World Population Challenge

Abhishek Hanchate, Akshay Kadu, Jaime Avila, Mayank Jaggi

***Abstract*- The challenge is to estimate the population of the given 212 countries from 2000 to 2016 based on the past data are given from the year 1960 to 1999. Lasso regression and correlation method were used to predict the population using at most 4 countries as predictors. The report explains the methodology used and network analysis explaining the relationship among various countries using Gephi.**

***Keywords*- Correlation, Regularized Least Squares, Lasso, Gephi, Linear Regression**

I. INTRODUCTION

Population growth of a country depends on multiple factors including (but not limited to) life expectancy, fertility rate, literacy, economic growth, government policy etc. To analyze the effect of the population of one country on the other, we perform Lasso Regression and Correlation Methodologies on a world population dataset by the World Bank. The data includes a population of 212 countries from the year 1960 to 2016.[1]

The goal is to identify the effect of the population of different countries on the population of the target country. This is achieved by taking a linear combination of the population of at most 4 countries in the same year.

II. REGULARIZED LEAST SQUARES METHOD

Lasso (least absolute shrinkage and selection operator) is a regression analysis method that performs both variable selection and regularization, in order to enhance the prediction accuracy and interpretability of the model. It achieves variable selection by adding a penalty of alpha*norm1 of beta. When alpha=0, we have least square approach. As we increase alpha value the solution becomes sparser that is reducing the number of countries as predictors for the given target country.[2]

Our approach is to perform a loop across the columns of the data frame wherein each column represented the country population. This is followed by looping alpha for each target country starting at 150 in steps of 50 until an apt alpha is found which is the value at which at most 4 predictors are selected. Lasso regression model is trained on the normalized data of the selected predictors and was used to predict the population of the target country from year 2000 to 2016. Three data frames were created storing the prediction values for each target country, MSE (Mean Square Error) and alpha value for each target country, and parameter coefficients of the predictors for each target country. Parameter coefficients for a sample of target countries can be seen in Table 1.

Target Country	Coeffecients	Predictor Countries
Canada	0.105092052	Argentina
	0.315375545	Australia
	0.010897269	Azerbaijan
	0.134929777	Iceland
Australia	0.067390931	Argentina
	0.397115286	Azerbaijan
	0.026012368	Canada
	0.011343564	Korea, Dem. People's Rep.
Bulgaria	-2.67411091	Aruba
	0.314152013	Bosnia and Herzegovina
	7.230366321	Curacao
	0.660382483	Hungary
Djibouti	0.002857248	Burundi
	0.481802573	Bahrain
	0.001295957	Iran, Islamic Rep.
	1.368731535	Marshall Islands

Table 1. Coefficient and Predictor countries Sample

III. CORRELATION METHOD

In statistics, correlation or dependence is any statistical relationship between two random variables, commonly referred to as the degree to which the pair of variables are related. A correlation coefficient is a numerical measure of some type of correlation often explaining the strength factor of the dependency. One of the most widely used correlations in feature selection is Pearson's Correlation. It gives a standardized value between -1 and 1 which indicates the strength of dependence between two variables. A correlation coefficient of 0 means that there is no relationship between the two variables, while a coefficient of 1 and -1 shows a perfect positive and negative relationship between the variables respectively. In feature selection, we select the 'n' variables with the highest correlation among them for a better prediction. A correlation matrix is an m-by-m square matrix or a table whose elements or cells are the pairwise correlation coefficients of 'm' vectors in $\mathbb{R}$n. When someone speaks of a correlation matrix, they usually mean a matrix of Pearson-type correlations. The diagonal elements of this matrix are 1s which indicates relationship of a variable with itself.

