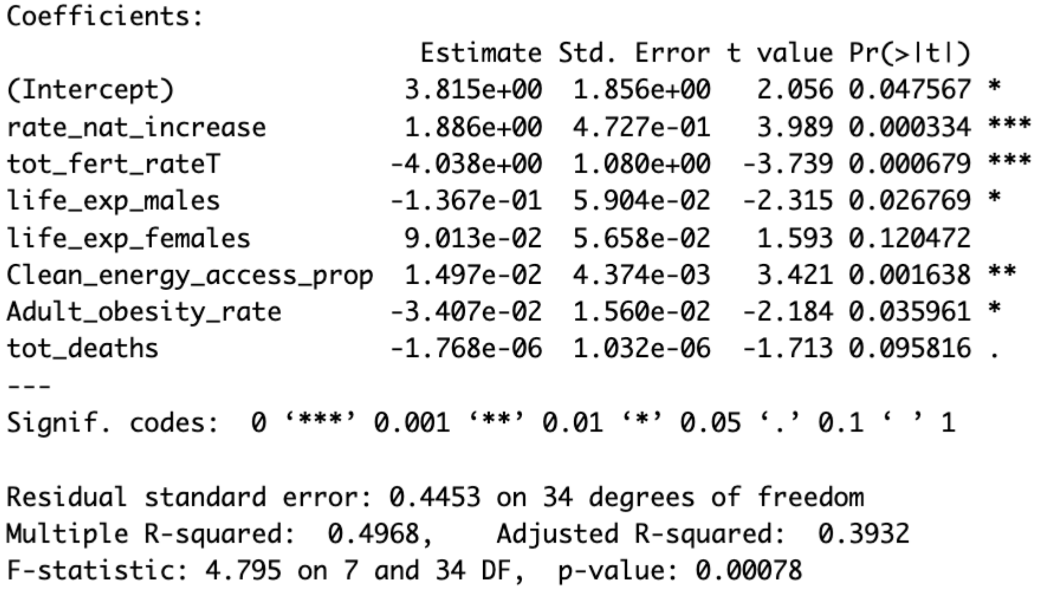


The interpretations of each of the coefficients are as follows: We expect on average, with all other variables remaining the same, a decrease in standardized happiness differential of -1.049 for every 1 unit increase in log(total fertility rate). We expect on average, with all other variables remaining the same, an increase in standardized happiness differential of 0.07486 for every 1 year increase in life expectancy of females. We expect on average, with all other variables remaining the same, an increase in standardized happiness differential of 0.0598 for every one unit increase in the death rate. We expect on average, with all other variables remaining the same, an increase in standardized happiness differential of 0.406 for every 1 unit increase in log(mortality_homicide rate). We expect on average, with all other variables remaining the same, an increase in standardized happiness differential of 0.0273 for every 1 unit increase in health expenditure percentage. We expect on average, with all other variables remaining the same, a decrease in standardized happiness differential of 0.0576 for every one unit increase in alcohol consumption in liters. We expect on average, with all other variables remaining the same, a decrease in standardized happiness differential of 0.672 for every 1 unit increase in log(CPI score).

These coefficients, including their signs do not align with expectations and in some cases do not align with the first model. CPI score for example has a negative coefficient, which is different from both the expectations of the World Happiness Report and from the previous model. Alcohol consumption accords with our expectation that higher consumption is a reflection of dissatisfaction or less happiness. It also might accord with the WH Report's understanding of healthy life expectancy. Health expenditure suggests greater social support, which is what we expected. Notably, social support in the form of health care is not accounted for by our universal health care variable, but by health expenditure, which is scaled rather than taking two values. Life expectancy for females does accord with our expectations and with the WH report's understanding of happiness. The negative total fertility rate transformed may correspond to freedom to make life choices in the case of women and access to birth control, and other adult life possibilities beyond child rearing. We cannot find an explanation for the positive values of mortality homicide rate, or death rate.

Third Group:

Lastly, we performed the same stepwise backwards regression for the lowest income set of countries.



