ECON-386

Dr. Steven Levkoff

Data Regression Final Project

05/15/2020

**Estimating Happiness Score for Countries**

Lauren Hidalgo, Conrad Smits, Ernesto, Aruna, Tony Woods, Yifan

**Table of Contents**

***Need to add pages***

**Introduction**

* Executive Summary
* Variables of Interest
* Dependant Variable
* Regions
* Structure of Dataset

**Cleanup Process**

* Data Cleanup
* Breakdown of Process
* Issues with Multicollinearity

**Model Proposals for Regression Task**

**Introduction**

**Executive summary**

In this project, we aim to create a model to examine the relationships between a country's happiness score against a set of independent variables. Using the World Happiness Report of 2019, and other sources of hard data, we are looking to estimate the regression between a country’s given happiness score, and other independent factors that we have identified as variables of interest. Along with a regression, our goal is to solve a qualitative problem with the collected data sets.

The happiness score was derived from the World Happiness Report for the years 2006 to 2015, and the independent variables were taken from Gapminder.org. The dataset consists of inputs from 156 countries, of which have been divided into ten separate regions, as well as six independent variables of the ten variables of interest (see section titled “Variables of Interest” below), and a dummy variable for the regions’ categories. Using the raw data from those 156 countries, each matching with a given happiness score given in the report, our next step was to identify our variables of interest from the website.

**Variables of Interest**

* **Democracy Score:** An index created using the data from the Economist Intelligence Unit to express the quality of democracies as a number between 0 and 100. The data is based on 60 different aspects of societies that are relevant to democracy, i.e. universal suffrage for all adults, voter participation, perception of human rights protection, and freedom to form organizations and parties.
* **Gini Coefficient:** A measure of income inequality. The index number ranges from zero to one. A zero meaning that there is no income inequality within the nation. A gini coefficient of .8 or .9 would mean that all of .8 percent of all the income in the country would be going towards a very small percentage of the population.
* **Child mortality:** This measures the number of deaths of children under 5 years of age, per 1000 live births.
* **Refugee share of Population**: The share of refugees as a percentage of the total population of the country of residence.
* **Percentage of Population with basic sanitation:** The percentage of people using at least basic sanitation service.
* **Mean Years of Schooling for Women 25-34:** The mean years a woman spends in school between the ages of 25 and 34 years.
* **Mean Years of Schooling for Men 25-34:** The mean years a man spends in school between the ages of 25 and 34 years.
* **Children and Elderly per 1000 Adults:** Measures the total dependency ratio in a country, which is the percentage of the population composed of children under 14 and adults over 65 years.
* **Population Density per Km²**: The average number of people per square kilometer.
* **Labor Force Participation 15+:** The percentage of people over 15 years old active in the labor force of the total population.

