International Journal of Health Research and Innovation, vol.4. no.2, 2016, 13-21
ISSN: 2051-5057 (print version), 2051-5065 (online)
Scienpress Ltd, 2016

Predictive models of health expenditure via regularization: Do low and upper middle income economies share common predictors?

Emmanuel Thompson[1] and Faustine Williams[2]

Abstract

Countries around the world are presently confronted with gargantuan health care challenges and huge variability in health spending. In the literature, income has been recognized as a crucial predictor of health expenditure. However, there is no agreement on which other variables may be connected to the remaining largely unexplained variation in health expenditure. Therefore, the aim of the present study was to investigate the link between health expenditure and some important predictors among low-income and lower middle-income economies. Regularized regression methods including the Lasso and the Elastic net, and the 2013 World Bank data were used to identify key predictors of health expenditure. The present study showed that the Elastic net algorithm produced a model with a better predictive power than the Lasso in the case of low-income economies. However, the Lasso produced a slightly superior predictive power compared to the Elastic net in the case of lower middle-income economies. Also, income remains a common predictor of health expenditure in both economies. Findings of the study would be valuable to governments seeking to lessen the impact of vast variability in health care spending on their economies by focusing on key predictors of health expenditure per capita, such as life expectancy and population density.

Keywords: Regularization, Lasso, Elastic net, Health expenditure per capita

[1]Department of Mathematics, Southeast Missouri State University, Cape Girardeau, Missouri, USA.
[2]Washington University in St. Louis School of Medicine, St. Louis, Missouri, USA.

1 Introduction

Today, countries around the world are not only confronted with gargantuan health care challenges, but also huge variability in health spending. For instance in poor countries, per capita health expenditure are only US$30, while in high income countries; it is over US$ 3,000 on average.[1] In 2013, the total global health spending was projected to go up by 2.8% before moving to an average of 5.2% per year from 2014 to 2018.[2, 3] All regions of the world are likely to experience rising health spending because of factors like population growth, ageing, prevalence of chronic diseases, improvement in treatment of diseases, better access to information among others.[2, 3]

In the literature, income has been recognized as a crucial predictor of health expenditure. However, there is no agreement on which other variables may be connected to the remaining largely unexplained variation in health expenditure.[4, 5] More so, available studies relating health care expenditure to income and other indispensable non-income predictors are mostly among organization for economic co-operation development (OECD) countries.[1, 2, 4] It is important to emphasize that wide variation exists in health spending for countries at different stages of economic development. While some countries spend over 12% of GDP on health, others spend below 3%.[1] Partly for this reason and for the fact that income has been identified as the key predictor of health care expenditure, it is possible that what determines health care expenditure could be the same for countries falling in the same economic class.

Xu et al. studied the factors related to total health expenditure, government health expenditure, and private out of pocket health expenditure taking into consideration income levels of 143 developing and developed countries.[1] Lv and Zhu used a semi parametric panel data model to analyze the relationship between income and health expenditure for 42 African countries at different levels of development.[6] This study seeks to also contribute to the literature by investigating the link between health expenditure and ten (10) predictors using regularized regression (table 1). Our primary objective was to identify key predictors of health expenditure and then propose a predictive model based on the World Bank revised economic classification of countries, particularly low income (LI) and lower middle income (LMI) economies using 2013 data.[7]

There are instances when the traditional statistical estimation methods such as the least squares (LS) tend to produce erroneous estimates of parameters causing inaccurate inferences in the presence of multicollinearity.[8] Regularized regression methods were therefore developed to surmount the weaknesses of LS method in terms of prediction accuracy. Among these methods are Ridge regression,[9,10] the Lasso,[11] and lately LARS,[12] Pathseeker[13] and the Elastic net.[14] The present study centers on the comparison between the Lasso and the Elastic net. Lasso stands for Least Absolute Shrinkage and Selection Operator and has the property of shrinking some of the regression coefficients to exactly zero. The Elastic net on the other hand

is an improved version of the Lasso. The method selects predictors like Lasso and shrinks the coefficients of correlated predictors similar to Ridge regression. While most previous studies concentrated much on panel data and time series models, the present study accentuates supervised learning and predictive modelling.