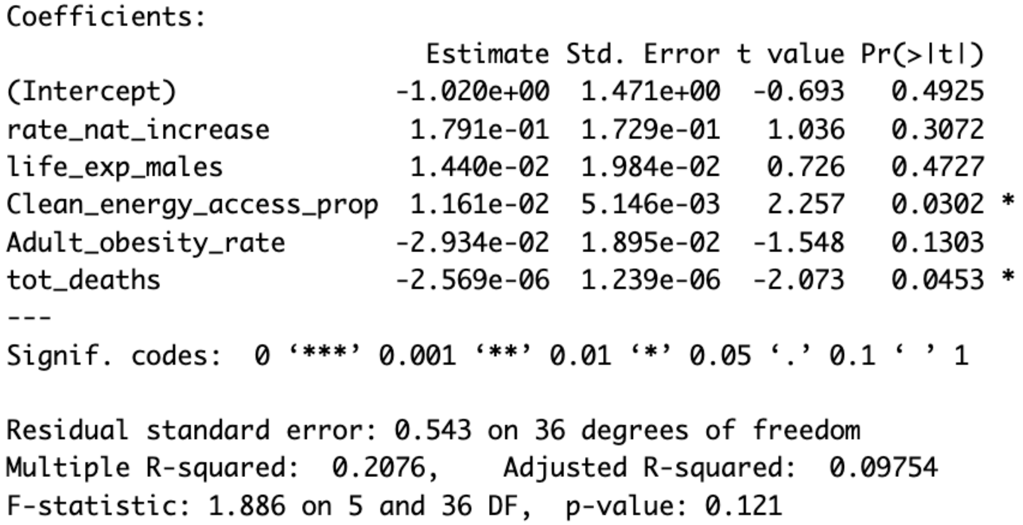
```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              3.815e+00  1.856e+00   2.056 0.047567 *
rate_nat_increase        1.886e+00  4.727e-01   3.989 0.000334 ***
tot_fert_rateT          -4.038e+00  1.080e+00  -3.739 0.000679 ***
life_exp_males          -1.367e-01  5.904e-02  -2.315 0.026769 *
life_exp_females         9.013e-02  5.658e-02   1.593 0.120472
Clean_energy_access_prop  1.497e-02  4.374e-03   3.421 0.001638 **
Adult_obesity_rate      -3.407e-02  1.560e-02  -2.184 0.035961 *
tot_deaths              -1.768e-06  1.032e-06  -1.713 0.095816 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4453 on 34 degrees of freedom
Multiple R-squared:  0.4968,    Adjusted R-squared:  0.3932
F-statistic: 4.795 on 7 and 34 DF,  p-value: 0.00078
```

Before going too far into the analysis, we looked at the VIF values for these variables expecting to find multicollinearity among the first four, and this was found. Here are the following VIF values for the first four explanatory variables: 28.478, 31.615, 20.732, and 22.407.

Although these variables would be appropriate as a linear combination and could be well served by principal components, in order to preserve interpretability we elected to remove total fertility rate and female life expectancy.

