Volume- 6
Issue- II
December- 2021

Rajarata University Journal

ISSN: 2362-0080
www.ruj.ac.lk/journals/

Category: Research Articled

Identification of Determinants of Life Expectancy at Birth across Nations Using Machine Learning Techniques

De Mel WAR & *Samarakoon NYJW

Department of Mathematics, Faculty of Science, University of Ruhuna, Matara, Sri Lanka

ARTICLE DETAILS

Article History

Published Online: 31/12/2021

Keywords

Ridge, Lasso, Principal Components Regression, Partial Least Squares Regression

[*]Corresponding Author

Email: neminisamarakoon95@gmail.com

ABSTRACT

Life expectancy at birth (LEB) gives implications regarding the overall development of a nation. So identification of prominent determinant factors that affect LEB, will lead to take relevant decisions regarding the development of a nation. Studies have been conducted to identify prominent determinant factors of LEB, using ordinary least squares procedure in linear regression models with a limited number of determinant factors. Problems regarding multicollinearity, prediction accuracy and model interpretation occur when using this procedure with a wider range of determinant factors. The machine learning techniques: shrinkage and dimensionality reduction techniques were applied to overcome these problems. 17 determinant factors were identified and applied to data obtained for 193 countries of United Nations Agencies for the year 2016. As shrinkage techniques ridge and lasso regression and as dimensionality reduction techniques principal components regression and partial least squares regression were applied. These regression techniques were compared concerning mean squared error, goodness of fit and ranking based on regression coefficient estimates. Ridge regression model turned out to be the best model with a good fit for data on hand, because it has the highest adjusted R^2 for the training data. Lasso regression model shows the highest adjusted R^2 and lowest mean squared error for the test data. So lasso regression model is the best predictive model.

1. Introduction

As the main factors for sustainable development, maintaining, expanding and improving the health of a nation can be considered [1]. Life expectancy at birth (LEB) plays a vital role in each of these instances. According to the United Nations Human Development Report, LEB is the number of years a newborn infant could expect to live if the prevailing patterns of age-specific mortality rates at the time of birth stay the same throughout the infant's life [2]. The level of LEB gives important implications for individual and aggregate human behavior, because it affects fertility behavior, economic growth etc. [3]. Improvements in LEB will lead to the improvements in areas like economy, health, sanitation, etc. [2]. So identification of the most prominent factors that affect LEB is an essential task because then on relevant decisions can be taken to increase LEB and in turn gain an overall development of the country.

A recent study has been conducted in understanding the impact of demographic changes: socioeconomic inequalities and the availability of health factors on LEB in 91 low and lower middle income countries in 2012 [1]. Total fertility rate, adolescent fertility rate, mean years of schooling, gross national income per capita, physician density and HIV prevalence rate are the determinant factors that were considered in the study. By applying stepwise multiple regression analysis, HIV prevalence rate, adolescent fertility rate and mean years of schooling were identified as the prominent factors that affect LEB in this study. This study suggests considering a wider range of determinant factors for the purpose of analysis.

To identify the prominent determinant factors that affect LEB, multiple linear regression models can be used. This can be done by considering LEB as the response variable and its determinant factors as predictor variables. When a wider range of determinant factors are considered in this instance problems regarding multicollinearity, prediction accuracy and model interpretation may occur [4]. So, a wider range of determinant factors of life expectancy at birth were considered in this study with the aim of overcoming problems that may arose due to a larger number of determinants.

The major objective in this study is to apply shrinkage and dimensionality reduction regression techniques to build regression models which identify determinant factors that affect LEB. The purpose of using these regression techniques is to overcome the problems mentioned. The other objective is to compare the shrinkage and dimensionality reduction methods using the regression models.

2 Material and Methods

2.1 Multiple Linear Regression

Multiple linear regression is the linear approach modeling the relationship between a dependent variable and more than one independent variable [5]. The multiple linear regression model can be expressed as follows,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

y – Dependent (Response) variable

$x_1, x_2, \cdots, x_p$ – Independent (Predictor) variables

$\beta_0, \beta_1, \cdots, \beta_p$ - Regression parameters

ε - Error term

p - No. of predictor variables

2.2 Shrinkage Techniques

In shrinkage techniques a multiple linear regression model is fitted with all predictors using a technique that constrains or regularizes the coefficient estimates towards zero. Depending on the method of shrinkage, two approaches ridge and lasso were considered.

- Ridge Regression

The idea of ridge regression was formulated by Hoerl and Kennard in 1970 [6]. Ridge regression shrinks the regression coefficients by imposing a penalty term to the residual sum of squares [7]. To obtain ridge estimates here a penalized residual sum of squares is minimized:

$$\hat{\beta}_{lasso} = argmin_\beta \left(\left(y_i - \beta_0 - \sum_{i=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right)$$

λ - Tuning parameter

Where $\lambda \geq 0$ is the tuning parameter (complexity parameter) that controls the amount of shrinkage. When the tuning parameter is 0 ($\lambda = 0$) the penalty term has no impact and the ridge estimates will be equivalent to least squares estimates. When $\lambda \to \infty$ the effect of ridge penalty increases and the ridge estimates approach to zero. The best value for λ is selected using cross validation [7].

The penalty term shrinks all the regression coefficients to zero, but not exactly to zero (except for $\lambda = \infty$), resulting all the predictors to include in a regression model. This will cause problems in model interpretation when the number of predictor variables is quite large.