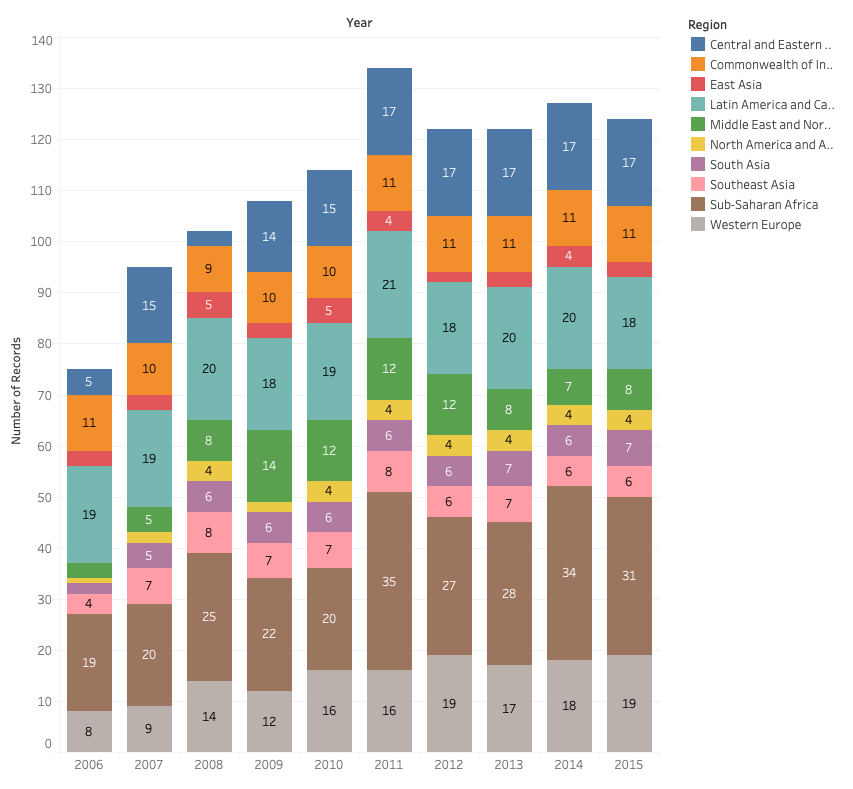
**Dependent Variable**

* **Happiness Score:** The score is based on multiple independent variables that ranks 156 countries by how happy their citizens perceive themselves to be.

**Regions**

We have included a dummy indicator for regions, including Western Europe, Central and Eastern Europe, Commonwealth of Independent States, Southeast Asia, South Asia, East Asia, Latin America and Caribbean, North America and ANZ, Middle East and North Africa, and Sub-Saharan Africa.

By doing this, we will be able to input variables into the model to retrieve an output of which region the record likely belongs to.



**Data Collection Process**

The data collected from the World Happiness Report was only the happiness score attached for each country. This was then exported into an Excel file where all the variables for the corresponding scores were attached. However, the independent variables for the World Happiness Report weren’t raw data that we could use. Therefore we decided to find our own independent variables for the 156 original countries that we found on a website called Gapminder. From there, we extracted the variables of interest into a seperate excel sheet. From here we started the process of cleaning the data. The data we collected is a mix of hard and soft data from websites. An example of the hard data is the happiness score for countries that the World Happiness Report decides because it was obtained through scientific observation and measurement. An example of the soft data we observed were the independent variables, our dataset that we extracted from was less structured.

**Structure of the Data Set**

We are using supervised learning in our data to best predict our dependent variable given all the independent variables we have. The reason it is supervised learning is due to the fact that both our x and y variables are known. Our data is structured as cross-section data by averaging across the time dimension for each country's metrics.

We have two datasets: “df” and “df\_dum”. The latter has a column region for the dummy variables, whereas the former does not. We have done this in order to be able to perform both quantitative and qualitative analysis on the data.

We have decided to not get rid of outliers in the original datasets. If desired, these may later be replaced with NA values, or left as is.

**Data Cleanup**

**Breakdown of Process**

Below are the steps we took in cleaning up our data. Please refer to the Rmarkdown file for the corresponding lines of code:

1. Line 10: Loaded data from Github into Rmarkdown.
2. Line 25: Created a key for the regions, each region corresponds with a unique number in order to be able to aggregate it.
3. Line 30: Changed the region data from categorical to numeric in order to be able use it in the model.
4. Line 33: Eliminated countries with no observations (Kosovo, Taiwan and Sudan) and no democracy values (South Sudan, Georgia, Belize).
5. Line 37: Converted the dataset from pooled cross-sectional to cross sectional by taking the mean of each column by country and aggregating this.
6. Line 41: Added a column for the region name titled “rname” in order to incorporate it into the dataset.
7. Line 45: Created dummy variables from the region name and assigned this to a new dataset titled “df\_dum”.
8. Line 48: Tested for multicollinearity using a covariance matrix, excluding regions and year from the matrix (section titled “Multicollinearity” for details).
9. Line 50: Dropped problem variables “men\_edu”, “sanitation”, “child\_mortality” & “elder\_child” found in multicollinearity test.
10. Line 56: Dropped unnecessary dummy variables to prevent multicollinearity.

**Issues with Multicollinearity**

We originally began with ten independent variables, which included all of those listed above, plus “Child Mortality”, “Percentage of Population with basic sanitation”, “Mean Years of Schooling for Men 25-34” and “Children and Elderly per 1000 Adults”. After running a correlation matrix, we saw high correlation coefficients of above .5 between these variables, although multicollinearity poses more of a problem in economics than in data analytics, we wanted a low threshold for multicollinearity in order to ensure our variables did not affect our model in any way. When deciding what variables to omit, we looked at which variables were more significant than others and took that into consideration. For example, we eliminated men’s education as opposed to women’s education since women’s education had more variability than men’s, and men’s was more static. Therefore we kept women’s as it would create a better model to guess the happiness score from.

**Bibliography**

<https://www.gapminder.org/data/>

<https://worldhappiness.report/ed/2020/>