Estimation of country's population using Lasso Regression

Michael Bass , Samarth Mistry and Samyuktha Sankaran

***Abstract*— This challenge is to estimate the population of the given 259 countries between the years 2000 and 2016 based on the data given for the years 1960 to 1999. The lasso regression was used to carry out the estimate.**

I. INTRODUCTION

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 statistical model it produces. Lasso was originally formulated for least squares models and this simple case reveals a substantial amount about the behavior of the estimator, including its relationship to ridge regression and best subset selection and the connections between lasso coefficient estimates and so-called soft thresholding. It also reveals that (like standard linear regression) the coefficient estimates need not be unique if covariates are collinear.

Though originally defined for least squares, lasso regularization is easily extended to a wide variety of statistical models including generalized linear models, generalized estimating equations, proportional hazards models, and M-estimators, in a straightforward fashion. LassoâĂŹs ability to perform subset selection relies on the form of the constraint and has a variety of interpretations including in terms of geometry, Bayesian statistics, and convex analysis.

II. DATA SET

The data set used here is the world bank data, the population of about 260 countries. The data from years 1960 to 1999 are being used to train the model and the population for the years 2000 to 2016 are being used for testing. The solution for a country is based on the best of five countries based on their coefficients.

III. RESULTS

A. Co-efficients

For each country, the five best features were selected, i.e., the other co-efficients, except these five are reduced to zero. A sample of first five countries are listed below.

B. Prediction

The test data set is fed to the trained model, to predict the population of the given countries between years 2000 and 2016. A sample of the first five countries between 2000 and 2004 are shown below.

$Country$	α_i	Features
Aruba	0.0026	Afghanistan
	-0.0033	Bulgaria
	-0.0007	Bosnia and Herzegovina
	0.5825	Grenada
	0.0196	Malta
Afghanistan	215.44	Aruba
	80.584	Antigua and Barbuda
	1.5730	Austria
	-148.589	Grenada
	9.8225	Luxembourg
Angola	0.2622	United Arab Emirates
	0.5245	Benin
	0.3133	Burkina Faso
	0.9837	Central African Republic
	1.8412	Comoros
Albania	1.5508	Bhutan
	0.00016	Europe & Central Asia
	0.6787	Lithuania
	0.0056	Poland
	0.0121	Russian Federation
Andorra	0.00042	Algeria
	0.11576	Micronesia, Fed. Sts.
	0.1050	Guam
	0.0610	Marshall Islands
	0.00216	Tajikistan

TABLE I
A. COEFFICIENTS

$Country$	Estimate	Year
Aruba	90853	2000
	92898	2001
	94992	2002
	97017	2003
	98737	2004
Afghanistan	20093756	2000
	20966463	2001
	21979923	2002
	23064851	2003
	24118979	2004
Angola	16440924	2000
	16983266	2001
	17572649	2002
	18203369	2003
	18865716	2004
Albania	3089027	2000
	3060173	2001
	3051010	2002
	3039616	2003
	3026939	2004
Andorra	65390	2000
	67341	2001
	70049	2002
	73182	2003
	76244	2004

TABLE II
B. PREDICTION

IV. VISUALIZATION

Data visualization is viewed by many disciplines as a modern equivalent of visual communication. It involves the creation and study of the visual representation of data.

To communicate information clearly and efficiently, data visualization uses statistical graphics, plots, information graphics and other tools. Numerical data may be encoded using dots, lines, or bars, to visually communicate a quantitative message. Effective visualization helps users analyze and reason about data and evidence. It makes complex data more accessible, understandable and usable. Users may have particular analytical tasks, such as making comparisons or understanding causality, and the design principle of the graphic (i.e., showing comparisons or showing causality) follows the task. Tables are generally used where users will look up a specific measurement, while charts of various types are used to show patterns or relationships in the data for one or more variables.

Here, each node represents a country and each country is connected to its five best features (i.e., countries). Figure 2 and 3 shows the distribution for countries in the African and European continents respectively. Figure 1 shows the distribution for the entire data set distributed geographically.

