Afrika Statistika

Afrika Statistika
Vol. 15 (3), 2020, pages 2431 - 2449.
DOI: http://dx.doi.org/10.16929/as/2020.2431.167

ISSN 2316-090X

Regularization-Based Bootstrap Ranking Model: Identifying Healthcare Indicators Among All Level Income Economies

Emmanuel Thompson [1*] **and Ahmad Mahmoud Talafha** [2]

[1] Department of Mathematics, Southeast Missouri State University, Cape Girardeau, Missouri 63701, USA
[2] Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, Ohio 43403, USA

Received on September 19, 2020; Accepted on November 1, 2020.

Abstract. This study considers the problem of uncertainty of concurrent variables selection among a potential set of healthcare expenditure predictors. It evaluates two regularization (shrinkage) methods: Least Absolute Shrinkage and Selection Operator (LASSO) and Elastic Net (ENET). To improve the accuracy of identifying important and relevant predictors of healthcare cost, the present study proposes a new methodology in the form of a bootstrapped-regularized regression with percentile rankings. A simulation study under various scenarios was implemented to learn the performance of the proposed methodology. The proposed methodology was applied to healthcare expenditure data for all level income economies: lower-income, lower-middle-income, upper-middle-income, and high-income.

Key words: penalization; LASSO; elastic net; adaptive LASSO; bootstrapping; ranking; percentile rank; health expenditure per capita
AMS 2010 Mathematics Subject Classification Objects : 62J07; 62P05

[*]Corresponding author: Emmanuel Thompson (ethompson@semo.edu)
Ahmad Mahmoud Talafha : talafha@bgsu.edu

Résumé (French abstract) Cette étude examine le problème de l'incertitude de la sélection simultanée des variables concurrentielles parmi un ensemble potentiel de prédicteurs des dépenses de santé. Il évalue deux méthodes de régularisation (rétrécissement) : LASSO (Least Absolute Shrinkage and Selection Operato)et ENET (Elastic Net). Afin d'améliorer l'exactitude de l'identification des prédicteurs importants et pertinents du coût des soins de santé, la présente étude propose une nouvelle méthodologie sous la forme d'une régression régularisée avec des classements percentiles. Une étude de simulation pout divers scénarios a été mise en oeuvre pour connaître la performance de la méthodologie proposée. La méthodologie proposée a été appliquée aux données sur les dépenses de santé pour toutes les économies à revenu de niveau : à faible revenu, à revenu intermédiaire inférieur, à revenu intermédiaire supérieur et à revenu élevé.

The authors.

Emmanuel Thompson, Ph.D., is associate professor at Department of Mathematics, Southeast Missouri State University, Cape Girardeau, Missouri 63701, USA.

Ahmad Mahmoud Talafha, M.Sc, is a Ph.D. student in statistics in the Department of Mathematics and Statistics at Bowling Green State University, USA.

1. Background

Over the last several decades, healthcare systems globally have encountered intensive developments and improvement. However, there is a remarkable variation across the world income economies in healthcare spending. In 2010, around US $\$6.5$ trillion was spent on healthcare across the globe. For instance, per capita health expenditure was US $\$12$ on average per person a year in Eritrea, while it was over US $8000 in the USA World Health Organization (2014). In 2012, as a percentage of GDP healthcare expenditure amounted to 6.4% in the Middle East and Africa, 10.7% in Western Europe, and 17.4% in the USA. All major regions of the world are likely to experience increases in healthcare spending as a result of ineffective drugs, the prevalence of chronic diseases, and improved treatment modalities Global health care outlook Common goals (2015).

A report by WHO in 2010 indicates that healthcare spending is rising faster than the rest of the global economy, and it accounts for 10% of global GDP. For instance, health expenditure in low-and middle-income economies is growing at 6% on average annually and is 4% in high-income economies World Health Organization (2014).

Healthcare spending can be classified into four categories - government expenditure, out-of-pocket payments, and other sources (voluntary health insurance, employer-provided health programs, and activities by non-governmental organizations). For a given country, on average, governments shoulder 51% of overall

healthcare expenditure while out-of-pocket accounts for more than 35% of the expenditure. When government spending on health increases, people are less likely to fall into poverty seeking health services. In middle-income countries, government health expenditure per capita has doubled since the year 2000. On average, governments spend US $60 per person on health in lower-middle income countries and close to US $270 per person in upper-middle income countries. In low and middle-income countries, new data suggest that more than half of healthcare spending is devoted to primary health care. Yet less than 40% of all spending on primary health care comes from governments. For countries to continue to strengthen health systems, policymakers, health professionals, and citizens need to identify appropriate predictive models and relevant predictors of healthcare expenditure.

In the literature, healthcare cost prediction models are usually estimated by the method of least squares. The difficulty with this method is that, it has poor extrapolation property and sensitive to outliers therefore inappropriate in certain real-world situations Kronick *et al.*(2002), Gregori *et al.*(2011). Running a model with various predictor variables usually causes inaccurate inferences in the presence of multicollinearity, which is created by the appearance of significant correlations among the predictor variables Patriche *et al.*(2011).