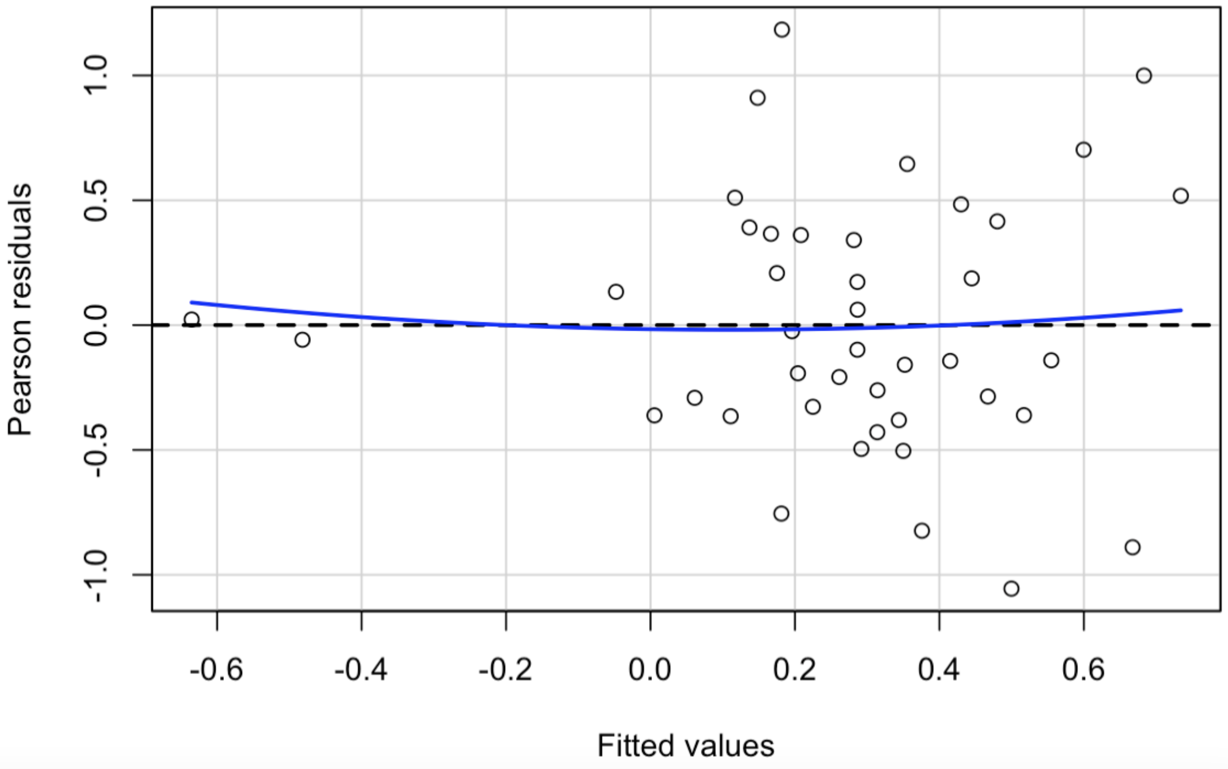
The following is our final summary output.



```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -1.020e+00  1.471e+00  -0.693   0.4925
rate_nat_increase        1.791e-01  1.729e-01   1.036   0.3072
life_exp_males           1.440e-02  1.984e-02   0.726   0.4727
Clean_energy_access_prop  1.161e-02  5.146e-03   2.257   0.0302 *
Adult_obesity_rate      -2.934e-02  1.895e-02  -1.548   0.1303
tot_deaths              -2.569e-06  1.239e-06  -2.073   0.0453 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.543 on 36 degrees of freedom
Multiple R-squared:  0.2076,    Adjusted R-squared:  0.09754
F-statistic: 1.886 on 5 and 36 DF,  p-value: 0.121
```

With an r-squared value of 0.2076 we understand that our model explained 20.76% of the variation in the response variable. This is a notable decrease in r-squared value from the previous model, however because of the issue of multicollinearity we could not be sure of the interaction effects between the collinear variables. After removing the total fertility rate and life expectancy of females, all VIF values were under 4. And the errors appeared to be normally distributed.



The interpretations of each of the coefficients are as follows: On average, with all other variables remaining the same, for every 1 unit increase in rate of natural increase we expect an increase of 0.1971 in our difference rank response variable. On average, with all other variables remaining the same, for every 1 year increase in life expectancy for males we expect an increase of 0.0144 in our difference rank response variable. On average, with all other variables remaining the same, for every 1 unit increase in clean energy access proportion we expect an increase of 0.0161 in our difference rank response variable. On average, with all other variables remaining the same, for every 1 unit increase in adult obesity rate we expect a decrease of 0.02934 in our difference rank response variable. On average, with all other variables remaining the same, for every 1 unit increase in total deaths we expect a decrease of 2.59 e-6 in our difference rank response variable.