- Lasso Regression

Least Absolute Shrinkage and Selection Operator (Lasso) was first formulated by Robert Tibshirani in 1996 [8]. The lasso is an alternative shrinkage method to the ridge regression that overcomes its disadvantage of model interpretation. The lasso estimate is given by,

$$\hat{\beta}_{lasso} = argmin_\beta \left(\left(y_i - \beta_0 - \sum_{i=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

λ - Tuning parameter

The L$_2$ ridge penalty $\sum_{j=1}^{p} \beta_j^2$ is replaced by the L$_1$ lasso penalty $\sum_{j=1}^{p} |\beta_j|$. Like ridge regression, lasso regression shrinks the coefficient estimates towards zero. Due to the L$_1$ penalty, a variable selection is performed, setting some regression coefficients exactly to zero [7].

Ridge and lasso yields a reduction in variance of the coefficient estimates at the cost of some bias, solving problems regarding prediction accuracy [9]. And problems regarding multicollinearity are reduced due to the penalty term in both occasions [10]. Unlike ridge regression, lasso gives models that are much better regarding interpretation due to its variable selection property [10].

2.3 Dimensionality Reduction Techniques

In dimensionality reduction techniques, the original predictors $x_1, x_2, \cdots, x_p$ are transformed to reduce the dimension of data and these transformed variables $z_1, z_2, \cdots, z_M$ are used to fit a least squares regression model.

Here $z_1, z_2, \cdots, z_M$ are $M(< p)$ linear combinations of the original p predictors [4]:

$$z_k = \sum_{j=1}^{p} \phi_{jk}\, x_j$$

$\phi_{1k}, \phi_{2k}, \cdots, \phi_{pk}$ are constants for $k = 1, \cdots, M$. Then least squares procedure can be used to fit the regression model:

$$y_i = \theta_0 + \sum_{k=1}^{M} \theta_k z_{ik} + \epsilon_i$$

[4], where $i = 1, \cdots, n$.

Here $\theta_1, \theta_2, \cdots, \theta_M$ are the regression coefficients. According to how the linear combinations $z_1, z_2, \cdots, z_M$ are constructed (i.e. the selection of ϕ_{jk}), the regression technique can be principal components regression or partial least squares regression.

- Principal Components Regression (PCR)

The first M components are derived by principal component analysis as in the above paragraph. The first principal component is the direction which the observations vary most. The second is the direction with the largest variance subjected to been uncorrelated with the first component. In this way we can compute up to p distinct independent principal components. The number of principal components (M) to retain can be done using cross validation. Then the components $z_1, z_2, \cdots, z_M$ are considered as predictors to build a linear regression model using least squares method.

Principal components regression makes it easier to build regression models using $M(< p)$ predictors. Feature selection does not occur here because each component is a linear combination of the original predictors.

In principal components regression, the response is not used in identifying the components. Although the components explain the predictors well they may not be the best components for predicting the response.

- Partial Least Squares Regression (PLSR)

Partial least squares regression is a supervised alternative to principal components regression [4]. This technique finds components that interpret both the response and the predictors.

To compute the first component z_1, the p predictors must be standardized first. Then, each ϕ_{j1} is set as the coefficient from simple linear regression of y on to x_j. In partial least squares regression, when constructing $z_1 = \sum_{j=1}^{p} \phi_{j1} x_j$, more weight is given on variables which are strongly correlated with the response.

To compute the second component each variable must be adjusted for z_1. This can be done by regressing each and every variable on z_1 and computing residuals. The residuals obtained are the remaining information that has not been explained by z_1. So using the orthogonalized data, z_2 is computed in a similar manner of computing z_1 using the initial data. This process can be repeated to calculate other components up to z_M. The number of components (M) to retain can be chosen from cross validation. Finally least squares procedure is used to fit a linear model using $z_1, z_2, \cdots, z_M$.

3 Results and Discussion

Seventeen determinant factors of LEB were identified according to the availability of data. The relevant data for LEB and determinant factors were obtained for 193 countries of United Nations Agencies for the year of 2016. Data were obtained from UNESCO Institute for Statistics, World Bank Open Data, World Health Organization Data Platform, Our World in Data, Socioeconomic Data and Applications Center (SEDAC), KOF Swiss Economic Institute and Gapminder Data (accessed online). Variable description is included in Table 1. Rstudio statistical software was used for the data analysis.

Table 1: Variable description and sources

Variable	Description
LEB	Average number of years that a newborn is expected to live if current mortality rates continue to apply.
Total fertility rate	Average number of children that would be born alive to a woman during her lifetime if she were to pass through her childbearing years.
Infant mortality rate	Number of deaths of children under one year of age per 1000 live births.
Gross National Income (GNI) per capita	Sum of value added by all resident producers plus any product taxes not included in the valuation of output plus net receipts of primary income per capita converted to international dollars using purchasing power parity rates.
Population growth	Increase in the number of individuals in a population annually as a percentage.
Urban population percentage	Number of people living in urban areas as a percentage of total population as defined by national statistical offices.

Table 1: Contd.

Table 1: Contd.