Income has been shown to be a very important variable in explaining variations in healthcare expenditures across countries. However, there is no consensus as to which other variables may be associated with the main outstanding unexplained variation in health expenditure. Farag *et al.*(2009) studied the fungibility of Official Development Assistance (ODA) for health and domestic government health expenditure based on panel data from 1995 to 2006 for 144 countries. It turns out that a 1% increase in GDP caused a corresponding 0.66% and 1% increase in domestic government health expenditure in low and middle-income countries, respectively. Lv and Zhu (2014) employed a semiparametric panel data model to examine the link between income and health expenditure for 42 African countries at different levels of development. Thompson and Williams (2016) analyzed and identified the key predictors of health expenditure among low and lower-middle income economies by shrinking ordinary least squares regression estimates.

In previous studies, regularization or shrinkage methods appear to perform well in predictive models since they reduce the effects of sampling variability of the sample mean and resolve the problem of severe multicollinearity. The development of accurate predictive models using shrinkage methods has been more recent. Thompson (2015) combined the linear regression model with shrinkage methods to identify key predictors of healthcare expenditure among 42 African countries. While Loginov, Marlow, and Potruch (2012), investigate classification algorithms, determine which factors drove costs up and how these factors affect the total cost of healthcare in the USA in 2012.

In the statistical learning literature, a variety of shrinkage methods such as LASSO, ENET, and their variates have been proposed. The LASSO model Tibshirani (1996) has been developed to overcome the limitations when there are numerous predictors analyzed. By shrinking variables with very unstable estimates towards zero, the LASSO approach can successfully exclude some unconnected variables and produce sparse models. The presence of many predictors in healthcare cost analysis can result in problems for usual model fitting techniques. The LASSO algorithm has been shown to be an effective tool to remove unimportant predictors in studying the relationship between several predictors and health care cost Thompson and Williams (2016) . However, in practice, the LASSO method produces undue biases when choosing important variables and is not consistent in terms of variable selection Fan and Li (2001), Leng *et al.*(2006). This implies that the set of predictors, chosen by LASSO, is not consistently comprised of the true set of relevant predictors. It remains challenging to develop robust techniques of predictor selection and enhance predictability for healthcare cost analysis. Therefore, we require a modeling tactic that combines bootstrapped-regularized regression methods with percentile rankings to better identify relevant and informative healthcare predictors to improve decisions associated with the management of healthcare expenditure.

The present study aims at identifying important predictors of healthcare expenditure for low-income, lower-middle-income, upper-middle-income, and high-income economies by proposing a new methodology. Simulation studies were initially carried out to verify the performance of the proposed methodology. We applied the proposed methodologies to the 2014 World Bank data with several development and economic indicators. The rest of the paper is organized as follows: Section 2 explains the methods and the associated algorithms. Section 3 describes the set-up of the simulation studies and the 2014 World Bank data. Section 4 presents the results of the study. Discussion and conclusions are presented in section 5.

2. Methods

Consider data of the form (x_i, y_i) with $i \in \{1, \ldots, n\}$, where n is the number of samples. The vector $x_i = (x_{i1}, \ldots, x_{ip})^\top$ corresponds to the predictor variables for sample i and p-dimensional vector of predictors, furthermore y_i is the response variable.

A multiple linear regression model has the form:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j \, x_{ij} + \varepsilon_i \tag{1}$$

where β_0 denotes the intercept, β_j denotes the regression coefficient for the $j-$th predictor, and ε_i denotes the noise term which assumes values in $\mathbf{N}(0, \sigma^2)$, and the coefficients in (1) are estimated by minimizing the ordinary least squared (OLS) criterion:

$$\widehat{\beta}^{(OLS)} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j \, x_{ij} \right)^2 \tag{2}$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^\top$.

LASSO Regression Model

The LASSO estimates of $\widehat{\beta}$ in Tibshirani (1996) are defined as,

$$\widehat{\beta}^{(LASSO)} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j \, x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{3}$$

where λ is a nonnegative tuning parameter and the term $\sum_{j=1}^{p} |\beta_j|$ is called ℓ_1 penalty of β. The LASSO simultaneously shrinks the components toward 0 as λ increases; some components are shrunk to exact 0 for some appropriately chosen λ and obtain a sparse subset of variables with non-zero regression coefficients.
In (3), the optimality condition can not be achieved directly since $|\beta_j|$ does not have a derivative at $\beta_j = 0$. A solution is given by coordinate-wise minimization (coordinate descent).

Algorithm 1 : Coordinate Descent Algorithm for LASSO and ENET

1. Initialize all the $\beta_j = 0$ for $j \in \{1, 2, ..., p\}$
2. Cycle over j till convergence.
3. Compute partial residuals $r_{ij} = y_i - \sum_{k \neq j} x_{ik} \, \beta_k$.
4. Regress r_{ij} on x_{ij} to obtain OLS estimate $\widehat{\beta}_j$.
5. Update β_j using $S(z, \gamma)$ with $z = \widehat{\beta}_j$ and $\gamma = \alpha\lambda$: $\beta_j \leftarrow \frac{S(\widehat{\beta}_j, \alpha\lambda)}{1+\lambda(1-\alpha)}$, where

$$S(z, \gamma) = \begin{cases} z - \gamma & : \text{if } z > 0 \text{ and } \gamma < |z|, \\ z + \gamma & : \text{if } z < 0 \text{ and } \gamma < |z|, \\ 0 & : \text{if } \gamma \geq |z|. \end{cases}$$

In **Algorithm 1**, the LASSO solution results when $\alpha = 1$.

However, it was shown by Zou (2005) that the LASSO could be somewhat unfair because it requires its components to be equally penalized in the ℓ_1 penalty. Therefore, Zou (2005) proposed adaptive LASSO estimator by imposing a cleverly chosen weight vector, $\widehat{w}_j$.