2 Method

This study utilized data from the World Bank open data website.[15] The variables used were as follows:

Table 1: Definition of variables

Variable	Definition
HEC	Health Expenditure per Capita, PPP (constant 2005 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)

Health expenditure per capita (HEC) was used as the response variable to capture health expenditure, while the rest were predictors of HEC. Some countries were excluded from the study due to missing data and the possibility to obtain them proved unsuccessful. All variables were transformed into natural logarithm (Ln); and the whole data was further partitioned into LI and LMI economies in line with the World Bank revised classification of countries by gross national income (GNI).[7] Both descriptive and inferential statistical analyses were performed on each of the income class data. The Pearson correlation coefficient was used to describe the strength of the linear association between the response variable and each of the predictors and also among the predictors. Lasso and Elastic net algorithms were used to identify important predictors of HEC. Prior to estimating the regression coefficients, each of the income class data was divided into training and testing sets. The training sets were used to estimate coefficients using both the Lasso and the Elastic net algorithms. The testing sets were then used to check the performance of the models emanating from the training sets. The mean square

error (MSE) was used as a metric to assess the error of prediction of the models. Table 2 shows the distribution of countries in terms of income class after excluding missing data.

Table 2: Frequency distribution of countries by income class

Category	Observations	Training	Testing
Low Income Economies	28	16	12
Low-Middle Income Economies	41	24	17
Total	**69**	**40**	**29**

2.1 Regularized Regression

Consider the general linear model of the Gaussian family where i represent countries.

$$y_i = \beta^T x_i + \epsilon_i,$$

where (x_i, y_i); $i = 1, 2, \ldots, N$ are a sample of N independent and identically distributed (i.i.d) random vectors, where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ik}) \in \mathcal{R}^k$ is the random vector of observations about k predictors for the i^{th} sample unit and $y_i \in \mathcal{R}$ is the corresponding response vector. ϵ_i is a stochastic error term capturing all factors that affect HEC but are not taken into consideration explicitly.[16] The vector of $(k+1)$ estimates $(\hat{\beta}_0, \hat{\beta})$ of regression coefficients were obtained by applying the coordinate descent[16] to solve the optimization problem whose objective function is given by

$$\min_{(\beta_0,\beta)\in\mathcal{R}^{k+1}} \sum_{i=1}^{N} \left(y_i - \beta_0 - x_i^T\beta\right)^2 + \lambda[(1-\alpha)\,\|\beta\|_2^2/2 + \alpha\|\beta\|_1^2]$$

where $\lambda \geq 0$ and $0 \leq \alpha \leq 1$. The Ridge regression[18] coefficients are obtained by setting $\alpha = 0$ which is not a subject of consideration in this study. When $\alpha = 1$, the optimized problem produces the Lasso regression coefficients and $0 < \alpha < 1$ results in the Elastic net regression coefficients. Because our models were based on the natural logarithm transformation, we unambiguously interpret the β's as coefficients of elasticity (CoE). CoE is the ratio of the percentage change in HEC to the percentage change in a specific predictor such as LIF.

3 Results

Tables 3 and 4 show the linear association among the variables for the LI and LMI

economies respectively. Each of the income class data showed evidence of high correlation among some of the predictors.

Table 3: Pearson correlation coefficients for the LI economies

	Ln(HEC)	**Ln(GDP)**	**Ln(CPI)**	**Ln(PP1)**	**Ln(PP2)**	**Ln(PP3)**	**Ln(PPD)**	**Ln(IMR)**	**Ln(EXR)**	**Ln(TBC)**	**Ln(LIF)**
Ln(HEC)	1										
Ln(GDP)	0.680	1									
Ln(CPI)	0.110	-0.074	1								
Ln(PP1)	-0.294	-0.451	0.007	1							
Ln(PP2)	0.292	0.426	0.007	**-0.982**	1						
Ln(PP3)	0.114	0.259	0.025	**-0.852**	0.77	1					
Ln(PPD)	0.430	0.275	0.287	-0.448	0.476	0.200	1				
Ln(IMR)	-0.409	-0.473	-0.216	0.470	-0.430	-0.413	**-0.576**	1			
Ln(EXR)	-0.121	-0.191	0.203	0.199	-0.203	-0.142	-0.111	0.165	1		
Ln(TBC)	-0.197	-0.100	0.266	-0.237	0.207	0.355	-0.217	0.102	-0.009	1	
Ln(LIF)	0.439	**0.612**	-0.041	**-0.595**	**0.557**	0.465	0.475	**-0.868**	-0.279	-0.174	1