The problem statement here is to predict the population for each variable (country) using four other as predictors with the highest correlation coefficients. We first generate a correlation matrix using the function *df.corr()*, where df is a pandas data frame with n variables. Once we have a correlation matrix, the first step is to find at most four top predictors with strongest correlation for each country. An iterative loop such as a 'for' loop is essential in such situations as we are supposed to find top predictors for all the countries in our data set. Since the matrix also gives correlation coefficient of 1 for relationship with itself, it is important for our algorithm to ignore such coefficients and therefore we replace them with zero to begin with. One of the convenient ways to determine the top predictors would be *df.nlargest()*, a function which returns the specified highest values of a vector. These top coefficients for each country can then be stored in a data frame. These predictors can then be selected and leveraged to make our desired estimation for a country's population using linear regression.

We use the countries associated with the strongest correlation coefficients above to predict the population of its respective country as a linear combination of population of the four best predictors. The regression coefficients generated by linear regression are also stored in a data frame.

IV. INSIGHTS AND VISUALIZATION

After finding the coefficients for both the Lasso regression and correlation parameters, the data stored in the csv files were used to create a visualization graph using a software called Gephi. Gephi is a network analysis and visualization tool used for exploration and useful observations about the given data. First, the data parameters for a specific prediction method is loaded into Gephi as a csv. The options that were used to fully load the data are: import it as a matrix (due to the country names appearing in both the x and y axis), undirected graph because direction is not an important metric, and not connected, as otherwise the software would create a fully connected network even if there is no relation between countries. One of the most common steps once the data is loaded is to apply one of the layout algorithms that are available to create a graph such that the nodes are positioned to improve readability and aesthetics. The algorithm chosen was the Yifan Hu proportional, which separates the nodes based on the proportional displacement of the nodes fairly fast compared to the other layouts. There is also a wide range of statistical metrics used for calculating various aspects of the network. Some of the ones that were calculated were modularity, which measures how well a network decomposes into modular communities. For example, the Lasso regression created 10 communities with a modularity of 0.67, whereas the Correlation network has a modularity of 1.15 and created 21 communities. A higher modularity indicates more sub-systems. Figure 1 shows the results for the Lasso regression while Figure 2 shows the correlation results.