Human Development Index (HDI)	Measure of a country's development. Formulated considering the indicators: life expectancy, education and per capita income.
Food production index	Measure of agricultural output and productivity.
Mean years of schooling	Mean years of total schooling across all education levels for adult population.
Unemployment rate	Number of people who are unemployed as a percentage of the labor force.
Health expenditure per capita	Expenditures on healthcare goods and services expressed in international dollars at purchasing power parity (PPP).
CO_2 emissions	Tones of CO_2 emitted per person based on territorial emissions.
Environmental Performance Index (EPI)	Scores calculated based on performance on high-priority environmental issues on protection of human health and protection of ecosystems.
Social globalization index	Index formulated considering data on personal contact, information flows and cultural proximity.
Political globalization index	Index formulated considering embassies in country, membership in international organizations, participation in U.N. security council missions and international treaties U.N.
Immunization DPT	Percentage of children ages 12-23 months who received DPT vaccinations before 12 months or a any time before the survey.
Gini coefficient	Measure of the inequality of income or wealth of a certain country.
Daily alcohol intake	Alcohol average daily intake in grams among drinkers with 95% CI

Correlation matrix heatmap (Figure 1) was plotted to see the pairwise correlation among the response and each predictor variable and pairwise correlation among predictor variables. As pairwise correlation among most of the predictor variables exceeds 0.8, a problem of multicollinearity is visualized [11]. This is confirmed further by VIF values of each predictor variable.

Due to the presence of multicollinearity, shrinkage and dimensionality reduction techniques were applied. Before applying these techniques the data were randomly split to training and test samples and standardized.

3.1 Ridge Regression

A range of 41 values from 10^{-2} to 10^{2} were selected for λ the tuning parameter. For each λ value a regression model was built. Figure 2 was created to take an idea on how the regression coefficient estimates vary with λ in each regression model built. In the figure the numbers 1 – 17 depict the predictor variables: "Total Fertility rate", "Infant Mortality rate", "GNI", "Population growth", "Urban population percentage", "HDI", "Food production index", "Mean years of schooling", "Unemployment rate", "Health expenditure per capita", "CO_2 emissions", "EPI", "Social globalization index", "Political globalization index", "Immunization DPT", "Gini coefficient", " Daily alcohol intake" respectively. The figure depicts that as λ increases, the regression coefficient estimates of these predictor variables approach to zero and none of the regression coefficient estimates is exactly zero. Next a best λ value which gives the minimum mean squared error for its model was selected as in the following approach.

The cross-validated estimate of mean squared error (MSE) versus $\log \lambda$ value was plotted (Figure 3). The number of non-zero coefficients for a given $\log \lambda$ is shown in the upper part of the plot, here it is 17 for each value. The grey bars at each point show MSE plus and minus one standard error. The first dashed line shows the location of the minimum mean squared error and the second shows the point selected by one standard error rule. Here $e^{-3.914395} = 0.01995262$ is selected as the best value for λ which is the λ value which gives the minimum mean squared error.

Ridge regression model is refitted using the λ value selected from cross validation. The variable importance plot (Figure 4) was plotted to identify the relative importance of predictors of life expectancy at birth based on the regression coefficients. According to Figure 4, the ridge regression model has following implications regarding the importance of variables on LEB. "Infant Mortality rate" is the variable which has the most negative importance on LEB. "Human Development Index" is the variable which has the most positive importance on LEB. Variables with least importance are closer to zero but not zero. The variable which has the least importance on LEB is "Political Globalization Index".