Table 4: Pearson correlation coefficients for the LMI economies

	Ln(HEC)	**Ln(GDP)**	**Ln(CPI)**	**Ln(PP1)**	**Ln(PP2)**	**Ln(PP3)**	**Ln(PPD)**	**Ln(IMR)**	**Ln(EXR)**	**Ln(TBC)**	**Ln(LIF)**
Ln(HEC)	1										
Ln(GDP)	0.752	1									
Ln(CPI)	-0.078	-0.063	1								
Ln(PP1)	-0.580	**-0.521**	0.138	1							
Ln(PP2)	0.526	**0.547**	-0.073	**-0.907**	1						
Ln(PP3)	0.632	**0.559**	-0.220	**-0.930**	**0.776**	1					
Ln(PPD)	0.089	0.174	-0.026	-0.246	0.187	0.365	1				
Ln(IMR)	-0.568	**-0.553**	0.251	**0.716**	**-0.635**	**-0.792**	-0.201	1			
Ln(EXR)	-0.218	0.017	0.020	0.055	0.058	-0.154	-0.021	0.116	1		
Ln(TBC)	-0.234	-0.293	-0.017	0.177	-0.150	-0.327	-0.075	**0.591**	0.079	1	
Ln(LIF)	0.301	0.451	-0.100	**-0.551**	**0.603**	**0.587**	0.150	**-0.796**	0.054	**-0.669**	1

Figures 1 and 2 display the paths of estimated coefficients for each of the predictors against $\log\lambda$ using Lasso (top-left panel) and the Elastic net (top-right panel) algorithms for the LI and LMI economies respectively. Each figure indicates how the coefficients entered the model as λ changes. For small values of λ, the estimated coefficients were close to that of the LS. From the paths of the estimated coefficients, it was difficult to identify appropriate values of λ in order to select the optimal model from both the Lasso and the Elastic net algorithms for the two income class data. Also figures 1 and 2 depict the cross-validation (CV) curves for the Lasso (bottom-left panel) and Elastic net (bottom-right panel) showed in red dotted lines with normal standard error bands around them.

