COMP 4442

14 March 2021

Executive Summary

This analysis explores the World Happiness Report dataset over the years 2015 – 2019. The factors that all 5 years have in common are GDP, life expectancy, generosity, social support/family, freedom to make choices, and perceptions of corruption. The combination of these result in a happiness score for each country. There were 156 – 158 countries in the dataset for each year. While most countries were represented all five years, some had fewer data points. The countries were ranked based on their overall resulting happiness score. The question this analysis is seeking to answer is which of the six factors (GDP per capita, social support, life expectancy, freedom to make choices, perceptions of corruption, or generosity) is most important to the overall happiness score across all the countries. This analysis was performed using Longitudinal Multilevel Modeling. Country was evaluated as a random effect which generalized the results over the full set of countries. This analysis was performed using the lmer function from the lmerTest package. Nested models were created, starting with the simplest model building iteratively until all factors were included. The lmer function was applied to each of these models. The models were then compared with the anova function to find which factors (GDP per capita, generosity, etc) were statistically significant. The resulting model contained all six factors, that is, each factor that makes up the happiness score was found to be significant. The model yielded the following results with the factors listed from most to least significant: GDP per capita, social support, freedom to make choices, perceptions of corruption, life expectancy, and generosity. The years 2017, 2018, and 2019 had negative coefficients and were statistically significant. 2016 had a positive coefficient, but with a pvalue of 0.09, was not significant. Since country was randomized, these can be considered worldwide results. After checking the assumptions of in residuals, they did not meet the requirements. Based on this, the coefficients should not used, particularly at the high and low points. The middle points of the data seem to be a better fit of the data. Further analysis of this data would attempt transformations to improve the residuals.

Data

This data is from the World Happiness Report for the years 2015-2019, available from Kaggle (https://www.kaggle.com/unsdsn/world-happiness). It was collected by using Gallup poll data from a representative sample from each country. The happiness score is based on respondent's answers considering their happiness if the worst possible life is 0 and the best possible life is 10. Each country was given an overall happiness score based on an average of these results. The score was subdivided into six factors, with those numbers representing the extent to which each of these impacts the overall score. The factors were GDP per capita, life expectancy, generosity, social support/family, freedom to make choices, and perceptions of corruption. Factors that were included in the datasets from some years and not others were Dystopia and residuals. Dystopia was presented to participants as the worst possible place on earth and asked to compare their country to that. Residuals were the variance not explained by the

model. Since these two data factors were not available for all years, they were excluded from the analysis. (FAQ, 2020) All the factors added together, the six included in the analysis and the two not included, total the happiness score for each country.

Exploratory Data Analysis

Exploratory data analysis was performed on this dataset. First, the one NA in the dataset was removed. Additionally, there were 7 countries with an entry only one year. Subsets of the data were created in order to create a random effect on year. The data with only a single point would not be sufficient for this since a minimum of two points are required to estimate a line. However, when removing only the data with one point was not adequate for the function to converge, multiple subsets were created to find a subset that would work. This model can handle unbalanced groups, so that was not a consideration. Subsets were created with countries present two or more times, three or more times, four or more times, and present all five years. Additional exploration showed that the happiness score, the explanatory variable, had a normal distribution and no outliers. The minimum value of score was 2.2693, which was represented by Central African Republic in 2017 and the maximum score was 7.769, represented by Finland in 2019. Some of the factor variables, such as perceptions of corruption and generosity had quite a few outliers and nonnormal data. However, since these were not the explanatory variable, that was not overly concerning.

Analysis Method

To address the research question of which factor was most important in the happiness score, the first step was to combine the five years of data into a single set. There were some inconsistencies in column names. In the 2015 and 2016 datasets, Country and Region were separate columns, while in later datasets, country and region names were combined. In the datasets from 2015, 2016, and 2017 there were columns for the Dystopia and Residual. Those were not present in the 2018 and 2019 columns. In the combined set, the Region, Dystopia, and Residual columns were all removed and a column representing year was added. The row with a N/A value was removed. While 140 countries were represented in all five years in the data, there were 29 countries that were present 1 to 4 times. A method that allows inconsistent groups in analysis was necessary. Additionally, the research question could theoretically be solved with a multiple linear regression type analysis. However, the assumption of independence is violated with multiple data points from the same country. Longitudinal Multilevel Modeling is a method that can account for the violation of independence, mixed numbers of data points in groups, and analyze data across time. Therefore, this model was chosen for this analysis.