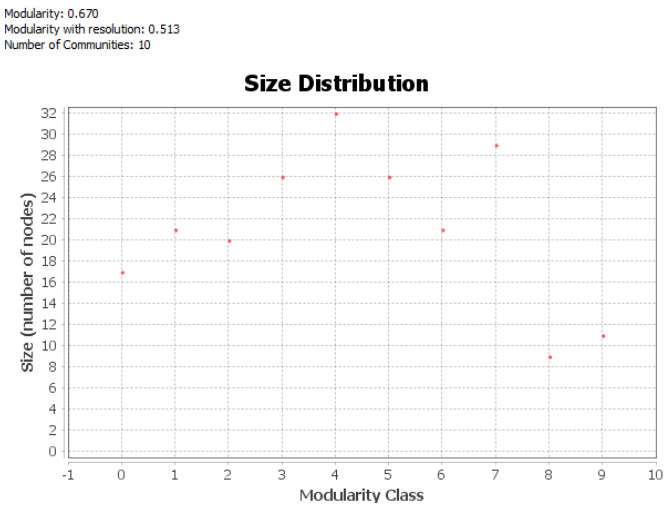

Figure 1. Graph of Modularity for Lasso

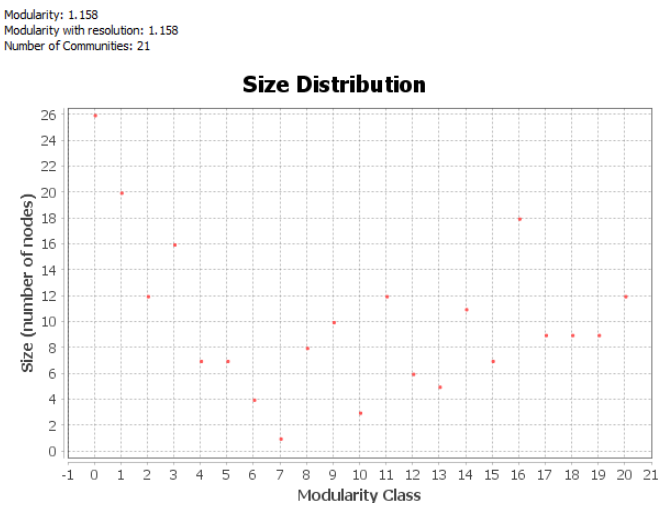

Figure 2. Graph of Modularity for Correlation

A few more settings were tuned to create a graph that is easier to analyze. Ranking was used to adjust the nodes for two options, the first option was to increase the color saturation and the second to increase the size of the nodes to indicate higher degrees. Using the modularity results, a color scheme was added to differentiate between the different types of communities that were found. The edges for each node also went through a similar process, with nodes that are thicker and more saturated in color indicate a higher weight (or how closely are two nodes together) between a particular set of nodes. Having applied all these settings, the following graphs were obtained for the Lasso and correlation methods, as shown in figures 2 and 3 (Due to the size of the networks, the nodes are hard to see in the figures, but the source file can be found in the GitHub repository along with higher resolution images). The different shapes of the graph is due different parameters between the Lasso and correlation methods.