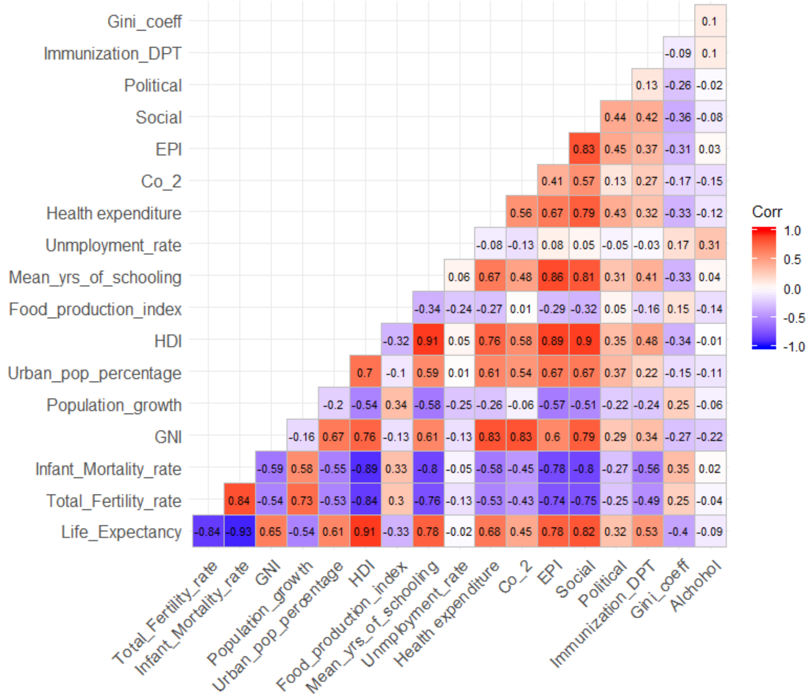

Figure 1: Correlation matrix heatmap

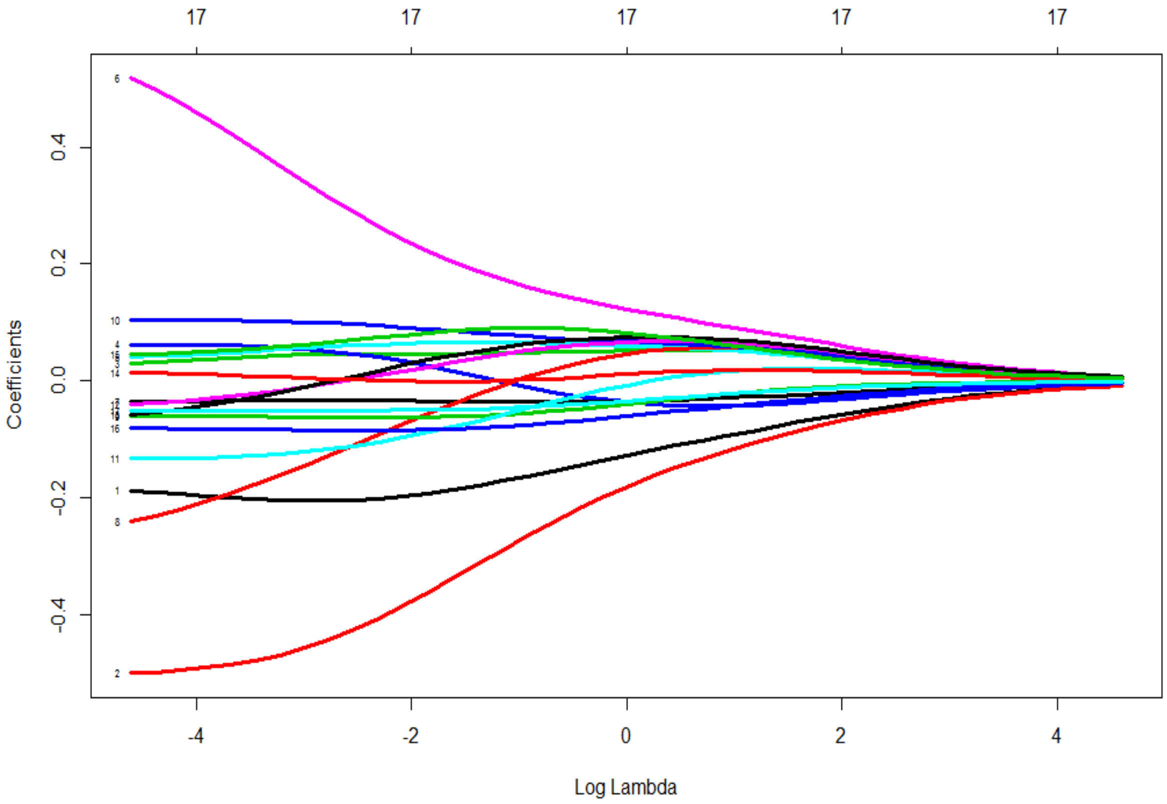

Figure 2: Plot of ridge regression coefficient estimates versus λ.

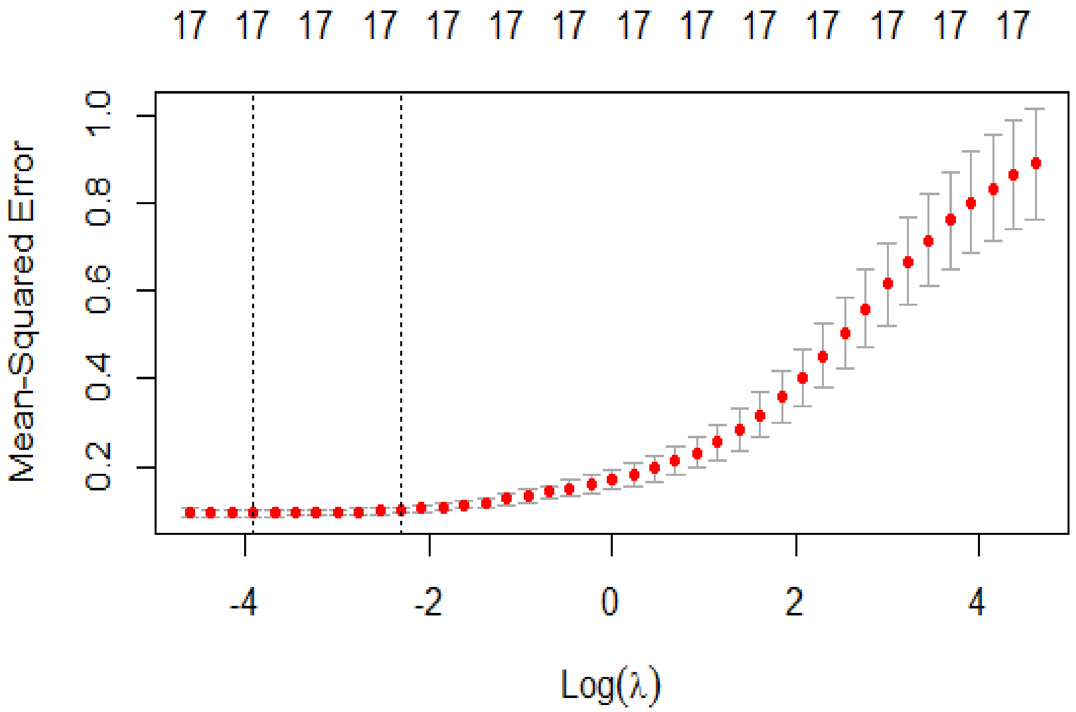

Figure 3: Cross-validated estimate of the mean squared prediction error for ridge as a function oflog λ.