Longitudinal Mixed Model

Longitudinal mixed models analyze measurements that are repeated over time (Bates). In these models, there are fixed effects and random effects. The fixed effects are the measures that are repeated in each observation. For example, in this data study, the fixed effects are factors that go into each score, such as GDP per capita, life

expectancy, etc. The random effects allow the results to be generalized over the population ("Lesson 18: Mixed Effects Models | STAT 485"). The random effects are those that the researcher would like to generalize over. In this case, the research question was to find the most impactful factors in the happiness score all around the world, so the country factor was randomized. It also would have been possible to randomize over year in this analysis if there was enough data; however, there were not enough samples and the model did not converge.

Model Data Requirements

To run multilevel models, data with a hierarchical or clustered structure are required (University of Bristol and Rasbash). This dataset meets this requirement with the repeated country data points over the years from 2015-2019.

Following the analysis, a researcher must check the model's residuals to see if inferences can be drawn from the conclusions. First, there should be no patterns in the residuals. Second, there should be homogeneity of variance in the residuals. Finally, the residuals should be normally distributed. If these assumptions are violated, transformations, such as log, ln, square root, and others can be applied to the outcome variables. The entire process of analysis is repeated and the residuals are rechecked to see if the assumptions are met. However, if these transformations are successful, they may make the interpretation of results difficult. (Palmeri)

Method Application

The lmer function in the lmerTest package was used for this analysis. The lmer function fits a linear mixed effects model to data via restricted (residual) maximum likelihood (REML) estimation (Cheung). Its output is a model containing the fixed and random effects. In this analysis, nested models were iteratively created starting from the simplest model and building until all the factors were included.

The first model is shown below. It contains score as the outcome variable, year as the only fixed effects variable, and country as the random variable:

```
m0 <- lmer(score~year + (1 | country ), world_happ)
```

The original intent was to randomize year as well as country in order to generalize the output over all the countries and all the years. The model with this representation is shown:

```
m0.1 <- lmer(score~year + (year | country ), world_happ_subset)
```

The variation to allow year to randomize was attempted for each of the models and data subsets created. However, none of these models converged. There was a warning indicating that the number of observations were less than or equal to the number of random effects for term (year | country); the random-effects parameters and the residual variance (or scale parameter) are probably unidentifiable. Therefore, the results are not generalizable over all the years, only the countries. Additionally, to improve the errors, the final model was run with the balanced dataset. That

is, the dataset that had countries present in all five years of the data. That reduced the available data to 140 countries and 700 observations. While from the model standpoint, this was not necessary because the model works with unbalanced groups. However, the error rates improved with balanced groups due to the limited amount of data available.

After creating a model with the lmer function, the models with the country randomized were each compared with the anova function. The models were contrasted, beginning with the most complex model to the second most complex:

```
m6.0 <- lmer(score~year + GDP.per.capita  + life.expectancy + perceptions.corruption + social.support +
freedom.choices + generosity + (1 | country ), world_happ)

m5.0 <- lmer(score~year + GDP.per.capita  + life.expectancy + perceptions.corruption + social.support +
freedom.choices + (1 | country ), world_happ)
```

The difference in these models is that the generosity term is present in model 6, but not model 5.

```
anova(m6.0,m5.0)
```

The result of this anova had a pvalue of 0.03524, which indicated that model 6 was a significant improvement over model 5. Therefore, model 6, with all six of the terms was found to be the best model. All other anova model comparisons indicated that model 6 was the best representation to use.

After model 6 was selected, tests were run on the residuals to ensure the validity of the results. The residuals plot showed some pattern and grouping in the middle of the graph. There are some issues as well with normality at the upper and lower tails.

After observing plot diagnostics, the results were viewed using the summary function and interpreted. A further discussion of the results will follow.