Adaptive LASSO Regression Model

The adaptive LASSO (ALASSO) is defined as,

$$\widehat{\beta}^{(ALASSO)} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j} \beta_j \, x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} w_j \, |\beta_j| \right\} \tag{4}$$

where $\sum_{j=1}^{p} \widehat{w}_j$ are the adaptive data-driven weight, that is, $\widehat{w}_j = \left(\left| \widehat{\beta}^{(Ridge)} \right| \right)^{-\psi}$, for some $\psi > 0$ and $\widehat{\beta}^{(Ridge)}$ is the ridge regression Tibshirani (1996) root-n consistent estimate of $\widehat{\beta}$.

Collinearity can harshly degrade the achievement of the LASSO method. It was shown in Zou and Hastie. (2005), the LASSO solution paths are unstable when predictors are highly correlated; therefore, the elastic net method links the LASSO method and ridge regression. It supports possessing a parsimonious model with adopting strength from correlated regressors, by imposing ℓ_1 and ℓ_2 penalties. Zou and Hastie. (2005) proposed the elastic-net as an improved version of the LASSO for analyzing high-dimensional data.

The computational details of the ALASSO solution obtained via the Least Angle Regression (LARS) algorithm Efron *et al.*(2004)

Algorithm 2 : LARS Algorithm for ALASSO

1. Define $x_{ij}^{**} = x_{ij} / \widehat{w}_j$ for $j \in \{1, 2, ..., p\}$.
2. Solve $\widehat{\beta}^{**} = \arg\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j} \beta_j \, x_{ij}^{**} \right)^2 + \lambda \sum_{j=1}^{p} w_j \, |\beta_j| \right\}$.
3. Output $\widehat{\beta}_j^{*} = \widehat{\beta}_j^{**} / \widehat{w}_j$ for $j \in \{1, 2, ..., p\}$.

ENET Regression Model

The elastic-net estimator is defined as follows:

$$\widehat{\beta}^{(ENET)} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \sum_{j} \beta_j \, x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \left[\frac{1}{2} \, (1 - \alpha) \, \beta_j^2 + \alpha \, |\beta_j| \right] \right\} \tag{5}$$

where $\sum_{j=1}^{p} \beta_j^2$ is called the ℓ_2 penalty. Elastic net uses coordinate decent algorithm to solve (5) and chooses $\alpha \in (0, 1)$. .

Bootstrap-Regularized Regression with Percentile Rankings

Here, we present the algorithm underlying our proposed methodology.

Algorithm 3 : Bootstrap-Regularized Regression Method

1. Draw a bootstrap sample of the form $(x^{\star b}_i, y^{\star b}_i)$ with $i \in \{1, \ldots, n\}$ and $b \in \{1, \ldots, B\}$ from (x_i, y_i) with replacement.
2. Apply K-fold cross-validation procedure on the selected pair $(x^{\star b}_i, y^{\star b}_i)$; the data set is randomly selected into training and test sets with a $70/30$ split between the two into K folds.
3. Fit the penalized regression method on the training set and distributively using each of the K folds as the test set and select the model with the smallest mean squared prediction error.
4. Apply steps 1–3 to each of b bootstrap samples
5. Compute coefficients set at

$$J = \left\{ j \in \{1, 2, ..., p\} : \widehat{\beta}_j{}^b \right\}$$

6. Generate the coefficients estimates matrix

$$\mathbf{E}_{B\times p} = \begin{pmatrix} J_{1\times p} \\ \vdots \\ J_{B\times p} \end{pmatrix}$$

7. Apply the absolute value to every element of $\mathbf{E}$. Sort all variables in $J_{p\times 1}$ in descending order and rank them with tied values given the average of the ranks that the ties would have attained.
8. Create a rank matrix $R_{p\times B}$ that consists of all the ranks of J's according to step 7.
9. Take the average ranks of each row of R, and compute the percentile rank of it.

3. Description of Simulation Studies and 2014 WHO Data

Simulation Studies

Extensive simulation studies were conducted to demonstrate the proposed methodology's validity when applied to LASSO, ALASSO, and ENET. The performance of the approach was measured by bootstrap mean square error (BMSE), and the algorithm that achieves the lowest BMSE is used as a basis of establishing variable importance.

Consider the linear model

$$y_i = \beta_0 1 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ is the true coefficient vector. Let $x_i = (x_{i1}, \ldots, x_{ip})^\top$ and the error term, ε, generated from $\varepsilon_i \sim \mathbf{N}(0, 1)$, $x_i \sim \mathbf{N}(0, \sigma^2 I_p)$, $\sigma_k^2 = Var x_k$ and $\sigma_{k\,l} = Cov x_k x_l$.