Life expectancy males accords with our understanding of healthy life expectancy as a positive indicator of happiness. Clean energy access likewise accords with our understanding of both social support and positive healthy life expectancy. A negative effect on adult obesity rate likewise follows. A negative value for total deaths and a positive value for rate of natural increase follow together, as this suggests a growing population, which might correspond with, if not present day wealth, then an industrializing, growing nation.

Conclusion and Next Steps

In sum, we were attempting to answer the question of differences in GDP and World Happiness rank. The World Happiness report comes to its own conclusions using GDP along with a number of other previously mentioned variables, but we wanted to examine this. Starting with a wide number of variables, we sought to build a model that could explain the differences in rank for the entire dataset of countries. However, such a model had a limited ability to demonstrate any significant relationships. Therefore, upon further analysis, we decided that unable to create a single model for all countries, we could subset the countries by wealth. That what adds to a nation's happiness with a high GDP might be different from the relationship to happiness for those with lower GDP's. Subdividing the countries into thirds, we attempted to use the same explanatory variables as our overall model, and this further demonstrated that the model was ineffective across the subgroups. In a cross-validation-esque attempt to see if the model we had worked across subgroups. It did not, particularly for groups one and three, the richest and the poorest nations.

Seeing that both an overall model, and an overall set of variables, would not be useful across subgroups, we ran backwards stepwise regression on each subgroup. Arriving at a different set of variables. Below is a table of the explanatory variables and their coefficients for each.

***Bold = p<0.1**

First		Second		Third	
Intercept (b0)	2.167	**Intercept (b0)**	-3.97924	Intercept (b0)	-1.020
Male life expectancy	0.05844	**log(Total Fertility Rate)**	-1.049	Rate Natural Increase	0.1791
Female life expectancy	-0.0476	**Female life expectancy**	0.07486	Male life expectancy	0.01440
Clean energy access proportion	-0.0598	Death Rate	0.05868	**Clean Energy Access Proportion**	0.0161
Adult obesity rate	0.02318	**log(mortality homicide rate)**	0.40644	Adult Obesity Rate	0.1303
log(Mortality suicide rate)	0.1865	Health Expenditure %	0.02730	**Total Deaths**	-2.569e-06
Universal HC	-0.3052	**Alcohol Consumption Liters**	-0.05757		
Total Deaths	1.338 e-6	**log(CPI)**	-0.67177		
log(CPI)	0.5408				
R^2:0.5598		R^2: 0.6607		R^2: 0.2076	

From this table it may be observed that no variables that appeared in more than one regression, with the exception of male life expectancy, had the same sign in both cases. This communicates that some of the factors associated with happiness according to the WH report, when money is accounted for, do not show across the board increases in a nation's happiness. For example we see a negative value for CPI in the case of our second group, and this variable did not show up in the last group.

Additionally, it is clear that the initial possible explanatory variables were most effective in the case of the middle subsection of data. This may be due to the fact that these nations were primarily in the middle for both World Happiness and GDP ranking and differentials would not be biased by a nation being at either end. Further, our data shows that the variables were least effective in the case of the third group. This may be due to the fact that there are other variables that may do a better job of explaining differences. Or that nations with lower GDP tend to be less stable and therefore may have greater inherent variation in happiness and circumstance than those at the higher end of the spectrum.

If we were to continue our research, one area we would tackle first would be to engage in further manipulations with our data. For one, we could re-group our countries: instead of making three groups by GDP, it would be interesting to see how groups split up by geographical region or date of last constitution differ in the significance of their predictor variables— perhaps countries with more recently updated supreme law are more reflective of its population and thus lead to increases in happiness; perhaps the climate of a region increases pollution and thus decreases well-being. Then, it would be interesting to adjust our response variable, the ranking differential, to account for potential biases in the calculation of its magnitude. For instance, a country ranking in the top 10 for both GDP and happiness can have a maximum differential of 9; countries who rank lower on both scales can have larger differentials. Thus, the under or over performing of a country in the top rankings may not be representing the differential in an equivalent way to how it is being represented in countries lower on the scale. Perhaps there would be a mathematical way to weigh our differentials to account for these potential differences. Another place of inquiry is to look at changes in happiness across time. The WH report has been published since 2012, and we could look at change based on these different factors, and different natures to identify causes for various trajectories.