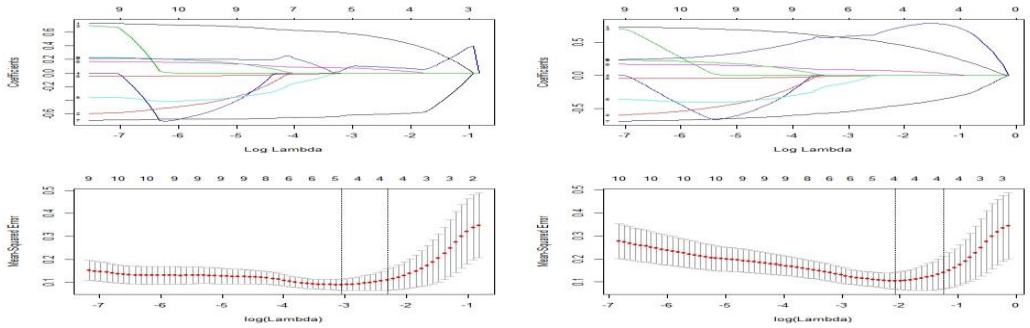

Figure 1: Estimated coefficient paths & CV curves for the LI economies

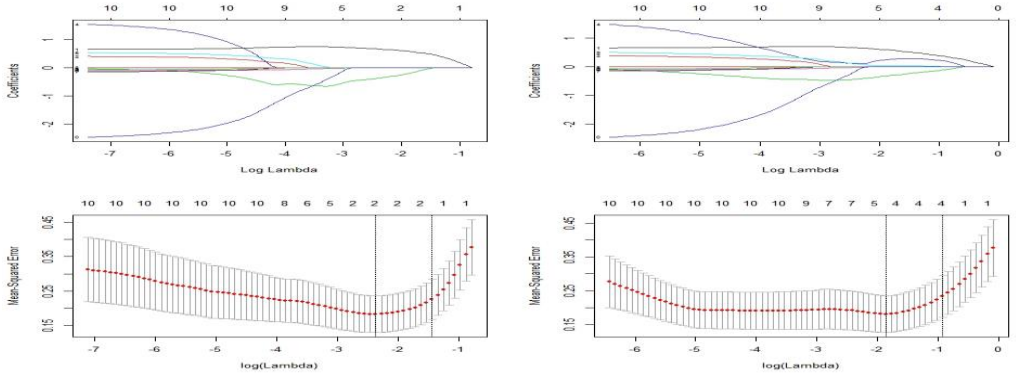

Figure 2: Estimated coefficient paths & CV curves for the LMI economies

Table 5 shows the non-zero estimated coefficients for the two income class data with their corresponding optimal values of λ and MSEs for both the Lasso and Elastic net. For LI economies, the Lasso algorithm estimated five (5) non-zero coefficients corresponding to predictors: Ln(GDP), Ln(PPD), Ln(IMR), Ln(EXR), and Ln(LIF). The Elastic net algorithm however, estimated four (4) non-zero coefficients matching the predictors: Ln(GDP), Ln(PPD), Ln(IMR), and Ln(LIF). The optimal values of λ and the MSEs on the basis of the testing sets were (0.047, 0.341) and (0.124, 0.334) for the Lasso and the Elastic net respectively. A comparison of the MSEs confirms that the Elastic net model provided a better fit. As a predictive model for HEC among LI economies, we recommend the Elastic net model. A closer look at the non-zero coefficients for the Elastic net in the case of LI economies reveals that, when LIF increases by 1%, HEC increases by 0.721%. Similarly, when GDP increases by 1%, HEC increases by 0.462%, and as PPD increases by 1%, HEC again increases by 0.066%. However, HEC decreases by 0.450% when IMR increases by 1%. Life expectancy at birth, GDP, and PPD cause HEC to increase. Life expectancy at birth has the highest impact on HEC,

followed by GDP then PPD. Infant mortality rate however, causes a decrease in HEC.

In the case of the LMI economies, the Lasso algorithm estimated two (2) non-zero coefficients as follows: Ln(GDP) and Ln(PP1). The Elastic net algorithm on the other hand estimated four (4) non-zero coefficients: Ln(GDP), Ln(PP1), Ln(PP2), and Ln(PP3). The ideal values of λ and MSEs on the basis of the testing sets were (0.094, 0.136) and (0.155, 0.137) for the Lasso and the Elastic net respectively. A comparison of the MSEs illustrates that the Lasso is not better in terms of prediction accuracy as compared to Elastic net; however on grounds of parsimony, we chose the Lasso model.

Taking into account the objective of the study, as well as the results of the two key predictive models for HEC, in LMI economies the Lasso model was a better predictor. A further assessment of the Lasso in the case of the LMI economies reveals that, as GDP increases by 1%, HEC increases by 0.658% however, for 1% increase in PP1, HEC rather decreases by 0.370%. While GDP causes HEC to rise, PP1 rather causes it to decrease.

Table 5: Non-zero estimated coefficients

	LI economies		**LMI economies**	
Parameters	**Lasso**	**Elastic net**	**Lasso**	**Elastic net**
Intercept	2.161	-0.202	1.280	0.673
Ln(GDP)	0.581	0.462	0.658	0.583
Ln(CPI)	-	-	-	-
Ln(PP1)	-	-	-0.370	-0.319
Ln(PP2)	-	-	-	0.245
Ln(PP3)	-	-	-	0.039
Ln(PPD)	0.075	0.066	-	-
Ln(IMR)	-0.618	-0.450	-	-
Ln(EXR)	-0.002	-	-	-
Ln(TBC)	-	-	-	-
Ln(LIF)	0.088	0.721	-	-
Optimal (λ)	**0.047**	**0.124**	**0.094**	**0.155**
MSE	**0.341**	**0.334**	**0.136**	**0.137**

