Journal of Kerman University of Medical Sciences

JKMU

Journal of Kerman University of Medical Sciences, 2019; 26 (6): 440-449

Penalized Lasso Methods in Health Data: application to trauma and influenza data of Kerman

Abolfazl Hosseinnataj, Ph.D.[1], Abbas Bahrampour, Ph.D.[2], Mohammad Reza Baneshi, Ph.D.[3], Farzaneh Zolala, Ph.D.[4], Roya Nikbakht, M.Sc.[5], Mehdi Torabi, M.D.[6], Fereshteh Mazidi Sharaf Abadi, M.Sc.[7]

1- Department of Biostatistics and Epidemiology, Modeling in Health Research Center, Faculty of Health, Institute for Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Iran
2- Professor, Department of Biostatistics, Physiology Research Center, Institute of Basic and Clinical Physiology Sciences & Modeling in Health Research Center, Faculty of Health, Institute for Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Iran
3- Professor, Department of Biostatistics and Epidemiology, Modeling in Health Research Center, Faculty of Health, Institute for Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Iran (Corresponding author; E-mail: rbaneshi2@gmail.com)
4- Associate Professor, Department of Biostatistics and Epidemiology, Social Determinants of Health Research Center, Institute for Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Iran
5- Department of Biostatistics and Epidemiology, HIV/STI Surveillance Research Center, and WHO Collaborating Centre for HIV Surveillance, Kerman University of Medical Sciences, Kerman, Iran
6- Associate Professor, Department of Emergency Medicine, Kerman University of Medical Sciences, Kerman, Iran
7- Department of Emergency Medicine, Kerman University of Medical Sciences, Kerman, Iran
Received: 28 April, 2019 Accepted: 21 December, 2019

ARTICLE INFO

Article type:
Original Article

Keywords:
Multi-collinearity
High dimension
Penalized regression
Lasso

Abstract

Background: Two main issues that challenge model building are number of Events Per Variable and multicollinearity among exploratory variables. Our aim is to review statistical methods that tackle these issues with emphasize on penalized Lasso regression model. The present study aimed to explain problems of traditional regressions due to small sample size and multi-colinearity in trauma and influenza data and to introduce Lasso regression as the most modern shrinkage method.

Methods: Two data sets, corresponded to Events Per Variable of 1.5 and 3.4, were used. The outcomes of these two data sets were hospitalization due to trauma and hospitalization of patients suffering influenza respectively. In total, four models were developed: classic Cox and logistic regression models, as well as their penalized lasso form. The tuning parameters were selected through 10-fold cross validation.

Results: Traditional Cox model was not able to detect significance of any of variables. Lasso Cox model revealed significance of respiratory rate, focused assessment with sonography in trauma, difference between blood sugar on admission and 3 h after admission, and international normalized ratio. In the second data set, while lasso logistic selected four variables as being significant, classic logistic was able to identify only the importance of one variable.

Conclusion: The AIC for lasso models was lower than that for traditional regression models. Lasso method has practical appeal when Events Per Variable is low and multicollinearity exists in the data.

Citation: Hosseinnataj A, Bahrampour A, Baneshi M.R, Zolala F, Nikbakht R, Torabi M, Mazidi Sharaf F. Penalized Lasso Methods in Health Data: application to trauma and influenza data of Kerman. *Journal of Kerman University of Medical Sciences*, 2019; 26 (6): 440-449.

Introduction

Two main issues that challenge the practice of regression modeling are number of independent variables candidate to the offered model, and presence of multicollinearity among the independent variables (1).

It has been shown that estimated parameters are robust when Event Per Variable (EPV) is at least 10. In the linear and Poisson regression models, EPV simply mean the sample size. In the logistic regression, EPV refers to the minimum of number of cases and controls. In survival analysis, EPV refers to the number of subjects experience the event of interest (2). It has been shown that the lower the EPV, the less stable the regression coefficients and its Standard Errors (SE).