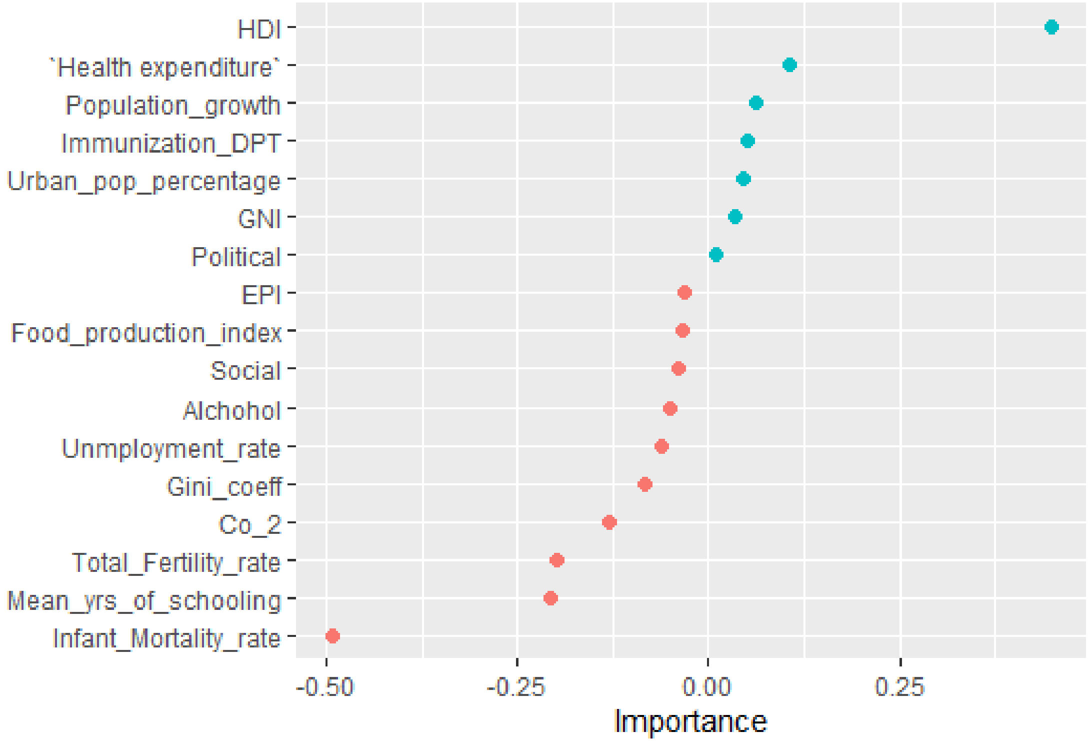

Figure 4: Variable importance plot for ridge regression

3.2 Lasso Regression

All the steps performed in ridge regression were also performed in lasso regression. The regression coefficient estimates vary with λ as in Figure 5. In the figure the numbers 1 – 17 depict the predictor variables:"Total Fertility rate", "Infant Mortality rate", "GNI", "Population growth", "Urban population percentage", "HDI", "Food production index", "Mean years of schooling", "Unemployment rate", "Health expenditure per capita", "CO_2 emissions", "EPI", "Social globalization index ", "Political globalization index ", "Immunization DPT", "Gini coefficient", " Daily alcohol intake" respectively. Just like in ridge regression, as λ increases, the regression coefficient estimates approach to zero. Also variable selection performed by lasso is depicted in Figure 5.

The cross-validated estimate of mean squared error (MSE) versus $\log \lambda$ value was plotted (Figure 6). The best value for λ selected by cross validation was $e^{-4.60517} = 0.01$.

With the selected value for the tuning parameter the lasso model was refitted. The relative importance of predictor variables is depicted as in Figure 7. Just like in ridge regression model, the variables with most positive and negative importance in lasso regression model are “Human Development Index” and “Infant Mortality rate” respectively. The variables which had least importance were dropped in this model: "EPI", "Social globalization index ", "Political globalization index " and “GNI”.

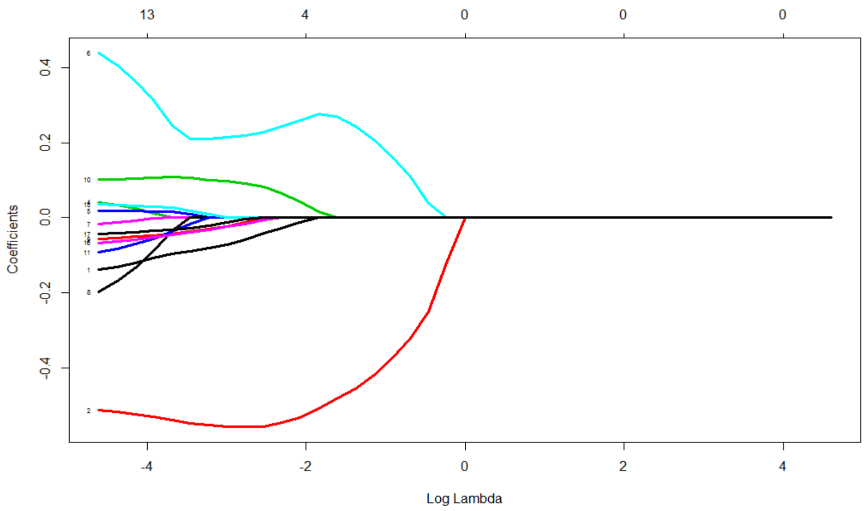

Figure 5: Plot of lasso regression coefficient estimates versus λ.

