Jack Rellamas
Dr. Gao
MATH-167R Sec 02
May 1, 2024

Inferential Analysis of the World Happiness Report Dataset

Introduction of Subquestion:

The subquestion that this report will explore is “What are the 3 strongest predictors of happiness out of the explanatory variables (GDP per capita, healthy life expectancy, freedom to make life choices, social support, generosity, and perception of corruption)? We want to find out what factors of a society, that were measured in this dataset, are most closely related to the happiness of a country. With this knowledge, one can then work to improve those factors that are most important to a happier society.

Data Analysis Plan:

The data analysis will start with a hypothesis test. The null hypothesis for each different predictor variable will be that there is no relationship between the predictor variables (listed in the subquestion) and happiness. The alternative hypotheses will be that there is a relationship between the predictor variables and happiness. If the null hypothesis is rejected, then we will look at the coefficient of the explanatory variables, β_1. The analysis will be done using linear regression and the significance levels will be given by the lm() function in R.

The columns for happiness score, GDP per capita, social support, etc, will need to be converted into quantitative data, or columns of doubles. As given by the original source, all of these variables are character columns. Additionally the strings are written with commas, so those need to be removed as well.

Assumptions:

For linear regression, there are four assumptions. They are linearity, independence, normality, and equal variance. Linearity can be checked with visual inspection of the plot between the independent variables and happiness. We know that our data is independent because each country is only surveyed once in the dataset. Normality can be checked with a quantile-quantile plot to see if the residuals are normally distributed. Equal variance can be checked with the residual plot, and seeing if the variance remains relatively constant.

Analysis and Interpretation:

Table 1: P-Values, Variable Coefficients, R-squared Values to Predict Happiness

	p-value	β_1	Adjusted R-squared
GDP per capita	< 2.2e-16	0.29628	0.5803
Social Support	< 2.2e-16	0.2005	0.6024
Healthy Life Expectancy	< 2.2e-16	0.12011	0.5448
Freedom to Make Life Choices	< 2.2e-16	0.083854	0.3862
Generosity	0.4444	4.859e-03	-0.002848
Perception of Corruption	1.74e-07	0.04883	0.1675

The 3 factors that stand out as better predictors of happiness are GDP per capita, social support, and healthy life expectancy. Their adjusted R-squared values are greater than 0.5, which means that the relationship between the independent variables and happiness explains more than 50 percent of the variation in the data. Additionally, they have the highest β_1 values out of the 6 explanatory variables.

It can also be seen that generosity and perception of corruption are not strong predictors of happiness. In the case of generosity, cannot reject the null hypothesis that is saying that there is a relationship between generosity and happiness.