4 Discussion and Conclusions

The study results signal the pertinence of regularization as a potentially statistical procedure for isolating essential predictors of HEC in the presence of multicollinearity. Major findings of the study were that, it identified the Elastic net and the Lasso models as critical in accurately estimating the predictors of HEC

among LI and LMI economies respectively. For the LI economies, GDP, PPD, IMR, and LIF were key predictors associated with the Elastic net. While LIF, GDP, and PPD in order of magnitude cause HEC to increase, MMR rather causes HEC to decrease. For the LMI economies, GDP and PP1 were key predicators connected with the Lasso. As GDP causes HEC to increase, PP1 rather causes it to decrease. Also, GDP remains a common predictor of HEC in both economies. The outcome of the current study should first guide policymakers and governments of LI and LMI economies to accurately predict health expenditure. Second, it should help them to have to a very significant extent some amount of control of health care cost by concentrating on the relevant predictors of HEC. Also, the findings herein have implications for future research. First, the current study did not take into consideration the upper middle and high income economies. The inclusion of these income economies to identify key predictors of HEC is justified. Second, Sparse principal component analysis (SPCA), a modified version of principal component analysis (PCA) produces a more parsimonious model. Future research needs to compare the Lasso, the Elastic net to the SPCA.

References

[1] Xu, Ke, Priyanka Saksena and Alberto Holly. 2011. "The determinants of health expenditure: A country-level panel data Analysis." Working paper. Washington, DC: Results for development institute.

[2] Deloitte. 2015 Global health care outlook: Common goals, competing priorities;2015 https://www2.deloitte.com/content/dam/Deloitte/global/Documents/Life-Sciences-Health-Care/gx-lshc-2015-health-care-outlook-global.pdf [accessed 30 June 2015].

[3] The Economist. 2015. "The economist intelligence unit's global outlooks (2015)." http://www.eiu.com/industry/healthcare [accessed 30 June 2015].

[4] Wilson, R. M. 1999. Medical care expenditures and gdp growth in OECD nations. American Association of Behavioral and Social Sciences Journal 2, 159-171.

[5] Baltagi, B. H. and Moscone, F. 2010. Health care expenditure and income in the OECD reconsidered: evidence from panel data. Economic Modelling, 27, 804-811.

[6] Lv, Z., and Zhu, H. 2014. Health care expenditure and gdp in african countries: evidence from semi-parametric estimation with panel data. The Scientific World Journal, 1-6.

[7] The World Bank, Country and lending group. 2016. [http://data.worldbank.org/news/new-country-classifications].

[8] Muhammad, I., Maria, J., and Muhammad, A. R. 2013. Comparison of shrinkable regression for remedy of multicollinearity problem. Middle-East Journal of Scientific Research, 14, 570-579.
[9] Hoerl, A.E. and Kennard, R.W. 1970a. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12, 55-67
[10] Hoerl, A.E. and Kennarq, R.W. 1970b. Ridge regression: applications to non-orthogonal problems. Technometrics 12, 69-82.
[11] Tibshirani, R. J. 1996. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B. 58, 267-288.
[12] Bradley, E., Hastie, T, Johnstone, I, and Tibshirani, R. 2004. Least angle regression. Annals of Statistics, 32(2), 407-499.
[13] Friedman, J. H., and Popescu, B. E. 2004. Gradient directed regularization for linear regression and classification. Stanford University, Department of Statistics. Technical Report.
[14] Zou, H. and Hastie, T. 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B 67, 301 – 320 67, 301 – 320.
[15] The World Bank, World Development Indicators. 2013. Healthcare expenditure per capita ppp for 2011, gdp per capita, ppp (constant 2011 international $), consumer price index (2010 = 100), inflation, consumer prices (annual %), population ages 0-14 (% of total), population ages 15-64 (% of total), population ages 65 and above (% of total), population density (people per sq. km of land area), mortality rate, infant (per 1,000 live births), net ODA received per capita (current US$), official exchange rate (lcu per US$, period average), incidence of tuberculosis (per 100,000 people), and life expectancy at age birth, [Data file]. Retrieved from http://data.worldbank.org/topic .
[16] Gujarati D. N. 1988. Basic econometrics. McGraw-Hill, New York.
[17] Friedman, J., Hastie, T., and Tibshirani, R. 2010. Regularization paths for generalized linear model via coordinate descent, Journal of Statistical Software, 33, 1-22.
[18] James, G., Witten, D., Hastie, T., and Tibshirani, R. 2014. “An introduction to statistical learning: with application in R.”, Springer, New York.