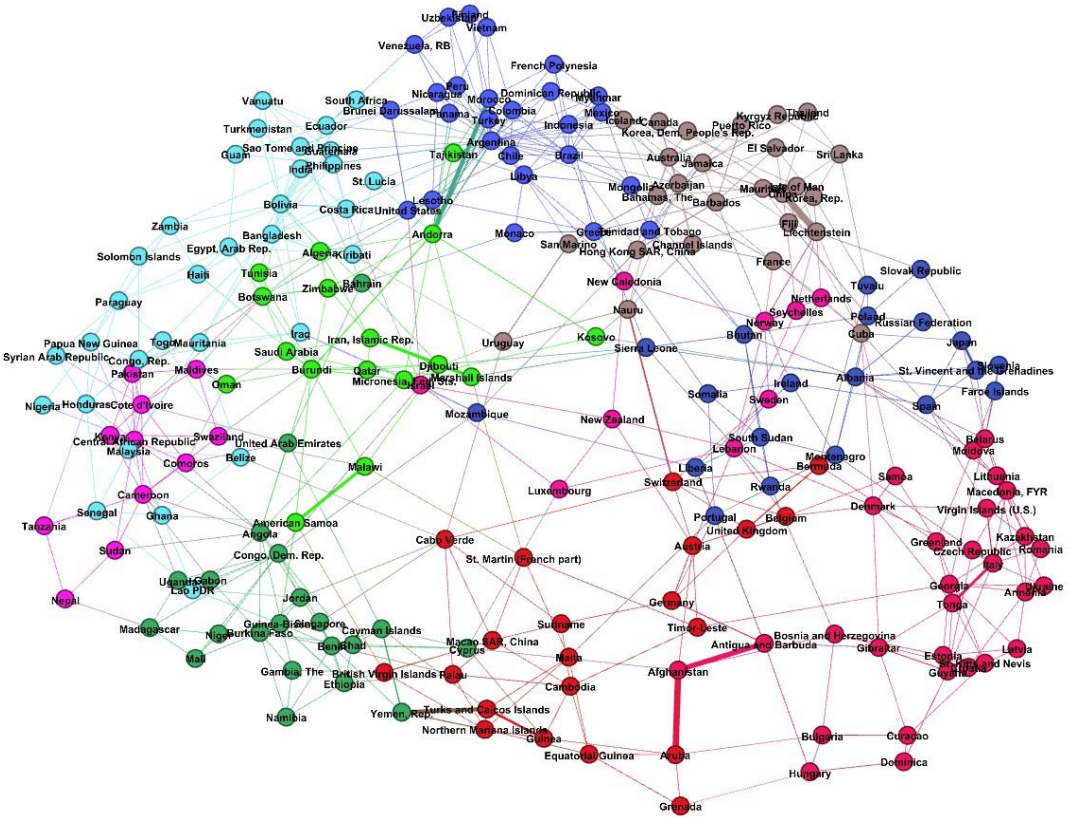

Figure 3. Lasso Regression Network

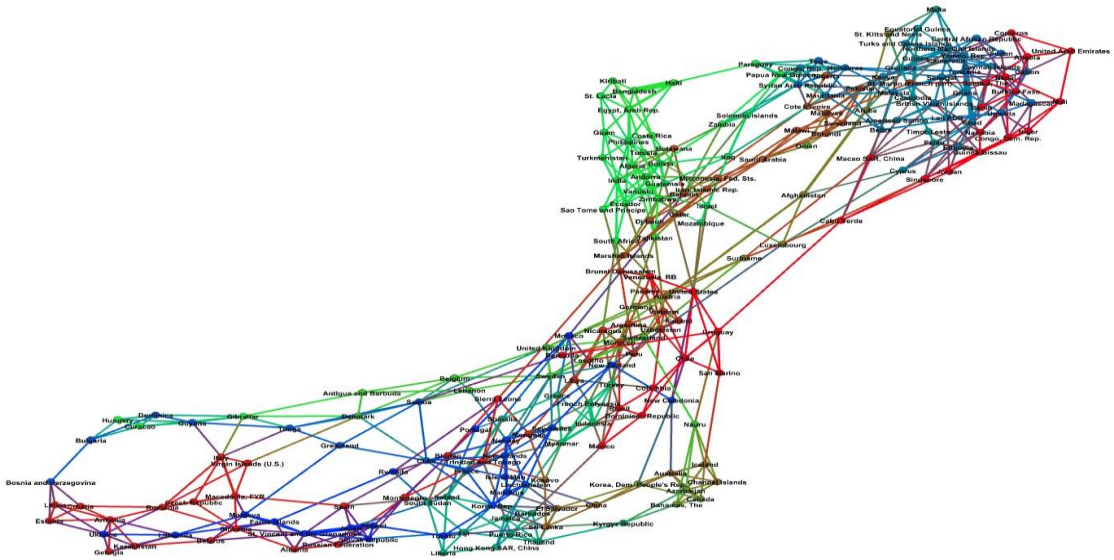

Figure 4. Correlation Network

One of the features that was of most interest in analyzing the graph is if geographical proximity played a role in how closely related two nodes are. Due to the size of the network, it is not very clear in Figure 3 and 4, but there is some correlation between a node and its neighbors, mainly the nodes in a particular neighborhood. For example, for Germany there was a correlation in the Lasso method with other European countries, like Austria and the United Kingdom. However, there were also other countries that are correlated that do not share a geographical proximity, such as Suriname. Another example in the correlation network is Colombia, which had close correlation with Brazil and Chile, but also with Turkey and Uzbekistan. A question that naturally arises is if geographical location does not play a strong correlation, then there must be other factors, such as language or GDP. Another interesting note is that how some countries have more connections to other nodes. For example, Argentina has more than 30 connections with other countries from almost all continents, whereas some have only three. In addition, some countries like Iran have a very high correlation (shown by the thicker and wider edges) with a country like Marshall Islands, but a much lower correlation with other neighboring countries like Saudi Arabia and Qatar whom they share many characteristics like religion, language, and demographics.

V. CONCLUSIONS

This challenge sought to find predict the population of a given country considering only the population of four other countries. Two different methods were used, Lasso regression and the correlation method. Another part of the assignment was to create a graphical representation of the connection between all the countries and try to understand why some countries are closely related to others. The data included population data from 1960 to 1999 for 212 countries, and predictions were made for 2000 to 2016. The graph network created with the software Gephi for both methods gave some insight into the correlation of countries. It was clear that geography played a factor, as some countries were closely correlated with neighboring ones. However, it is also important to note that this was not the defining parameter, as some countries were related with countries from other continents. This challenges serves as a basis to dive further into the analysis of correlation between countries and further research is needed to obtain some characteristics that links certain countries together.

REFERENCES

[1] World Population data by the World Bank.
Available at: https://www.kaggle.com/c/ais-world-population-challenge/data [Accessed 25 November 2019]
[2] An Introduction to Statistical Learning with Applications in R by Gareth James, Robert Tibshirani.
[3] Gephi the Open Graph Viz Platform
Available at: https://gephi.org/ [Accessed 25 November 2019]