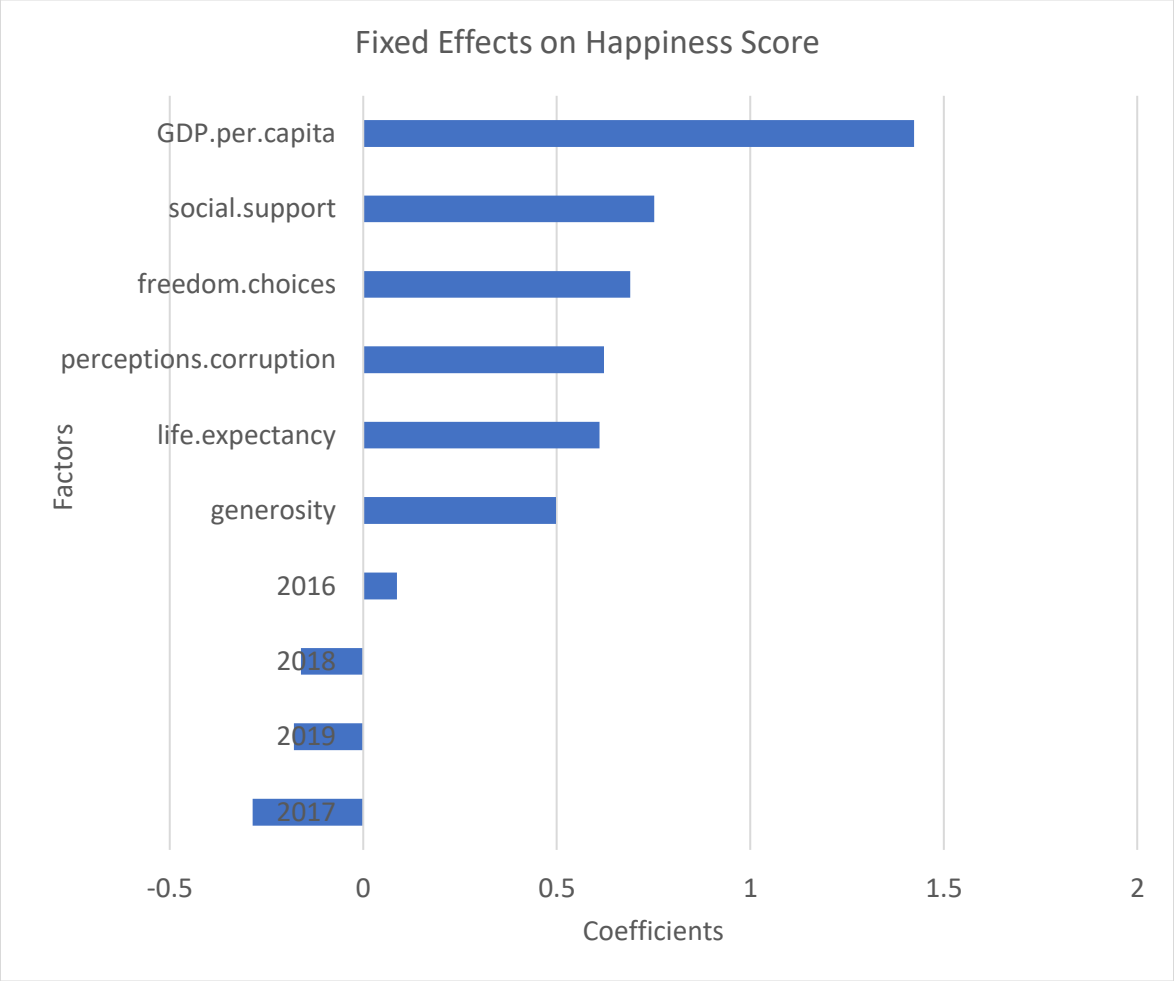
Results

The following results were obtained from the summary function on model 6:

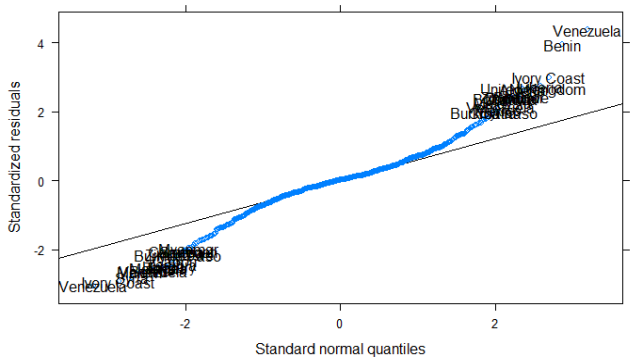
```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: score ~ as.factor(year) + GDP.per.capita + life.expectancy +
    perceptions.corruption + social.support + freedom.choices +      generosity + (1 | country)
   Data: world_happ_subset_balanced

Random effects:
 Groups   Name        Variance Std.Dev.
 country  (Intercept) 0.24451  0.4945
 Residual             0.05619  0.2370
Number of obs: 700, groups:  country, 140

Fixed effects:
```