Lastly, we could look more into points that are outlying or at least interesting. We have gathered the points for each subset from the influence plot:

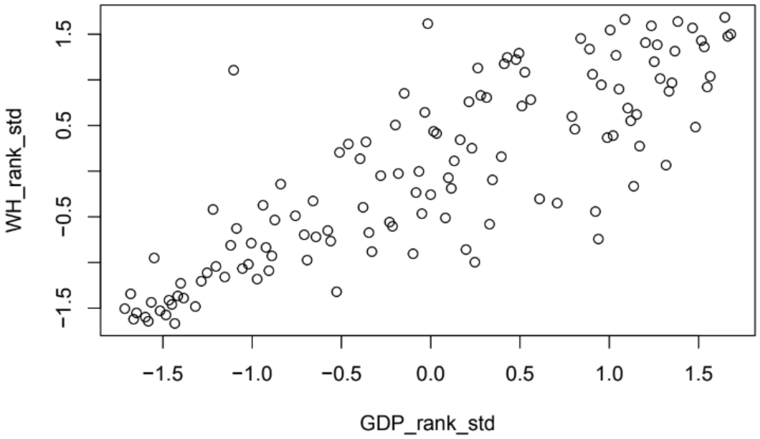
First: United Arab Emirates, Japan, Panama, Romania
Second: Ukraine, Botswana, Eswatini, Gabon
Third:Tunisia, Algeria, Nicaragua, India

With further investigation, perhaps they could point us towards variables we are not accounting for at present.

Appendix

1: Understanding the phenomenon we are exploring (through a scatter plot, fitted linear model, and residual plot analysis of World Happiness and GDP). Our results suggest that more information is necessary to explain the variation in happiness, which fuels our research.

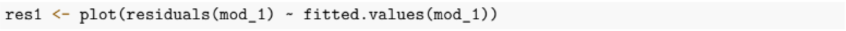
```
plot(WH_rank_std ~ GDP_rank_std)
```