We added the latitudes and longitudes for every country given in the data set to see if we could get any inference. We plotted the graph using the GeoLayout in Gephi but could not say anything definitively as there are connections all over the place. To see how it follows for all countries of a particular continent, we added an attribute continent to every node to filter the nodes in Gephi. Where,

- Avg. Degree is the average no. of arcs incident to a node in the graph.
- Network Diameter is the average graph distance between all pairs of nodes. Connected nodes have a graph distance of 1. The diameter is the longest graph distance between any two nodes in the network. (i.e. How far apart are the two most distant nodes).
- Network Density measures how close the network is to 'complete'. A complete network has all possible connections and its density is 1.
- Avg. Clustering Coefficient gives an overall indication of the clustering in the network. It indicates how nodes are embedded in their neighborhood.

We see the graph of North American and European countries along with the groupings where we see clear clusters of High-income countries of Europe the Caribbean, Middle-income countries and Low-income country groupings. And the graph of countries from Asia, South America, Africa and Oceania countries exhibit good cohesiveness.

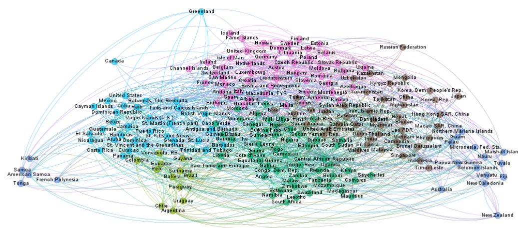

Fig. 1.   World

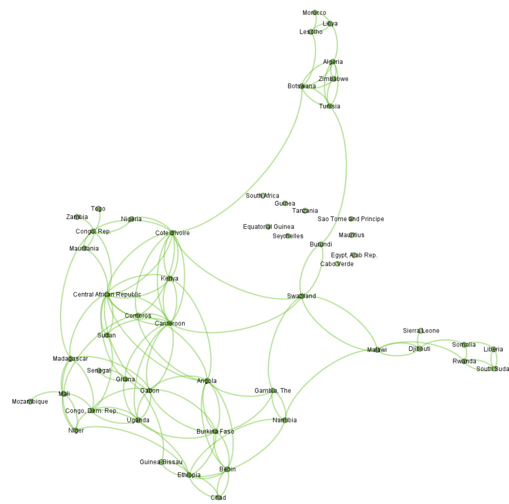

Fig. 2.   Africa

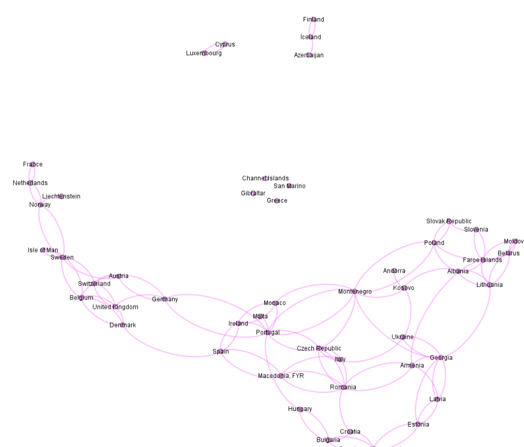