The second challenge is the presence of multicollinearity. When there is multicollinearity, the effect of correlated variables may interfere with each other; for example, two independent variables may be significant in the univariate regression analysis but they may not have a significant effect in multivariate analysis. Other problems with this type of data include the insignificancy of the important variables, change of direction and intensity of the coefficients of independent variables, and inflation in variance of parameters (3).

Nowadays due to the advancement of science in the variety of fields such as genetics, bioinformatics and microarray data generation in cancer research, data sets are available in which the number of independent variables is much more than the number of observations ($p >> n$). such data sets are referred to as high dimensional data (4). Main characteristics of high dimensional data are that EPV is extremely low and multicollinearity exists (5). In the case of high dimensional data and correlated independent variables, estimated regression coefficients are unbiased but inclusion or exclusion of few cases remarkably affects the regression coefficients. This means that variance is high. The idea of bias-variance trade-off can be explained from another angel. Assume we have an under-fit when all egression coefficients are zero. Mean Square Error (MSE) would be high no matter whether the model is applied on training or test sets. This means that the bias is high but the variance is low. On the other hand, when a model is over-fitted, it shows good performance in train but poor in test set. Here the bias is low but the variance is high (6). Therefore, modeling of these data sets needs special statistical tools.

The traditional approach, to tackle these problems, is to offer a reduced set of independent variables to the multifactorial model. The reduced set is selected through a series of univariate regression. Main problems are that an insignificant variable in univariate analysis might show a different behavior in the presence of other variables. In addition to that, sometimes the researcher wishes to adjust the effect of several confounders regardless of their statistical significance (7). Another solution is to combine independent variables and offer the combined variable(s) to the multifactorial model. Methods such as Principal Component Analysis (PCA) and Partial Least Square (PLS) are frequently applied to create small number of components out of huge number of variables. These methods are usually not welcomed because of difficult of interpretation (8).

In the recent years penalized regression models, also known as shrinkage methods, are proposed. Shrinkage methods aim at reduction of variance by incorporating a tolerable degree of bias in parameter estimation. These methods shrink the regression coefficients towards zero by applying a penalty term. Two of the most important shrinkage methods are Ridge and Lasso (9).

These models are applicable even when EPV is low and strong multicollinearity exists.

One of the biggest problems with Ridge method is to keep the coefficients of all variables in the model. This means that the coefficients of none of the variables would be exactly zero. If the goal of the researcher is to adjust the effect of all of confounder variables in the model, the ridge method is useful. However, when the aim is to develop a parsimonious model for prediction purposes, Lasso regression is applied. This method shrinks regression coefficients towards zero and therefore can be used as a common tools for variable selection in high dimension data.

Therefore, the purpose of this paper is to introduce Lasso regression as the most modern shrinkage method. We will explain how the method works in a practical way. Two data sets are used to demonstrate the practicality of the method and compare to traditional methods.

Subjects and Methods

First data set

The first data set comprised of 280 patients with multiple trauma that were referred to Bahonar Hospital, Level II Trauma Center, in southeast Iran. Multiple trauma patients with Injury Severity Score > 16 and older than 18 years, from 1 September 2015 to 1 September 2016, were enrolled in this study. Patients referred after an hour with a history of chronic lung, kidney, heart, or liver diseases or diabetes, anticoagulant medication consumption, drug or alcohol intoxication, or shock (except hemorrhagic shock) were excluded from the study. Time to death during hospitalization was used as outcome. Information of 36 independent variables was available. Some of the most important independent variables include blood sugar (BS) on admission and 3 h after admission, as well as their difference (i.e. ΔBS), INR (International Normalized Ratio), serum lactate, RR (Respiratory Rate), FAST (Focused Assessment with Sonography in Trauma) and etc.