The pairwise correlation between the k$^{\text{th}}$ and l^{th} predictors is given by:

$$cor(k,l) = \frac{\sigma_{k\,l}}{\sigma_k\,\sigma_l} = \rho^{|k-l|} \quad \text{for } k,l = 1,\dots p$$

The main effects of predictors were simulated according to the following scenarios:

- Scenario 1 : The linear model was defined with $n = 50$, $p = 10$ and pairwise correlation $\rho = 0.90^{|k-l|}$, for $k,l = 1,\dots 10$. The true $\beta_{p\times 1}$ were as follows:

$$\beta = (10, 10, 5, 5, \underbrace{0, ..., 0}_{6})^{\top}.$$

- Scenario 2 : Same as Scenario 1, except that the pairwise correlation is $\rho = 0.30^{|k-l|}$ for $k,l = 1,\dots 10$.
- Scenario 3 : The linear model was defined with $n = 50$, $p = 40$ and pairwise correlation $\rho = 0.90^{|k-l|}$ for $k,l = 1,\dots 40$. The true $\beta_{p\times 1}$ were as follows:

$$\beta = (10, 10, 5, 5, \underbrace{1, ..., 1}_{5}, \underbrace{0, ..., 0}_{31})^{\top}.$$

- Scenario 4 : Same as Scenario 3, except that the pairwise correlation is $\rho = 0.30^{|k-l|}$, where $k,l = 1,\dots,40$.
- Scenario 5 : The linear model was defined with $n = 100$, $p = 40$ and pairwise correlation $\rho = 0.90^{|k-l|}$, for $k,l = 1,\dots 40$. The true $\beta_{p\times 1}$ were as follows:

$$\beta = (10, 10, 5, 5, \underbrace{1, ..., 1}_{5}, \underbrace{0, ..., 0}_{31})^{\top}.$$

- Scenario 6 : Same as Scenario 5, except that the pairwise correlation is $\rho = 0.30^{|k-l|}$, for $k,l = 1,\dots 40$.

2014 WHO Data

The data used in this analysis were obtained from The World Bank (2014) open data website. The data consist of 11 variables where Health Expenditure per Capita (HEC) denotes the response variable, and the rest represent the predictor variables (Table 1) from low income (LI), lower-middle income (LMI), upper-middle income (UMI), and high income (HI) economies. Some countries were excluded from the analyses due to missing data, and the possibility to obtain them proved futile.

The logarithmic transformation scale was used to transform all variables (dependent and independent) before analyzing the data. Tables 8, 9, 10, 11 in Appendix report descriptive statistics of the natural logarithm of the variables of the income economies.

The data were partitioned into training and testing sets as well as LI, LMI, UMI, and HI economies aligned with the World Bank revised classification The World

Table 1. List of Variables with Description

Variable	Description
HEC	Health Expenditure per Capita, PPP (constant 2011 international $)
GDP	Gross Domestic Product per Capita, PPP (constant 2011 international $)
CPI	Consumer Price Index (2010 = 100)
PP1	Population Age 0-14 (% of total)
PP2	Population Ages 15-64 (% of total)
PP3	Population Ages 65 and above (% of total)
PPD	Population Density (people per sq. km of land area)
IMR	Mortality Rate Infant (per 1,000 live births)
EXR	Official Exchange Rate (LCU per US$, period average)
TBC	Incidence of Tuberculosis (per $100,000$ people)
LIF	Life Expectancy at Birth, total (years)

Bank (2014). The Pearson correlation coefficient via heatmap was utilized to illustrate the strength of multicollinearity between HEC and each of the predictors and also among the predictors themselves. Bootstrap-Regularized Regression approaches were used to identify essential predictors of HEC. The performance of the approaches was measured by BMSE. Table 2 shows the frequency for each income economy after excluding missing data.

Table 2. The frequency of data among all LI, LMI, UMI, and HI economies

Income class	Observation	Training	Testing
Low	25	17	8
Lower-middle	45	31	14
Upper-middle	44	30	14
High	31	21	10

The Pearson correlation coefficient, ranges between -1 and $+1$ and assesses the direction and strength of the linear association among the variables. Figure 1 shows that some of the predictors were correlated among the LI economies. The correlation between GDP and HEC is about 0.65, which indicates that there is a moderate positive relationship between the variables. Also, there is a very strong negative linear correlation between PP2 and PP1 as well as PP3 and PP1.

A strong positive linear relationship exists between GDP and HEC and between PP2 and PP3 as well for the LM economies. However, a moderate but negative linear relationship exists between IMR and HEC. All these results are shown in Figure 2.

Figure 3, indicates for UMI, a strong positive linear relationship exists between GDP and HEC. Similarly, a moderately strong positive relationship exists between

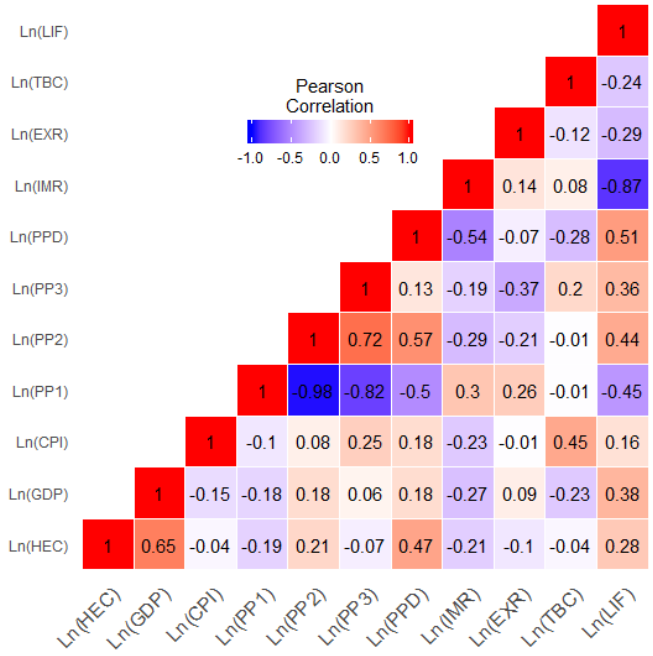

Fig. 1. Correlation heatmap of the Low economies coefficients

LIF and PP3. For the same income group however, a strong but negative linear association exists between PP2 and PP1 as well as between PP3 and PP1. Figure 4, indicates a positive linear relationship between GDP and HEC among HI economies. For the same income group, a negative but moderate linear relationship between IMR and HEC was observed as well as a negative strong association between LIF and IMR.