Fig. 3.   Europe

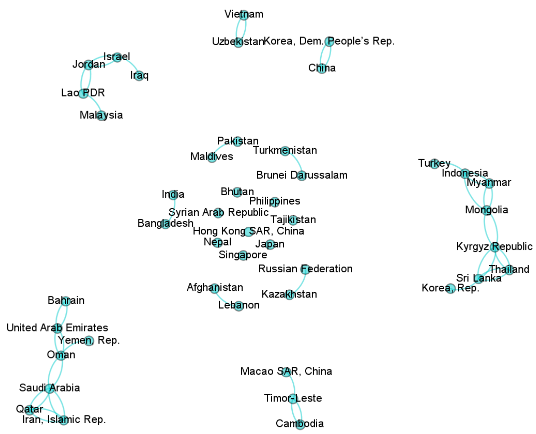

Fig. 4. Asia

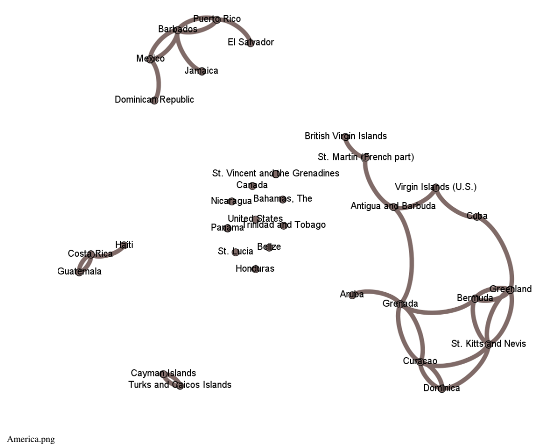

Fig. 5. North America

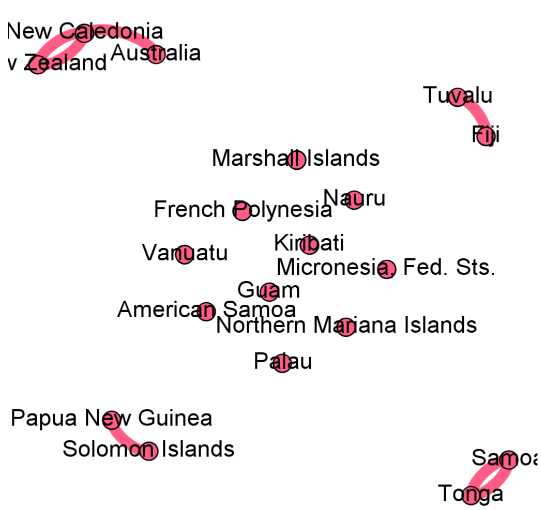

Fig. 6. Oceania

X	Avg. degree	Network diameter	Graph Density	Avg.clustering coefficient
Entire graph	5	11	0.019	0.206
Africa	2.321	10	0.045	0.183
Europe	2.060	15	0.042	0.093
Asia	1.022	6	0.023	0.115
North America	0.939	4	0.019	1.797
Oceania	0.368	2	0.020	0.000
South America	1.250	4	0.114	0.235
Asia + South America Africa + Oceania	3.147	15	0.025	0.205
Asia + South America + Africa + Oceania with Groupings	3.897	18	0.022	0.244
North America + Europe	2.554	11	0.031	0.084
North America + Europe with Groupings	3.047	23	0.024	0.193
Groupings only	1.891	18	0.042	0.330

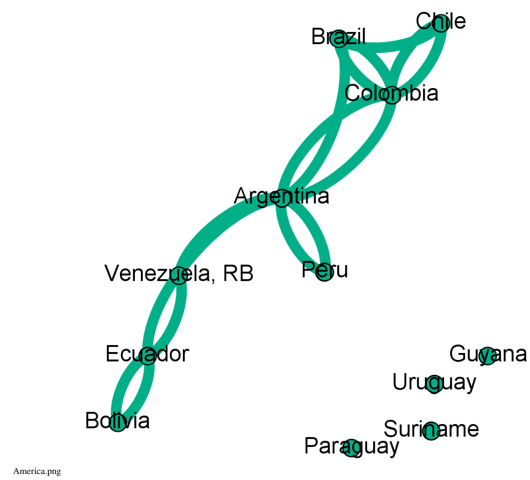

Fig. 7. South America

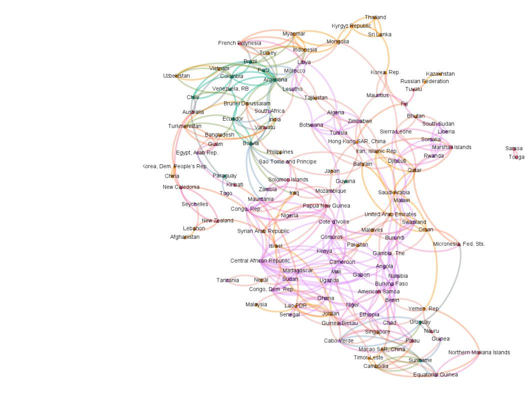

Fig. 8. Asia, South America, Africa and Oceania

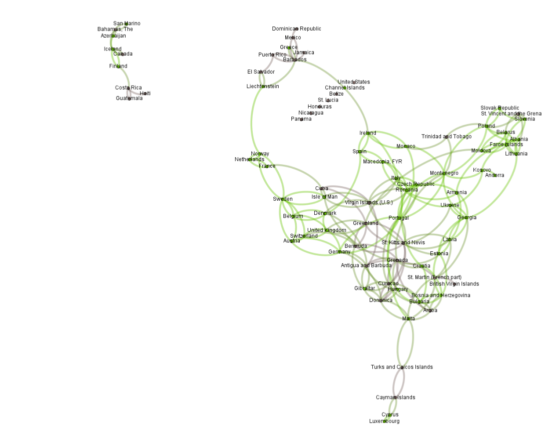

Fig. 9. North America Europe

V. REGULARIZATION STRATEGY

Lasso regularization achieves its goal by adding to the least squares minimization an alpha*norm1 of Beta. When alpha approaches 0, the solution is equal to that of least squares. As alpha increases, the solution becomes sparser. However, for equations, different alpha values will yield varying degrees of sparsity. This lead us to realize that we must search the alpha space for the alpha value that meets our constraints and yields the most optimal solution.