Second data set

An H1N1 influenza happened in 2015 in Kerman. Binary outcome was hospitalization due to influenza (H1N1). The case group included 85 patients who were admitted to the hospital and the control group included 51 patients who referred to the flu-like symptoms and were discharged after an ambulatory examination and outpatient treatment. Both groups were interviewed alike with regard to the risk factors of the disease. Totally, 15 independent variables were received such as age, sex, diabetes status (yes or no), asthma status (yes or no), pulmonary diseases (yes or no), smoking and etc.

Statistical Methods

Lasso regression is the widely used model among the shrinkage methods and its parameters are derived by maximizing the following equation

$$l(\beta)_c = \ln(l(\beta)) - \lambda \sum_{j=1}^{p} |\beta_j|$$

Where: ln (l(β)) is the natural logarithm of the likelihood function and the penalty term $\lambda \sum_{j=1}^{p} |\beta_j|$ shrinks the unimportant regression coefficients towards zero. When λ equals zero, the equation reduces to the classic regression and the traditional maximum likelihood methods are applied to estimate the coefficients. Therefore, all coefficients will remain in the model and thus the model will have low bias and high variance (10).

Therefore, λ is called the tuning parameter which has values greater than or equal to zero. Increase in λ is associated with

more weight to the penalty term and therefore more coefficients would become zero (11). Thus the model's bias will increase and the variance decreases.

An important step in Lasso regression analysis is selection of appropriate value of λ through cross validation. A range of values, say 0 to 10+6, for λ is proposed. In this method, the data is divided into k equal sized subsamples. In each step, one part of data is used as test and the rest as train set. The model is constructed using train set and is applied on test set. MSE is calculated for all values of λ. λ and the model corresponded to the lowest MSE is selected (12-14).

It is worth mentioning that the statistics used to assess the performance of the model depends on the outcome type. In linear, logistic and Poisson indices such as MSE and R^2 are used. In the survival regression, deviance is usually used (15).

In trauma data that the response variable is follow up time (the period between hospitalization to the death or discharge), the Lasso Cox regression is used to identify important independent variables. To check the proportional hazard (PH) assumption, the time interaction test was applied. In the influenza data which has a binary response variable, Lasso logistic regression is used.

To demonstrate the performance of the Lasso models, we compared them with the traditional models Cox and logistic methods. Models were compared in terms of goodness of fit [Akiake Information Criterion (AIC)], and number of variables shows significant association with the outcome. The analyses were performed using the R software. R codes are displayed in the appendix.

Results

First data set: hospitalization of trauma patients

Data set comprised of 280 patients of which 54 were hospitalized. Number of independent variables was 36, corresponded to EPV of 1.5. The highest correlation between independent variables was 0.95. Out of 630 pairwise correlations, 211 ones were above 0.2 and 85 ones were above 0.5. Degree of correlation between some of independent variables is depicted in figure 1, left panel. In this figure, blue and red colors are used for positive and negative correlation, respectively. The larger the circle, the correlation is stronger.

