(ECEN 689-606) Challenge 3

Insights from Population Estimation using Lasso Regression

Ki Yeob Lee, Prabhneet Arora and Rishi Laddha

I. Abstract

Graphical Representation using Gephi has been performed to provide insights on the estimation of the population of a country based on the population of (at most) five countries using Lasso Regression. Demographics, fertility rates, Life expectancy and migration are the main factors listed as motivating aspects for predicting the population of a country.

II. Introduction

Intuitively, population estimation calculation is based on calculating the population size for a year between the census period and or for the ongoing year. [1] A projection of the population size for the future date-in-time based on the historical data has been a conventional method of population estimation for years.

A similar estimation method has been carried out for 258 countries using the Lasso Regression technique, instead of using the demographics for its own country, a country's population is estimated by the population of at most five other countries.

III. Methodology

Data collection: Data for 258 countries for 1960-1999 was provided for analysis of population for the mentioned 258 countries for years 2000-2016.

Algorithm: Lasso (Least Absolute Shrinkage and Selection Operator) has been a proven regression analysis method that performs variable selection and regularization simultaneously to enhance the prediction accuracy and interpretability of statistical model it produces. [2] It was originally introduced in geophysics literature in 1986 but was independently rediscovered and popularized by Robert Tibshirani in 1996 who coined out the term and provided further insights into the observed performance. [3][4]

During that period, ridge regression was the most popular technique for improving prediction accuracy.[2] Ridge regression condenses prediction error by shrinking large regression coefficients in order to reduce overfitting, but it does not accomplish covariate selection and therefore does not make the model more interpretable. Lasso can achieve both goals by forcing the sum of the absolute value of the regression coefficients to be less than a fixed value, which forces certain coefficients to be set to zero,

effectively choosing a simpler model that does not include those coefficients using the below-mentioned equation (where λ is a hyperparameter). [2]

$$\min\left(\left|\left|Y - X\theta\right|\right|_2^2 + \lambda||\theta||_1\right)$$

From the figure 1, one can see that the constraint region defined by the $l1$ norm is a square rotated so that its corners lie on the axes, while the region defined by the $l2$ norm is a circle. [2]

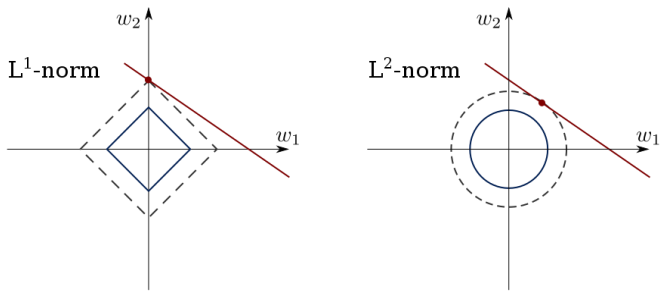

Figure 1: Forms of the constraint regions for lasso and ridge regression. [2]

Therefore, using multiple λ values as represented in the equation above, the model is tuned to provide lowest mean square error on the predictions.

Visualization: Gephi was used for the visualization of the results by importing the node and edge files into the software as separate entities. The network was further visualized for any relevance to data models in population estimation.

IV. Results

The Fruchterman Reingold layout generated by the visualization tool consisting of 258 nodes is shown in Figure 3. The layout automatically pushes the less connected nodes to the outside. A more in-depth analysis as shown in Figure 2 (left and right) proved that the estimated population of a country was not purely co-incidental using any randomly chosen 5 countries. It was rather based on the demographics of the neighboring countries. Some European countries with similar geographic conditions and demographics predicted the population of other European countries. Likewise, it was true for many other countries that fell within the same criteria.

Figure 2: Population estimation of Croatia, Bosnia & Herzegovina(left) and United Kingdom (right) by five other European countries.

Figure 3: Gephi: Fruchterman Reingold layout of 258 nodes

V. Conclusion

The results clearly proved that the demographics and geographical conditions of countries proved to be a valid point for estimation of the population of a country. It is often observed that neighboring countries share these characteristics and therefore, the results of our estimation are in parallel with our intuition. But it is certainly possible that far off countries with similar demographics too may be a good estimate for the population of a country (Figure 4).

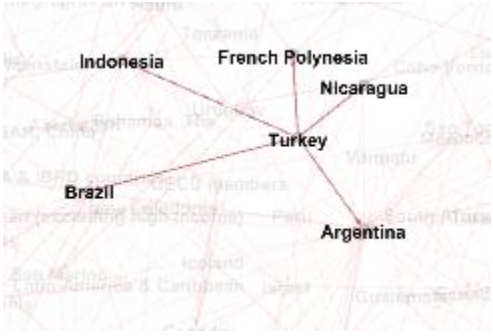

Figure 4: Estimation of Turkey's population by non-neighboring countries

VI. References

[1] Wormald, B. (2015, May 12). Main Factors Driving Population Growth. Retrieved from http://www.pewforum.org/2015/04/02/main-factors-driving-population-growth/

[2] Lasso (statistics). (2018, September 22). Retrieved October 02, 2018, from https://en.wikipedia.org/wiki/Lasso_(statistics)

[3] Tibshirani, Robert (1996). "Regression Shrinkage and Selection via the lasso". Journal of the Royal Statistical Society. Series B (methodological). Wiley. 58(1): 267–88. JSTOR 2346178.

[4] Santosa, Fadil; Symes, William W. (1986). "Linear inversion of band-limited reflection seismograms". SIAM Journal on Scientific and Statistical Computing. SIAM. 7 (4): 1307--1330.