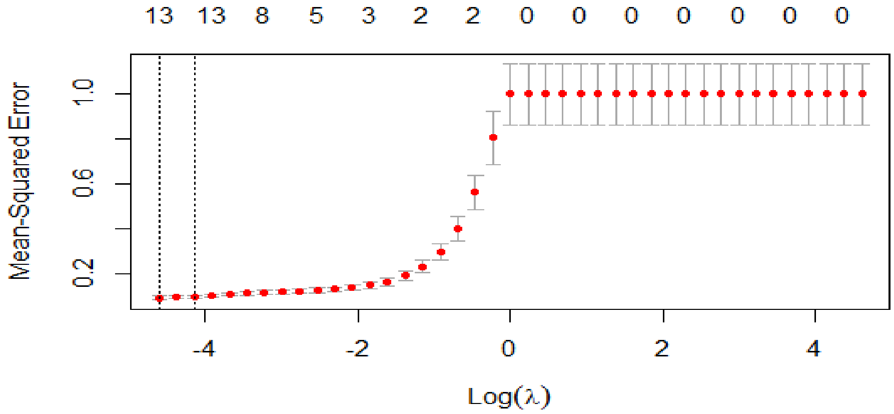

Figure 6: Cross-validated estimate of the mean squared prediction error for lasso as a function of $\log \lambda$.

3.3 Principal Components Regression

First the 10 fold cross validation error for each principal component was calculated. The possible number of components to be calculated goes up to 17 (the number of predictors). The proportion of variance explained in the predictors and in the response, by each number of components retained was calculated (Table 2). This percentage of variance is the quantity of information explained about the predictors or response using the number of principal components retained.

To select the number of components to retain, the *selectncomp* command which uses the randomization method in the pls package was used. Here in this method cross validation is used to get the results. The number of components to be retained was 12. Figure 8 shows the cross validation plot.

Finally, the model was fitted using 12 components. The relative importance of predictor variables are depicted as in Figure 10. The predictor variable with highest importance on LEB in this model is “Infant mortality rate”. And the one with least importance is “Political globalization index”

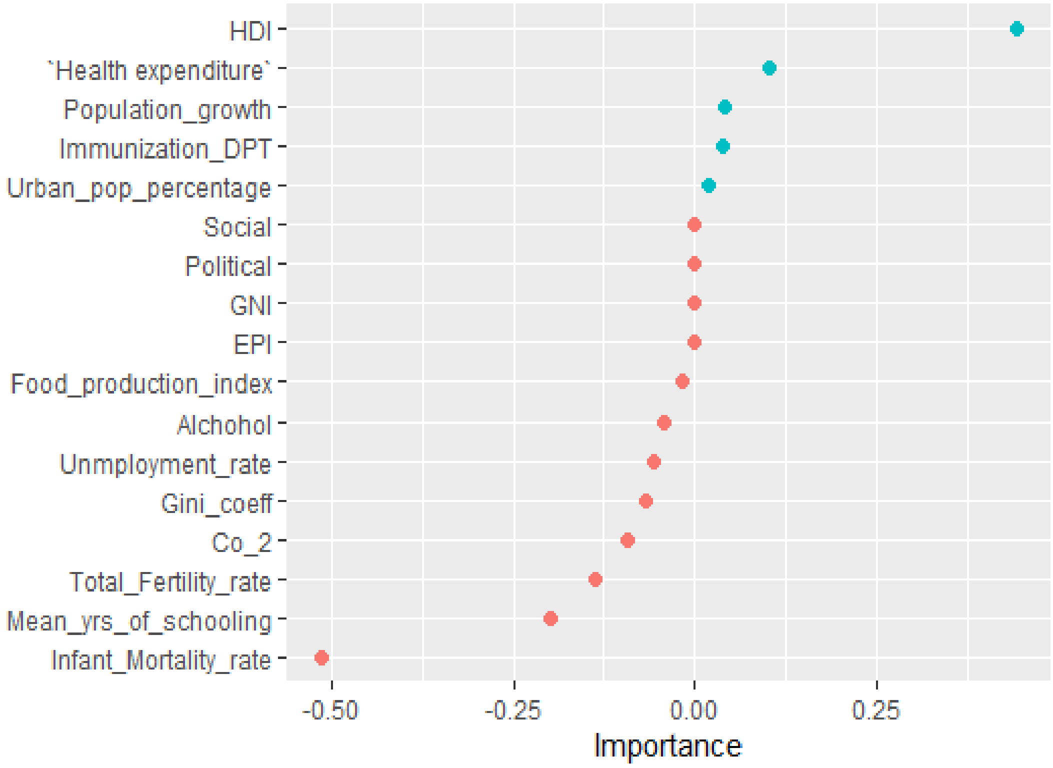

Figure 7: Variable importance plot for lasso regression

Table 2: Proportion of variance explained in predictors and response in principal components regression

No. of components	1	2	3	4	5	6	7	8	9
X	46.88	58.64	66.11	72.37	77.85	82.70	87.04	89.96	92.53
LEB	78.78	78.96	79.31	82.05	83.09	83.09	83.25	83.60	85.34

No. of components	10	11	12	13	14	15	16	17
X	94.67	96.12	97.29	98.12	98.79	99.44	99.80	100.00
LEB	85.34	87.96	91.28	91.86	92.16	92.25	92.38	93.25

3.4 Partial Least Squares Regression

In building the regression model from partial least squares regression, the same procedure used in principal components regression was used. The proportion of variance explained in the predictors and in the response, by each number of components retained is as in the Table 3.