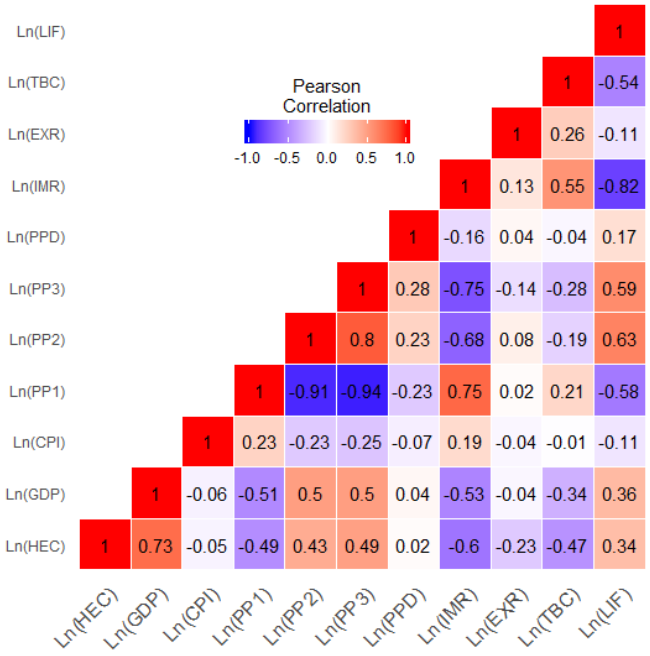

Fig. 2. Correlation heatmap of the Lower-middle economies coefficients

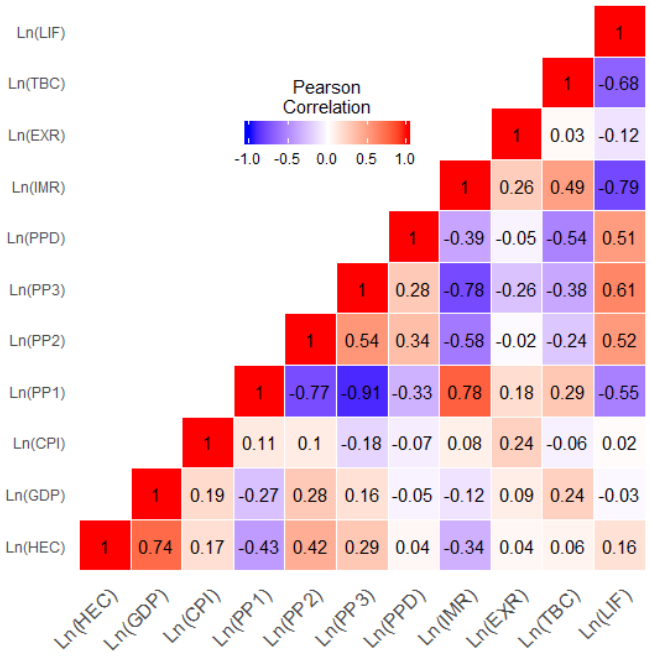

Fig. 3. Correlation heatmap of the upper-middle economies coefficients

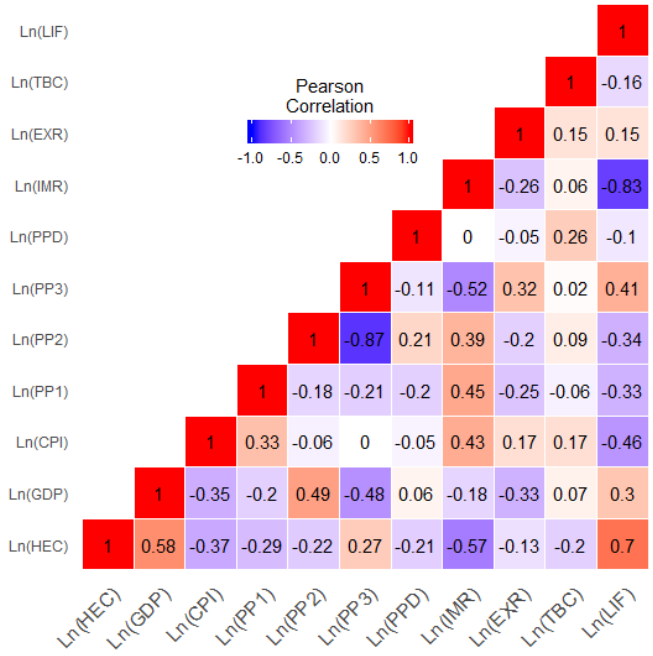

Fig. 4. Correlation heatmap of the high economies coefficients

4. Results

Performance Evaluation based on Simulation Studies

The performance in terms of predictive powers of the Bootstrap-LASSO, Bootstrap-ALASSO, and Bootstrap-ENET models using 10,000 iterations (Table 3) reveal that, the Bootstrap-LASSO model performs better than both the Bootstrap-ALASSO and Bootstrap-ENET models for 4 out 6 of the simulation scenarios. While Bootstrap-ENET models performs better in 2 out 6 of the simulation scenarios. In the case of highly correlated predictors (scenario 1, 3 and 5), the Bootstrap-ENET model stands out to be the best in identifying the important predictors in 2 out 3 simulation scenarios, highlighting the usefulness of such an approach in eliminating insignificant predictors leaving only the important ones. Models with least BMSE among each scenario were chosen for further analysis.