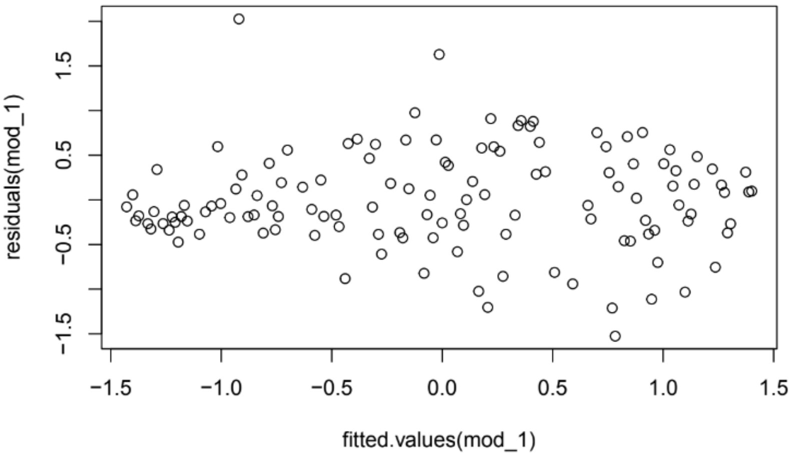
```
mod_1 <- lm(WH_rank_std~GDP_rank_std)
summary(mod_1)
```

```
##
## Call:
## lm(formula = WH_rank_std ~ GDP_rank_std)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.52639 -0.31305 -0.06676  0.34281  2.02637
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.672e-16  4.918e-02    0.00        1
## GDP_rank_std 8.338e-01  4.938e-02   16.89   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

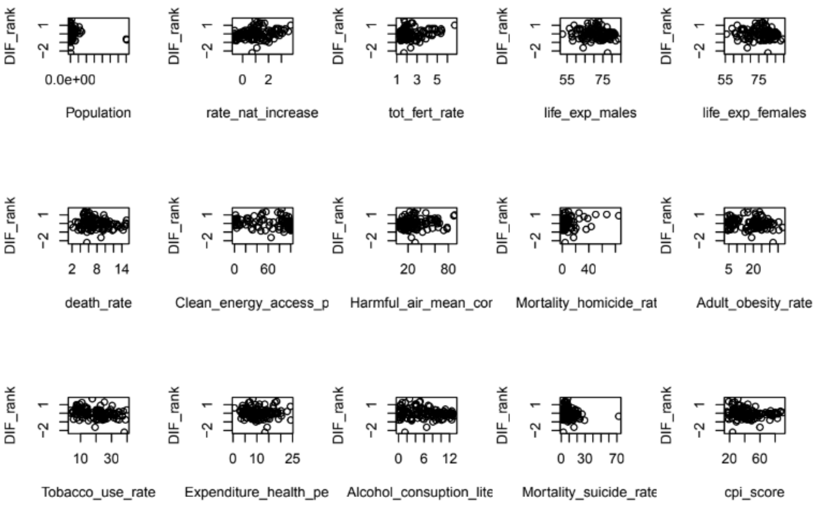
```
##
## Residual standard error: 0.5543 on 125 degrees of freedom
## Multiple R-squared:  0.6952, Adjusted R-squared:  0.6928
## F-statistic: 285.2 on 1 and 125 DF,  p-value: < 2.2e-16
```

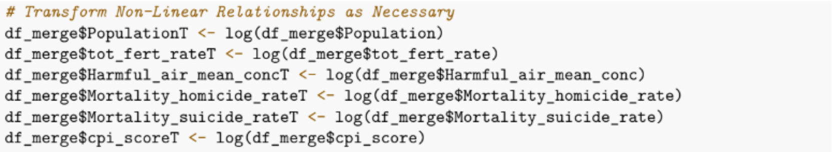