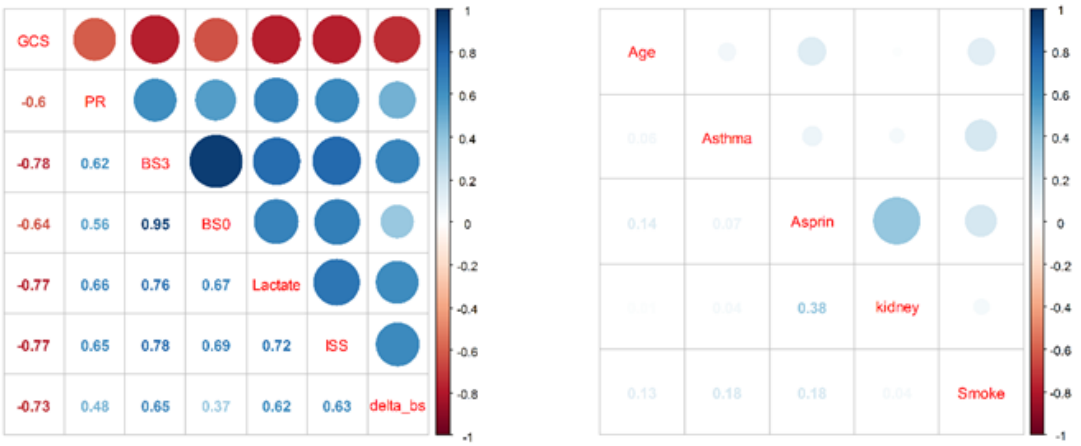

Figure 1. Degree of correlation between some of independent variables on trauma data (left panel) and influenza data (right panel)

In figure2, left panel shows model deviance against different values of log (λ). According to this graph, the lowest deviance was associated to log (λ) of -3.65 (i.e. λ=0.026).

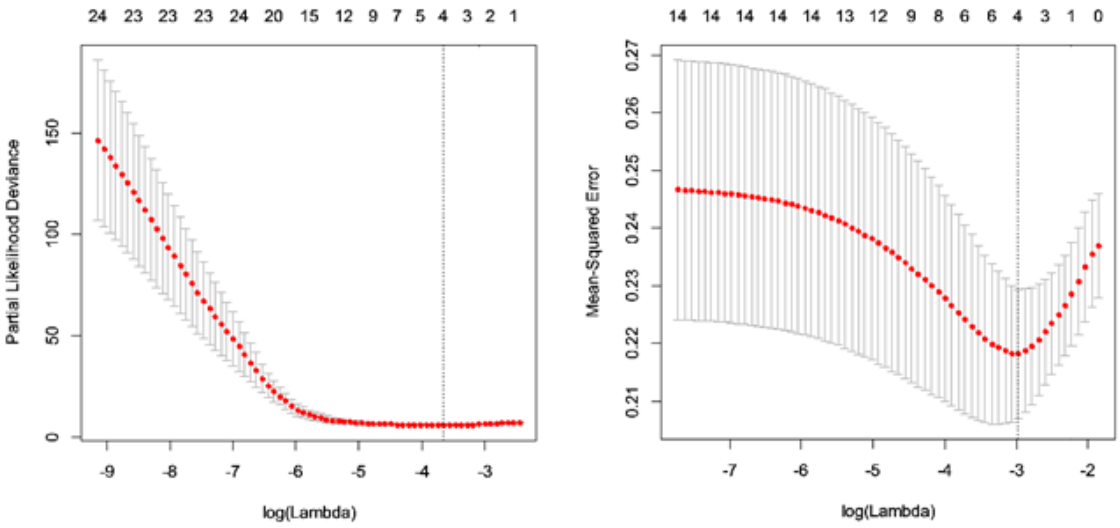

Figure 2. Partial likelihood deviance (left panel, trauma data) and mean square error (right panel, influenza data) versus logarithm of tuning parameters

The left panel of figure 3 is a scatter plot of regression coefficients versus log (λ).

As the amount of log (λ) increases, less number of variables will remain in the model.