Table 3. BMSE by Scenario and Model ($B = 10,000$)

	BMSE		
Model / Scenario	LASSO	ALASSO	ENET
Scenario 1	1.366*	71.706	1.374
Scenario 2	1.176*	40.277	1.193
Scenario 3	1.821	76.378	1.819*
Scenario 4	2.910*	67.768	3.219
Scenario 5	1.732	112.916	1.691*
Scenario 6	0.900*	7.107	0.914

* The smallest among all methods

Percentile ranks in Table 4 established relative variable importance among the rest of the variables for each of the scenario 1, 2 and 3. The higher the percentile rank for a variable, the more weight it has received compared to variables in the same scenario. For example, a value 90-100 means that the variable in question is among 10% most important variable; the other 90% of the variables have achieved less impact. Scenario 1 shows that the non-zero variables were identified within the 60-70 percentile, whereas the zero ones were below 60 percentile. Even with a higher collinearity degree in Scenario 2, the proposed method was able to detect the important non-zero variables among 60-70 percentile. Scenario 3 showed that almost all the non-zero variables were among the 20% most important variable.

In Table 5 , Scenario 4 illustrated that increasing the degree of the pairwise correlation still enable the algorithm to filter the non-zero variables among 20% most important variables. Scenario 5 and 6 were able to select the non-zero variables among almost 20% most important variables.

Table 4. Percentile rank for Variables identified using Bootstrap rank-based procedure among Scenario 1-3

	Variable Selected		
Percentile Rank Range	Scenario 1	Scenario 2	Scenario 3
90-100	x_2	x_2	$x_1 - x_4$
80-90*	x_1	x_1	x_5, x_6, x_8, x_9
70-80*	x_3	x_4	$x_7, x_{10}, x_{21}, x_{28}$
60-70*	x_4	x_3	$x_{11}, x_{13}, x_{25}, x_{27}$
50-60*	x_5	x_8	$x_{12}, x_{20}, x_{32}, x_{38}$
40-50*	x_9	x_5	$x_{15}, x_{24}, x_{33}, x_{37}$

* Not included in the interval

Table 5. Percentile rank for variables identified using Bootstrap rank-based procedure among Scenario 4-6

	Variable Selected		
Percentile Rank Range	Scenario 4	Scenario 5	Scenario 6
90-100	$x_1 - x_4$	$x_1 - x_4$	$x_1 - x_4$
80-90*	$x_5 - x_8$	$x_5 - x_7, x_9$	$x_5, x_7 - x_9$
70-80*	$x_9, x_{18}, x_{29}, x_{30}$	$x_8, x_{10}, x_{13}, x_{14}$	$x_6, x_{13}, x_{18}, x_{19}$
60-70*	$x_{16}, x_{19}, x_{20}, x_{27}$	$x_{11}, x_{12}, x_{17}, x_{18}$	$x_{23} - x_{25}, x_{35}$
50-60*	$x_{31}, x_{32}, x_{34}, x_{37}$	$x_{15}, x_{16}, x_{19}, x_{21}$	$x_{21}, x_{26}, x_{30}, x_{34}$
40-50*	$x_{12}, x_{17}, x_{26}, x_{35}$	$x_{20}, x_{25}, x_{28}, x_{37}$	$x_{15}, x_{22}, x_{27}, x_{38}$

* Not included in the interval

Performance evaluation based HEC data

The proposed methodology was applied to the healthcare data among all level of income. This has also been used for illustration purpose in previous studies Thompson (2015), Thompson and Williams (2016) . The result from applying $10,000$ iterations of Bootstrap-LASSO procedure, Bootstrap-ALASSO procedure, and Bootstrap-ENET procedure indicate that both LASSO and ENET procedures provide a competitive BMSE value (Table 6). However, the method resulted in the smallest BMSE among each income class was chosen for further analysis.

Based on percentile rank in Table 7, GDP was among 10% important variables in LI, LMI, and UMI income class while LIF was the 10% important variable in HI income class. It also shows that PP2 received attention from all income classes, which stands in $20 - 40\%$ important variable. It is worth mentioning that EXR was the least important variable among LI, UMI, and HI.

5. Conclusions and future work

The motivation behind this research was to formulate a more efficient means of accurately selecting important predictors of HEC. Two newly proposed variable se-

Table 6. Bootstrap Mean squared error (BMSE) for each penalized regression method among all income group ($B = 10,000$)

	BMSE estimation		
Method / Income class	LASSO	ALASSO	ENET
Low	0.1838	0.1803	0.1736*
Lower-middle	0.1827*	0.2283	0.1830
Upper-middle	0.1173*	0.1575	0.1184
High	0.1256*	0.1702	0.1278

* The smallest among all methods

Table 7. Percentile rank for variables identified using Bootstrap rank-based procedure among all income group