```r
res1 <- plot(residuals(mod_1) ~ fitted.values(mod_1))
```



2: Variable transformations

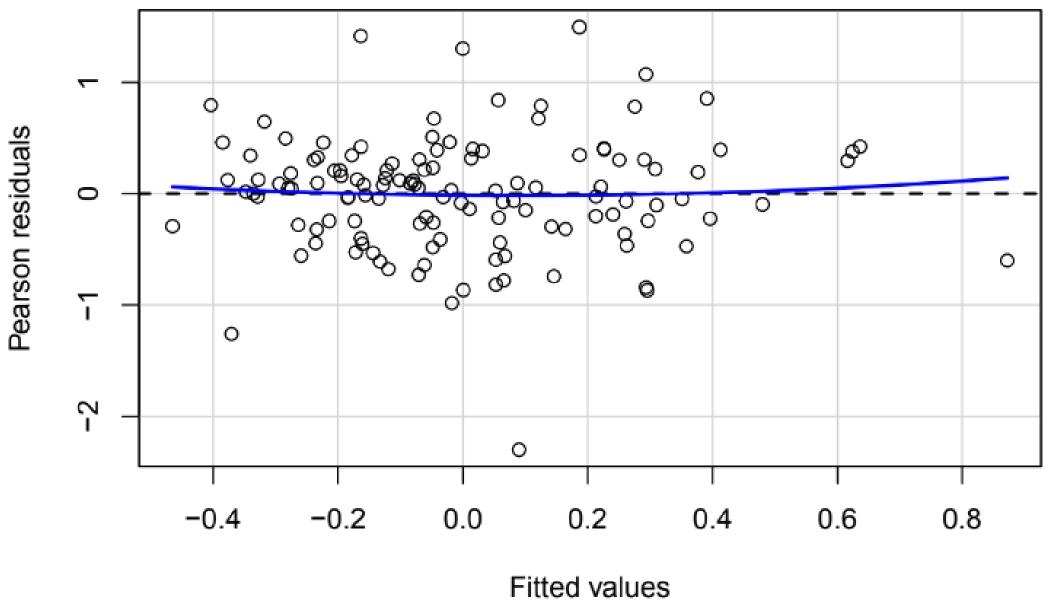




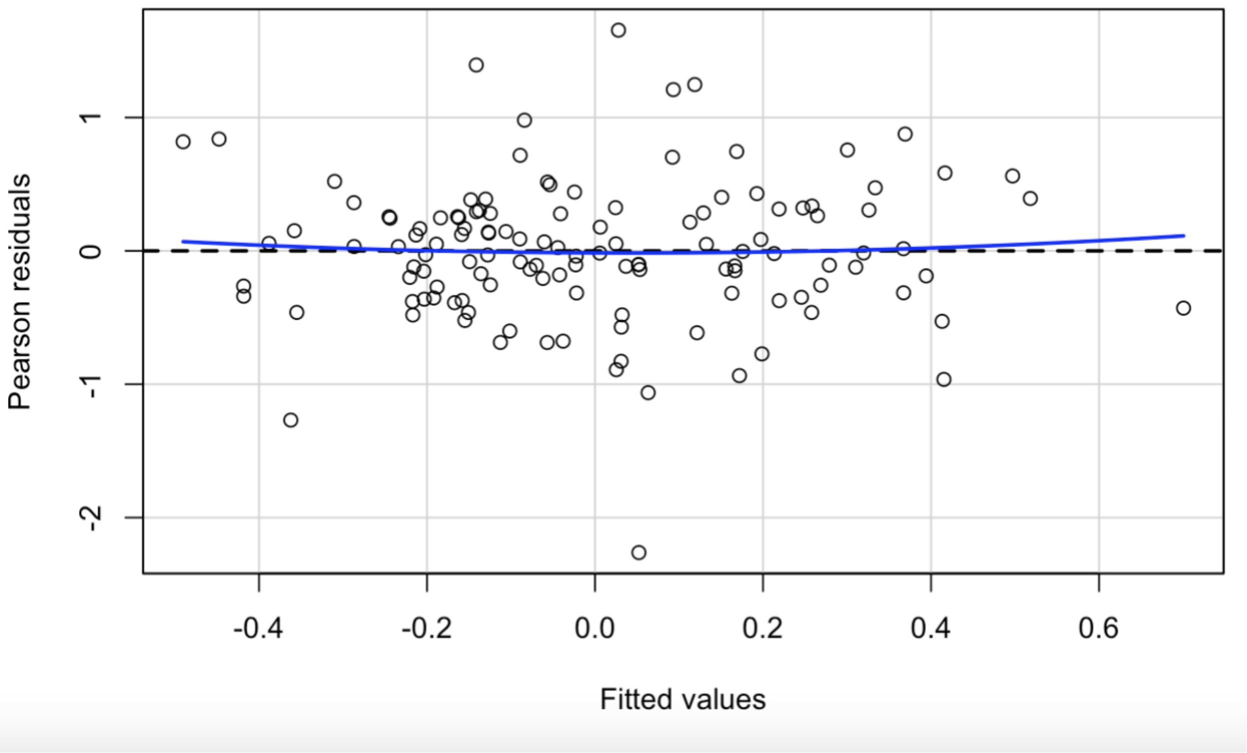
```r
# Transform Non-Linear Relationships as Necessary
df_merge$PopulationT <- log(df_merge$Population)
df_merge$tot_fert_rateT <- log(df_merge$tot_fert_rate)
df_merge$Harmful_air_mean_concT <- log(df_merge$Harmful_air_mean_conc)
df_merge$Mortality_homicide_rateT <- log(df_merge$Mortality_homicide_rate)
df_merge$Mortality_suicide_rateT <- log(df_merge$Mortality_suicide_rate)
df_merge$cpi_scoreT <- log(df_merge$cpi_score)
```

3: Preliminary results residual plot

```
residualPlot(mod_bwAIC)
```



4: Adjusted overall model residual plot.


Pearson residuals
1
0
-1
-2
-0.4
-0.2
0.0
0.2
0.4
0.6
Fitted values