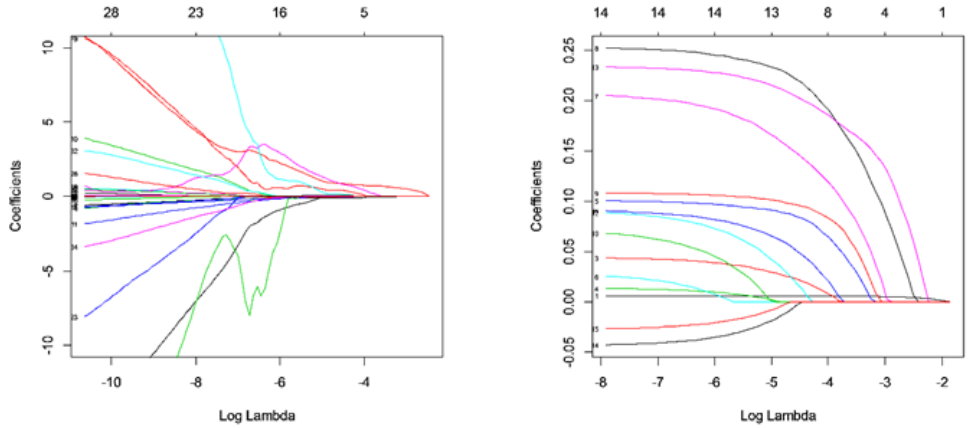

Figure 3. The standardized cox (left panel, trauma data) and logistic lasso regression coefficient (right panel, influenza data) versus values of tuning parameters

At the optimum λ of 0.026, four variables retained in the final model: INR, FAST, ΔBS and RR. Estimated regression coefficients and hazard ratios are provided in Table 1. As it is shown in table 1, increase in INR was associated with increase in hazard ratio (HR=2.01). FAST was associated with increase in the likelihood of mortality (HR=1.06). The hazard ratio increased by 1.06 for each unit change in ΔBS. Increase in RR was associated with decrease in hazard ratio (HR=0.96). AIC criteria for this model were 65.94.

Table1. The standardized significant coefficients in Cox and logistic lasso models

Cox lasso model			Logistic lasso model		
variable	value	HR	**variable**	value	OR
INR[a]	0.698	2.01	**pulmonary diseases**	0.682	1.978
FAST[b]	0.058	1.06	**diabetes**	0.419	1.52
ΔBS[c]	0.046	1.047	**age**	0.022	1.022
RR[d]	-0.038	0.963	**asthma**	0.010	1.01

[a]: International Normalized Ratio
[b]: Focused Assessment with Sonography in Trauma
[c]: difference between blood sugar 3 h after admission and on admission
[d]: Respiratory Rate

AIC corresponded to traditional Cox model was 166.97. Moreover, none of variables reached significant level. Also, regression coefficient for two variables was not calculated due to non-convergence. Variance of coefficient for three variables was higher than 1,000.

Second data set: hospitalization due to influenza

The EPV in this data set was 3.4. Maximum pairwise correlation between independent variables was 0.4. Figure 2.b presents different amounts of log (λ) versus MSE. The best value for λ was 0.051. Regression coefficients and odds ratio are shown in figure 3 right panel. Patients with pulmonary diseases (OR=1.98), diabetes (OR=1.52), asthma (OR=1.01) and also older patients (OR=1.02) were more likely to be hospitalized.

Using the optimal amount of λ, the model was fitted and finally the four variables of age, underlying asthma, diabetes and pulmonary diseases had a significant effect. AIC criteria for lasso logistic model were 48.94. Corresponding figure for traditional logistic model was 195.53. Traditional model was able to capture just significance of age.

Discussion

Our results demonstrate the benefit of lasso models, over traditional models. The performance of the lasso models was superior to traditional models in terms of AIC. The AIC for lasso models was lower than that for traditional regression models. It has been shown that in the case of correlated predictors, which is a common problem in high dimension data, traditional methods are not appropriate (16). This is because such methods are highly likely to select noise variables as being significant.

In this study we demonstrate the practicality of penalized regression models with emphasize on lasso method. While penalized methods introduce bias in parameter estimation, these methods control the variance. Another advantage of the Lasso method is that the method simultaneously achieves two goals of regression methods: selection of important variables and parameter estimation (17). The tuning parameter forces some regression coefficients equal to zero and therefore, only strong predictors remain in the model.