Our approach was to perform an initial fast coarse search followed by an iterative fine-grained search. For each country we created a line space of alphas from 0 to 5000, with 25 points. We performed lasso regression with this set of alphas and evaluated the sparsity of the results. Given our constraint of using only 5 countries we must find an alpha the correlates with 5 non-zero coefficients. If there were no acceptable solutions from the alphas between 0 to 5000, we then exponentially increase our search space as follows:

- 0 to 5000
- 5000 to 10000
- 10000 to 20000
- 20000 to 40,000
- etc.

Once we find sparsity values of 5, we evaluate each alpha that results in 5 coefficients, and determine which yields than minimum mean squared error (MSE) and call this our current alpha. Next, we proceed to the fine grain search.

For the fine-grained search, we refer to our line space, and select the alpha to the left of the current alpha as the leftmost bound of the next line space, and the alpha to the right of the current alpha as the rightmost bound of the next line space. Then, we create a line space between these points with 100 points. Again, we apply the lasso regression, and select the optimal alpha. We perform this fine-grained search iteratively and can do this as many times as needed. For our results, we applied the fine-grained search twice.

We had to modify the coarse search slightly as sometimes no results had a sparsity of 5. As the search space increases, and since the coarse search uses only 25 points per line space, the distance between points increases. As a result, it may be possible for the result to include sparsity greater than 5 and sparsity less than 5, but no 5. If this happens, we narrow in on the region that surpassed 5 until we find an alpha with sparsity of 5.

A. Alternate Approaches

We had to modify the coarse search slightly as sometimes no results had a sparsity of 5. As the search space increases, and since the coarse search uses only 25 points per line space, the distance between points increases. As a result, it may be possible for the result to include sparsity greater than 5 and sparsity less than 5, but no 5. If this happens, we narrow in on the region that surpassed 5 until we find an alpha with sparsity of 5.

B. Ridge Regression

Ridge regression belongs a class of regression tools that use L2 regularization. The other type of regularization, L1 regularization, limits the size of the coefficients by adding an L1 penalty equal to the absolute value of the magnitude of coefficients. This sometimes results in the elimination of some coefficients altogether, which can yield sparse models. L2 regularization adds an L2 penalty, which equals the square of the magnitude of coefficients. All coefficients are shrunk by the same factor (so none are eliminated). Unlike L1 regularization, L2 will not result in sparse models.

A tuning parameter (λ) controls the strength of the penalty term. When $\lambda = 0$, ridge regression equals least squares regression. If $\lambda = \infty$, all coefficients are shrunk to zero. The ideal penalty is therefore somewhere in between 0 and ∞.

VI. CONCLUSION

The lasso has a major advantage over ridge regression, in that it produces simpler and more interpretable models that involve only a subset of the predictors. Lasso leads to qualitatively similar behavior to ridge regression, in that as Îż increases, the variance decreases and the bias increases.The lasso implicitly assumes that a number of the coefficients truly equal zero. Consequently, it is not surprising that ridge regression outperforms the lasso in terms of prediction error in this setting. In general, one might expect the lasso to perform better in a setting where a relatively small number of predictors have substantial coefficients, and the remaining predictors have coefficients that are very small or that equal zero. Ridge regression will perform better when the response is a function of many predictors, all with coefficients of roughly equal size. However, the number of predictors that is related to the response is never known a prior for real data sets. As with ridge regression, when the least squares estimates have excessively high variance, the lasso solution can yield a reduction in variance at the expense of a small increase in bias, and consequently can generate more accurate predictions. Unlike ridge regression, the lasso performs variable selection, and hence results in models that are easier to interpret. There are very efficient algorithms for fitting both ridge and lasso models; in both cases the entire coefficient paths can be computed with about the same amount of work as a single least squares fit. We will explore this further in the lab at the end of this chapter.

REFERENCES

[1] An Introduction to Statistical Learning with Applications in R by Gareth James and Daniela Witten.

[2] Understanding Machine Learning: From Theory to Algorithms by Shai Ben-David and Shai Shalev-Shwartz.