	Variable Selected			
Income class / Percentile Rank Range	LI	LMI	UMI	HI
90-100	GDP	GDP	GDP	LIF
80-90*	LIF	IMR	PP2	GDP
70-80*	PP3	LIF	PP1	PP2
60-70*	PP2	PP2	IMR	CPI
50-60*	PPD	TBC	LIF	PP3
40-50*	CPI	CPI	CPI	PP1
30-40*	IMR	PP1	PP3	TBC
20-30*	PP1	EXR	PPD	IMR
10-20*	TBC	PP3	TBC	PPD
0-10*	EXR	PPD	EXR	EXR

* Not included in the interval

lection algorithms, Bootstrapped-LASSO regression with percentile rankings and Bootstrapped-ENET regression with percentile rankings were investigated in this study. The extensive simulation study revealed that, the proposed methodology eliminated unimportant predictors and established relative variable importance among the rest of the variables in a comparison group. The application of these algorithms in the empirical study concluded that governments should investigate further and reduce the effect of large variability in HEC spending, such as LIF, PPD, and IMR before any effort is made to reduce the total level of government spending.

In this study, the emphasis has been on a simulation study with a relatively medium dimensionality in terms of the predictor space. However, results on how the proposed methodology behaves with very large number of predictors or when the number of predictors exceeds the number of observations i.e $p >> n$ were not considered. Nevertheless, the bootstrapped-regularized regression with percentile rankings can be applied to tackle this type of situation. For future analysis, this

study could be extended to investigate the behavior of additional regularization problem such as adaptive ENET. In summary, the proposed methodology has the potential to aid governments to better manage the large variability in HEC spending on their national expenditure.

Acknowledgments

The authors would like to thank the editorial team for their comments and the anonymous reviewers for their insightful suggestions and careful reading of the manuscript.

Appendix

Table 8. Descriptive statistics of the natural logarithm of the variables of the Low Income economies

	Mean	SD	Median	Min.	Max.	Skewness	Kurtosis	Shapiro-Wilk Statistic	Shapiro-Wilk p-value
Ln(HEC)	7.854	0.561	7.785	6.739	9.149	0.237	-0.519	0.980	0.821
Ln(GDP)	10.551	0.475	10.552	9.703	11.702	0.307	-0.567	0.982	0.874
Ln(CPI)	4.707	0.063	4.693	4.598	4.928	1.530	3.331	0.862	0.001
Ln(PP1)	2.916	0.211	2.964	2.570	3.325	0.038	-1.161	0.953	0.191
Ln(PP2)	4.240	0.078	4.225	4.114	4.446	0.975	0.803	0.919	0.022
Ln(PP3)	2.175	0.888	2.589	-0.012	3.233	-1.160	0.083	0.818	0.000
Ln(PPD)	4.475	1.805	4.821	1.116	8.951	0.040	-0.231	0.969	0.479
Ln(IMR)	1.646	0.601	1.548	0.531	2.874	0.021	-0.949	0.981	0.848
Ln(EXR)	1.561	2.137	1.149	-1.257	6.959	1.013	0.116	0.888	0.004
Ln(TBC)	2.371	1.053	2.398	-0.274	4.443	-0.171	-0.196	0.986	0.946
Ln(LIF)	4.362	0.044	4.361	4.255	4.426	-0.294	-0.867	0.943	0.097

Table 9. Descriptive statistics of the natural logarithm of the variables of the Lower-middle Income economies

	Mean	SD	Median	Min.	Max.	Skewness	Kurtosis	Shapiro-Wilk Statistic	Shapiro-Wilk p-value
Ln(HEC)	6.772	0.439	6.833	5.598	7.599	-0.535	0.014	0.966	0.220
Ln(GDP)	9.520	0.390	9.560	8.523	10.359	-0.434	0.080	0.978	0.560
Ln(CPI)	4.801	0.214	4.758	4.591	5.853	3.501	13.280	0.568	0.000
Ln(PP1)	3.203	0.286	3.235	2.626	3.710	-0.267	-0.884	0.964	0.185
Ln(PP2)	4.191	0.064	4.197	4.026	4.291	-0.860	0.192	0.926	0.008
Ln(PP3)	1.992	0.483	1.910	1.074	2.982	0.195	-0.684	0.979	0.601
Ln(PPD)	3.933	1.416	4.178	1.058	7.198	-0.161	-0.301	0.972	0.355
Ln(IMR)	2.613	0.603	2.613	1.435	4.250	0.307	0.014	0.972	0.352
Ln(EXR)	3.124	2.562	2.384	-0.283	10.164	0.798	-0.270	0.917	0.004
Ln(TBC)	3.689	1.198	3.580	1.335	6.726	0.457	0.111	0.965	0.199
Ln(LIF)	4.290	0.070	4.314	4.048	4.375	-1.648	2.410	0.813	0.000

Table 10. Descriptive statistics of the natural logarithm of the variables of the Upper-middle Income economies

	Mean	SD	Median	Min.	Max.	Skewness	Kurtosis	Shapiro-Wilk Statistic	Shapiro-Wilk p-value
Ln(HEC)	5.588	0.610	5.641	4.478	6.682	-0.116	-1.072	0.967	0.218
Ln(GDP)	8.506	0.470	8.473	7.543	9.337	-0.001	-1.016	0.971	0.310
Ln(CPI)	4.864	0.167	4.831	4.650	5.699	2.687	11.356	0.761	0.000
Ln(PP1)	3.485	0.274	3.518	2.694	3.855	-1.070	0.778	0.908	0.002
Ln(PP2)	4.111	0.096	4.127	3.922	4.309	-0.146	-1.015	0.973	0.363
Ln(PP3)	1.501	0.438	1.454	0.826	2.760	1.081	0.877	0.912	0.002
Ln(PPD)	4.152	1.214	4.342	0.632	7.110	-0.412	0.703	0.970	0.282
Ln(IMR)	3.388	0.588	3.364	2.128	4.311	-0.299	-0.901	0.963	0.165
Ln(EXR)	4.161	2.642	4.111	-0.342	9.959	0.458	-0.510	0.962	0.145
Ln(TBC)	4.973	1.033	5.106	1.705	6.748	-0.789	0.715	0.957	0.093
Ln(LIF)	4.205	0.097	4.225	3.962	4.327	-0.996	0.302	0.900	0.001