In the trauma data, the Cox lasso regression selected four variables RR (Respiratory Rate), INR (International Normalized Ratio), FAST and ΔBS. Torabi et al analyzed this data applying standard Cox regression model with backward elimination process. They have found significance of ΔBS, HR (Human Resources), and INR but not RR. Torabi et al showed that patients with higher difference between blood sugar 3 h after admission and on admission (ΔBS) are more likely to die during hospitalization (18). Also, other parameters such as INR, FAST and RR can be helpful in predicting hospital mortality in multiple trauma patients. Increase in FAST and INR is associated with increase of the hazard of death. The opposite was true with respect to RR. In addition, the mentioned factors had significant relation with mortality of trauma patients (19-22).

In the influenza data, 4 risk factors were determined as factors influencing the odds of hospitalization. Older patients, and those suffering asthma, diabetes and pulmonary diseases were more likely to be hospitalized. In other studies, these factors were significantly associated with hospitalization (23-26).

Lasso regression model has some disadvantages. For example, the maximum number of independent non-zero

coefficients in the model is equal to the sample size. Another important limitation is that all coefficients have the same penalty; in other words, less important and important variables shrink to a similar degree (27). Several alternative methods have been proposed to solve this problem; one of the most important is the adaptive lasso. In this method, for each coefficient, a separate penalty is considered (28). Also, the elastic net method is useful when there is high autocorrelation and categorical variables in the data. In this method, the penalty term is a combination penalty applied in Lasso and Ridge (29).

Another limitation of lasso model is that it does not pave the way for formal hypothesis testing. The ordinary lasso does not address the uncertainty of parameter estimation; standard errors for β's are not immediately available (30). When $p > n$ or even p close to n, parameter estimates unstable, since standard errors are likely to be high (31). There is room to incorporate Bayesian analysis in the context of penalized regression methods. In this method, statistical distributions for model parameters (such as variable coefficients, variance, and tuning parameters) are considered, which are referred to as the prior distribution. With the product of the prior distribution and the likelihood function, the posterior distribution of the data is obtained and the analysis and estimation of the parameters are obtained through the posterior distribution (32). Currently, using simulated data, we are trying to integrate Bayesian inference with penalized likelihood methods, to better understand situations in which Bayesian methods improve the model fitness.

Conclusion

Two advantages of lasso regression compared to traditional regression are: first, this model is useful for high dimension data and second, lasso regression model can be used for multicollinearity problems. Also, lasso regression can be applied for prediction rsponse as traditional regression. For example hazard ratio can be calculated for a new case by lasso cox and lasso logistic regression.

Acknowledgement

This research is part of the first author's PhD dissertation. We are very grateful to Kerman University of Medical Sciences for providing data.

References

1. Austin PC, Steyerberg EW. The number of subjects per variable required in linear regression analyses. J Clin Epidemiol 2015; 68(6):627-36.
2. Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. Am J Epidemiol 2007; 165(6):710-8.
3. Lin FJ. Solving multicollinearity in the process of fitting regression model using the nested estimate procedure. Quality & Quantity 2008; 42(3):417-26.
4. Clarke R, Ressom HW, Wang A, Xuan J, Liu MC, Gehan EA, et al. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nat Rev Cancer 2008; 8(1):37-49.
5. Li H, Gui J. Partial Cox regression analysis for high-dimensional microarray gene expression data. Bioinformatics 2004; 20(suppl 1):i208-15.
6. Pinsky PF, Magder LS. Evaluating the tradeoff between bias and variance through use of prior

probabilities. Commun Stat Simul Comput 1997; 26(2):399-421.

7. Slinker BK, Glantz SA. Multiple regression for physiological data analysis: the problem of multicollinearity. Am J Physiol 1985; 249(1 Pt 2):R1-12.

8. Hammami D, Lee TS, Ouarda TB, Lee J. Predictor selection for downscaling GCM data with LASSO. Journal of Geophysical Research 2012; 117(D17116).

9. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 1996; 58(1):267-88.

10. Tian GL, Tang ML, Fang HB, Tan M. Efficient methods for estimating constrained parameters with applications to lasso logistic regression. Comput Stat Data Anal 2008; 52(7):3528-42.