The number of components selected was 4 and the cross validation plot is given in Figure 9.

The partial least squares model was fitted using 4 components and the variable importance is given in Figure 10. According to partial least squares regression model, the predictor variable with highest importance is “Infant mortality rate”. The variable with least importance is “Environmental performance index”.

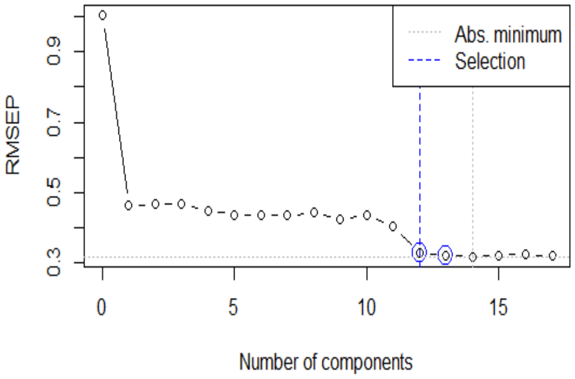

Figure 8: Cross validation plot for principal components regression

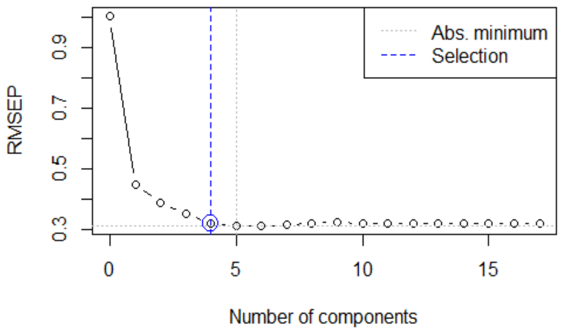

Figure 9: Cross validation plot for partial least squares regression

Table 3: Proportion of variance explained in predictors and response in partial least squares regression.

No. of components	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
X	46.83	52.97	60.09	66.51	71.04	75.62	79.64	82.3	85.62
LEB	80.50	87.70	90.52	92.10	92.56	92.77	92.92	93.1	93.17
No. of components	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	
X	87.57	89.82	91.30	95.51	96.58	97.75	99.35	100.00	
LEB	93.23	93.24	93.25	93.25	93.25	93.25	93.25	93.25	

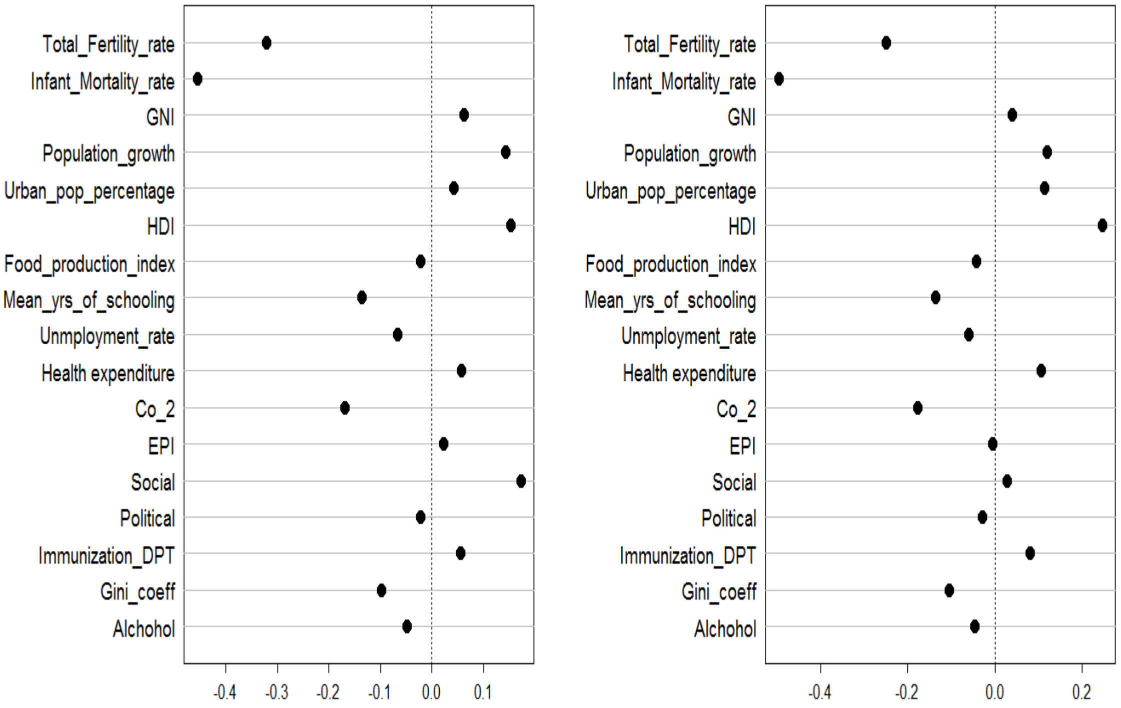

Figure 10: Variable importance plot for principal components regression (left) and partial least squares regression (right)

3.5 Comparison of the Regression Techniques

The regression coefficient estimates can be plotted as in Figure 11 for the regression techniques to get a better idea of the variability of the coefficient estimates for each technique through visualization. The predictor variable index is in the following order: “Total Fertility rate", "Infant Mortality rate", "GNI", "Population growth", "Urban population percentage", "HDI", "Food production index", "Mean years of schooling", "Unemployment rate", "Health expenditure per capita", "CO_2 emissions", "EPI", "Social globalization index ", "Political globalization index ", "Immunization DPT", "Gini coefficient", " Daily alcohol intake". The predictor variables with least variability in coefficient estimates are “Unemployment rate” and “Daily alcohol intake”. The predictor variable with highest variability in coefficient estimates is “Human Development Index”.