Table 11. Descriptive statistics of the natural logarithm of the variables of the High Income economies

	Mean	SD	Median	Min.	Max.	Skewness	Kurtosis	Shapiro-Wilk Statistic	Shapiro-Wilk p-value
Ln(HEC)	4.513	0.454	4.538	3.217	5.411	-0.698	0.905	0.959	0.395
Ln(GDP)	7.281	0.351	7.367	6.400	7.784	-0.778	-0.155	0.932	0.097
Ln(CPI)	4.856	0.186	4.834	4.649	5.326	0.892	0.018	0.897	0.016
Ln(PP1)	3.769	0.092	3.767	3.509	3.916	-1.228	1.745	0.868	0.004
Ln(PP2)	3.978	0.060	3.980	3.855	4.115	0.414	0.277	0.954	0.301
Ln(PP3)	1.073	0.202	1.046	0.783	1.670	1.248	1.455	0.886	0.009
Ln(PPD)	4.354	1.151	4.325	1.981	6.131	-0.273	-0.788	0.960	0.406
Ln(IMR)	3.963	0.288	3.985	3.434	4.538	0.068	-0.631	0.973	0.726
Ln(EXR)	5.961	1.568	6.203	2.975	8.856	-0.194	-0.834	0.930	0.086
Ln(TBC)	5.048	0.721	5.176	3.555	6.312	-0.378	-0.905	0.948	0.229
Ln(LIF)	4.095	0.080	4.100	3.924	4.241	-0.504	-0.318	0.956	0.349

References

C. Leng, Y. Lin, and G. Wahba. A NOTE ON THE LASSO AND RELATED PROCEDURES IN MODEL SELECTION. *Statistica Sinica* . 2006; 16:1273-1284

J. Fan and R. Li. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. J. Am. *Stat. Assoc*. 2001; 96:1348-1360

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. LEAST ANGLE REGRESSION. *The Annals of Statistics* . 2004; 32:407–499

World Health Organization, WHO. Global Health Expenditure Atlas. 2014. http://www.who.int/health-accounts/. Accessed 23 Dec 2017

World Development Indicators. 2014. https://data.worldbank.org/data-catalog/world-development-indicators. Accessed 23 Dec 2017

H. Zou, A. Hao, and H. Zhang. ON THE ADAPTIVE ELASTIC-NET WITH A DIVERGING NUMBER OF PARAMETERS. *The Annals of Statistics*. 2009; 37:1733–1751

H. Zou. The Adaptive Lasso and Its Oracle Properties. J. Am. Stat. Assoc. 2005; 101:1418-1429

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. J. R. Statist. Soc. B. 2005; 67:301-320

R. Tibshirani. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological) . 1996; 58:267-288

M. V. Loginov, E. Marlow, and V. Potruch. Predictive Modeling in Healthcare Costs using Regression Techniques. The 47th Actuarial Research Conference . 2012; undefined:37

E. Thompson. HEALTH CARE EXPENDITURE IN AFRICA – AN APPLICATION OF SHRINKAGE METHODS. International Journal of Mathematics and Statistics Studies. 2015; 3:15-20

E. Thompson and F. Williams. Predictive models of health expenditure via regularization: Do low and upper middle income economies share common predictors?. International Journal of Health Research and Innovation . 2012; 4:13-21

Lv, Zhike, and Huiming Zhu. Health care expenditure and GDP in African countries: evidence from semiparametric estimation with panel data. *The Scientific World Journal*, vol. 14, 2014

M. Farag, A. Nandakumar, S. S. Wallack, G. Gaumer, and D. Hodgkin. Does Funding From Donors Displace Government Spending For Health In Developing Countries?. Health Affairs . 2009; 4:1045-1055

C. V. Patriche, R. G. Pîrnău, and B. Roşca. Comparing Linear Regression and Regression Trees for Spatial Modelling of Soil Reaction in Dobrovăţ Basin (Eastern Romania). Bulletin UASVM Agriculture . 2011; 68:264-271

D. Gregori, M. Petrinco, S. Bo, A. Desideri, F. Merletti, and E. Pagano. Regression models for analyzing costs and their determinants in health care: an introductory review. International Journal for Quality in Health Care . 2011; 23:331-41

Kronick, Richard and Gilmer, Todd P and Dreyfus, Tony and Ganiats, Theodore G and Director, Project. CDPS-Medicare: The Chronic Illness and Disability Payment System Modified to Predict Expenditures for Medicare Beneficiaries. 2002. http://cdps.ucsd.edu/. Accessed 23 Dec 2017

Deloitte, Global health care outlook Common goals, competing priorities. 2015. https://www.deloitte.com/content/dam/Deloitte/global/Documents/Life-Sciences-Health-Care/gx-lshc-2015-health-care-outlook-global. Accessed 02 Jan 2018