11. Jang DH, Anderson-Cook CM. Influence Plots for LASSO. [cited ?????] Available from: https://www.osti.gov/pages/biblio/1337112-influence-plots-lasso.

12. Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. Ann Stat 2006; 34(3):1436-62.

13. Benner A, Zucknick M, Hielscher T, Ittrich C, Mansmann U. High-dimensional Cox models: the choice of penalty as part of the model building process. *Biom J* 2010; 52(1):50-69.

14. Roberts S, Nowak G. Stabilizing the lasso against cross-validation variability. Computational Statistics & Data Analysis 2014; 70:198-211.

15. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for some traditional and novel measures. Epidemiology 2010; 21(1):128-38.

16. Murphy TB, Dean N, Raftery AE. Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications. Ann Appl Stat 2010; 4(1):396-421.

17. Huang H. Controlling the false discoveries in LASSO. Biometrics 2017; 73(4):1102-10.

18. Torabi M, Mazidi Sharaf Abadi F, Baneshi MR. Blood sugar changes and hospital mortality in multiple trauma. Am J Emerg Med 2018; 36(5):816-19.

19. Isbell C, Cohn SM, Inaba K, O'Keeffe T, De Moya M, Demissie S, et al. Cirrhosis, operative trauma, transfusion, and mortality: a multicenter retrospective observational study. Cureus 2018; 10(8):e3087.

20. Cirocchi R, Grassi V, De Sol A, Renzi C, Parisi A, Parisi G, et al. Diagnostic, therapeutic and health-care management protocol for major abdominal trauma at the "Santa Maria" Hospital of Terni. Analysis of the results after two years. Ann Ital Chir 2018; 89:540-51.

21. Froberg L, Helgstrand F, Clausen C, Steinmetz J, Eckardt H. Mortality in trauma patients with active arterial bleeding managed by embolization or surgical packing: an observational cohort study of 66 patients. J Emerg Trauma Shock 2016; 9(3):107-14.

22. Duane TM, Ivatury RR, Dechert T, Brown H, Wolfe LG, Malhotra AK, et al. Blood glucose levels at 24 hours after trauma fails to predict outcomes. J Trauma 2008; 64(5):1184-7.

23. Ono S, Ono Y, Matsui H, Yasunaga H. Factors associated with hospitalization for seasonal influenza in a Japanese nonelderly cohort. BMC Public Health 2016; 16:922.

24. Czaja CA, Miller L, Alden N, Wald HL, Cummings CN, Rolfes MA, et al. Age-related differences in hospitalization rates, clinical presentation, and outcomes among older adults hospitalized with influenza-U.S. Influenza Hospitalization Surveillance Network (FluSurv-NET). Open Forum Infect Dis 2019; 6(7): pii: ofz225.

25. Homaira N, Briggs N, Oei JL, Hilder L, Bajuk B, Snelling T, et al. Impact of influenza on hospitalization rates in children with a range of chronic lung diseases. Influenza Other Respir Viruses 2019; 13(3):233-9.

26. Tempia S, Walaza S, Moyes J, Cohen AL, von Mollendorf C, Treurnicht FK, et al. Risk factors for influenza-associated severe acute respiratory illness hospitalization in South Africa, 2012-2015. Open Forum Infect Dis 2017; 4(1):ofw262.

27. Radchenko P, James GM. Variable Inclusion and shrinkage algorithms. J Am Stat Assoc 2008; 103(483):1304-15.

28. Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc 2006; 101(476):1418-29.

29. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B 2005; 67(2):301-20.

30. Mallick H, Yi N. Bayesian methods for high dimensional linear models. J Biom Biostat 2013; 1:005.

31. Reid S, Tibshirani R, Friedman J. A study of error variance estimation in lasso regression. Statistica Sinica 2016; 26:35-67.

32. Park T, Casella G. The bayesian lasso. J Am Stat Assoc 2008; 103(482):681-6.