To check the performance of each regression model mean squared error of prediction of LEB and goodness of fit of the regression models were calculated. R^2 and Adjusted R^2 were considered as measures of goodness of fit and these were calculated for training and test data. Mean squared error was calculated for test data. The results of the calculations performed are given in Table 4.

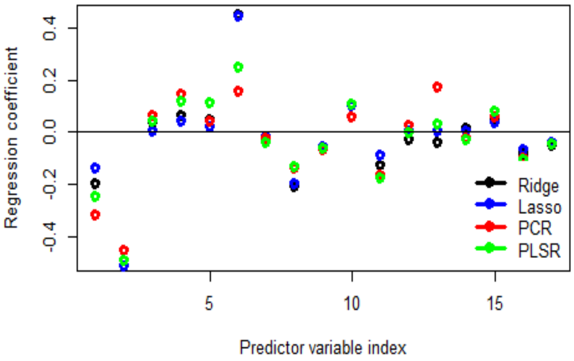

Figure 11: Plot of regression coefficient estimates for each regression technique

Table 4: Adjusted R^2, R^2 and Mean squared error for each regression technique

	Ridge	Lasso	PCR	PLSR
Adjusted R^2 (Training)	0.9200692	0.9193041	0.8996973	0.9091508
R^2 (Training)	0.9305217	0.9273737	0.9128138	0.9210311
Adjusted R^2 (Test)	0.9244373	0.9399817	0.9005909	0.9034632
R^2 (Test)	0.9511991	0.9562367	0.9357983	0.9376533
Mean squared error	0.04780499	0.04287021	0.06289145	0.06107431

4 Conclusions

From the results of ridge and lasso regression it is evident that regression coefficient estimates shrink to zero as the shrinkage parameter (λ) increases. In ridge regression none of the coefficient estimates are exactly zero. The predictor variables that were dropped from the lasso regression model were gross national income per capita, social globalization index, political globalization index and environmental performance index performing a variable selection. These variables that were dropped in lasso regression model had the least importance in ridge regression model. Accordingly, we conclude that the impact on LEB by these predictor variables is quite smaller.

Twelve components were used to fit the model in principal components regression and four components in partial least squares regression. The no. of components required for partial least squares regression is lesser. This is because the response is not supervised in the identification of the components in principal components regression.

Range of adjusted R^2 values varies from 89.9% to 92.0% for the training data and 90.0% to 93.9% for the test data. The best model with a good fit for the data on hand is the ridge regression model because it has the highest adjusted R^2. The lasso regression model has lowest mean squared error and highest adjusted R^2 for the test data. Therefore lasso is the best predictive model among these models. The reason for lasso to have the lowest mean squared error is the variable selection property of lasso, because the predictors that do not have a significant impact on the response were dropped. Among principal components regression and partial least squares regression techniques partial least squares regression performs to be the best when considered with mean squared error and adjusted R^2 for the test data.

The predictor variable with the highest regression coefficient estimate in all the techniques is infant mortality rate. As such necessary steps to decrease infant mortality rate should be taken into consideration.

References

1. Bayati M, Akbarian R, Kavosi Z. Determinants of life expectancy in eastern mediterranean region: a health production function. International journal of health policy and management. 2013 Jun;1(1):57.
2. Mondal MN, Shitan M. Relative importance of demographic, socioeconomic and health factors on life expectancy in low-and lower-middle-income countries. Journal of epidemiology. 2014 Mar 5;24(2):117-24.
3. Mondal MN, Shitan M. Impact of socio-health factors on life expectancy in the low and lower middle income countries. Iranian journal of public health. 2013 42(12):1354.
4. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. New York: springer; 2013 Feb 11.
5. Melkumova LE, Shatskikh SY. Comparing Ridge and LASSO estimators for data analysis. Procedia engineering. 2017;201:746-55.
6. Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics. 1970;12(1):55-67.
7. Melhem M, Ananou B, Ouladsine M, Pinaton J. Regression methods for predicting the product’s quality in the semiconductor manufacturing process. IFAC-PapersOnLine. 2016; 49(12):83-8.
8. Tibshirani R. The lasso method for variable selection in the Cox model. Statistics in medicine. 1997,16(4):385-95.
9. Acharjee A, Finkers R, Visser RG, Maliepaard C. Comparison of regularized regression methods for omics data. Metabolomics. 2013,3(3):1.
10. Tranmer M, Elliot M. Multiple linear regression. The Cathie Marsh Centre for Census and Survey Research (CCSR). 2008, 5(5):1-5.
11. Alibuhtto MC, Peiris TS. Principal component regression for solving multicollinearity problem.