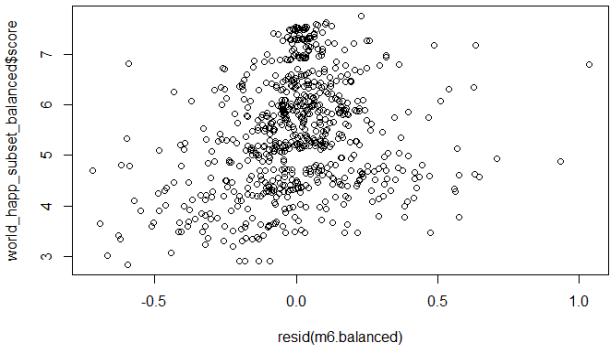
	Estimate	Std. Error	df	t value	Pr(>|t|)	
(Intercept)	2.52972	0.15287	235.90565	16.549	< 2e-16	***
as.factor(year)2016	0.08684	0.05154	672.16348	1.685	0.092511	.
as.factor(year)2017	-0.28625	0.05217	659.73828	-5.487	5.85e-08	***
as.factor(year)2018	-0.16062	0.04459	688.40337	-3.602	0.000339	***
as.factor(year)2019	-0.17948	0.04573	669.08985	-3.925	9.59e-05	***
GDP.per.capita	1.42364	0.15339	411.82725	9.281	< 2e-16	***
life.expectancy	0.61087	0.22814	526.07969	2.678	0.007647	**
perceptions.corruption	0.62179	0.27991	637.15861	2.221	0.026674	*
social.support	0.75214	0.13375	608.89320	5.624	2.85e-08	***
freedom.choices	0.68983	0.20260	670.91197	3.405	0.000701	***
generosity	0.49831	0.23841	561.59916	2.090	0.037055	*

The fixed effects indicated the greatest effect on the happiness score was from GDP per capita. Next was social support, then freedom to make choices, perceptions of corruption, followed closely by life expectancy, and lastly, generosity. The years 2017, 2018, and 2019 had statistically significant negative coefficients. 2016 had a slightly positive coefficient, however, it was not significant. The random effect, country, had a standard deviation of 0.49. Double that amount is approximately 1.0. Because 95% of data is within 2 standard deviations of the mean, this means 95% of country effect data will fall within 1.0 point of the estimate on the 10-point happiness scale.



Next, the assumptions were checked to test the validity of the output:





Variance comparison results:

Analysis of Variance Table

Response: Model.6Bal.Res2
Df Sum Sq Mean Sq F value Pr(>F)
score 1 0.3296 0.32964 39.468 5.869e-10 ***
Residuals 698 5.8298 0.00835

Each of these tests indicate concerns about the validity of results. The residuals are not distributed normally, and the residuals versus score plot shows patterns instead of being evenly distributed. Also, the ANOVA function results indicated that the homogeneity of variance assumption is also violated. Based on this, there can not be a lot of faith in the coefficients, particularly those at the extremes. Those in the middle appear to better model the data. Research of other analyses of this data show similar results to those that were indicated in this analysis. While there were not any direct comparisons available that combined these five years of data, other results were consistent that GDP per capita is the most important component in the overall happiness score. (Yang)

Further analysis of this data would add more years of data to have a sufficient amount to randomize over the year variable. If the residuals continued to violate assumptions with more data, transformations would be applied to the explanatory variable to see if they improve.

Works Cited

Bates, Douglas. “Lme4.” R-Forge, lme4.r-forge.r-project.org. Accessed 7 Mar. 2021.

Cheung, Mike. “Reml Function | R Documentation.” *R Documentation*, www.rdocumentation.org/packages/metaSEM/versions/1.2.5/topics/reml. Accessed 8 Mar. 2021.

FAQ, Sustainable Development Solutions Network, 2020, worldhappiness.report/faq/.

“Lesson 18: Mixed Effects Models | STAT 485.” PennState: Statistics Online Courses, Pennsylvania State University, online.stat.psu.edu/stat485/lesson/18. Accessed 7 Mar. 2021.

Palmeri, Michael. “Chapter 18: Testing the Assumptions of Multilevel Models.” *A Language, Not a Letter: Learning Statistics in R*, ademos.people.uic.edu/Chapter18.html#1_preface. Accessed 13 Mar. 2021.

University of Bristol, and Jon Rasbash. “What Are Multilevel Models and Why Should I Use Them?” *Centre for Multilevel Modelling | University of Bristol*, 19 Sept. 2017, www.bristol.ac.uk/cmm/learning/multilevel-models/what-why.html.

Yang, Sydney. “Student Association for Applied Statistics.” *World Happiness Report EDA*, 2017, saas.berkeley.edu/rp/world-happiness-report-eda.