CELL JOURNAL (Yakhteh)

www.celljournal.org

Supplementary Information for

Regularized Machine Learning Models for Prediction of Metabolic Syndrome Using *GCKR, APOA5,* and *BUD13* Gene Variants: Tehran Cardiometabolic Genetic Study

Nadia Alipour, Ph.D.[1], Anoshirvan Kazemnejad, Ph.D.[1]*, Mahdi Akbarzadeh, Ph.D.[2], Farzad Eskandari, Ph.D.[3], Asiyeh Sadat Zahedi, MSc.[2], Maryam S Daneshpour, Ph.D.[2]*

1. Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran
2. Cellular and Molecular Endocrine Research Centre, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran
3. Department of Statistics, Faculty of Statistics, Mathematics and Computer, Allameh Tabataba'i University, Tehran, Iran

Regularized machine learning models

Consider the general linear regression model

$$y_i = b_0 + \sum_{j=1}^{k} b_j x_{ij} + \varepsilon_j$$

where y_i (i=1,…, n; n is the sample size) is the ith observation of the response variable, x_{ij} is the ith observation of the jth covariate ((j=1,…, k), b_j's), are the regression coefficients, ε_j are i.i.d. random error terms which $\varepsilon_j \sim N(0, \sigma^2)$. The ordinary least squares estimator of $b=(b_0, b_1, \ldots, b_p)'$, is obtained by minimizing the residual sum of squares (RSS), i.e.,

$$b_{ols} = \underset{b}{argmin} \sum_{i=1}^{n} (y_i - b_0 - \sum_{j=1}^{k} b_j x_{ij})^2,$$

Regularized machine learning (RML) models avoid the overfitting problem by penalizing the model complexity and adding a nonnegative regularization term to the log-likelihood function and, consequently, shrink the values of regression coefficients (1, 2).

$$RML = RSS + \lambda\, \omega(b)$$

Where $\lambda\, \omega(b)$ is the regularization term. As the amount of shrinkage is controlled by the regularization parameter $\lambda \geq 0$, choosing λ is an essential part of model fitting. A larger value of λ, leads to a greater amount of shrinkage. This study used a 10-fold Cross-validation method to select the optimal λ value for our models. Considering several regularization terms that have been previously proposed (1, 3-6), we applied five popular regularized machine learning models to select relevant features for MetS predicting models include the least absolute shrinkage and selection operator (LASSO) (1) , ridge regression (RR) (3), elastic net (ENET) (6), adaptive LASSO (aLASSO) (5), and adaptive elastic-net (aENET). Supplementary table 1 shows the penalty factors and features of applied methods in the present study.

Received: 25/April/2023, Revised: 08/June/2023, Accepted: 19/June/2023
*Corresponding Addresses: P.O.Box: 14115-111, Department of Biostatistics, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran
P.O.Box: 19195-4763, Cellular and Molecular Endocrine Research Centre, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran
Emails: kazem_an@modares.ac.ir, daneshpour@sbmu.ac.ir

Royan Institute
Cell Journal (Yakhteh)

Table S1: Penalty factors and features of regularized machine learning methods

Models	λ Penalty factor (ω(b))	Features
LASSO		Perform variable selection by shrinking some predictor coefficients to be exactly zero.
		Improve the interpretability of the model.
	$\lambda \sum_{j=1}^{k} \lvert b_j \rvert$	Does not capture any grouping effect.
		May fail when there is a strong correlation between predictors.
Ridge		Shrinks the coefficients' estimates towards zero but does not force them to be exactly zero.
		Does not perform variable selection, meaning all features are retained even if they are not important.
	$\lambda \sum_{j=1}^{k} bj^2$	Helps in reducing the impact of multicollinearity among predictors.
		Works well when there are many relevant features.
ENET		Combines the strengths of LASSO and ridge regression.
	$\lambda_1 \sum_{j=1}^{k} \lvert b_j \rvert + \lambda_2 \sum_{j=1}^{k} b_j^2$	Simultaneously perform variable selection and shrink coefficients of correlated variables.
		Works well when there are many relevant features and some of them are highly correlated.
		Can handle correlated predictors better than LASSO.
aLASSO		Have all the good properties of the lasso.
		Simultaneously estimate and variable selection.
	$\lambda \sum_{j=1}^{k} \lvert w_j b_j \rvert$	Has the oracle properties.
		This method used adaptive weights for penalizing different coefficients to overcome the inconsistency of LASSO.
aENET		Has the oracle properties.
	$\lambda_1 \sum_{j=1}^{k} \lvert w_j b_j \rvert + \lambda_2 \sum_{j=1}^{k} w_j b_j^2$	Can handle the collinearity.
		Combines the advantages of both elastic net and adaptive LASSO.

LASSO; Least absolute shrinkage and selection operator, ENET; Elastic Net, aLASSO; Adaptive LASSO, and aENET; Adaptive Elastic Net.

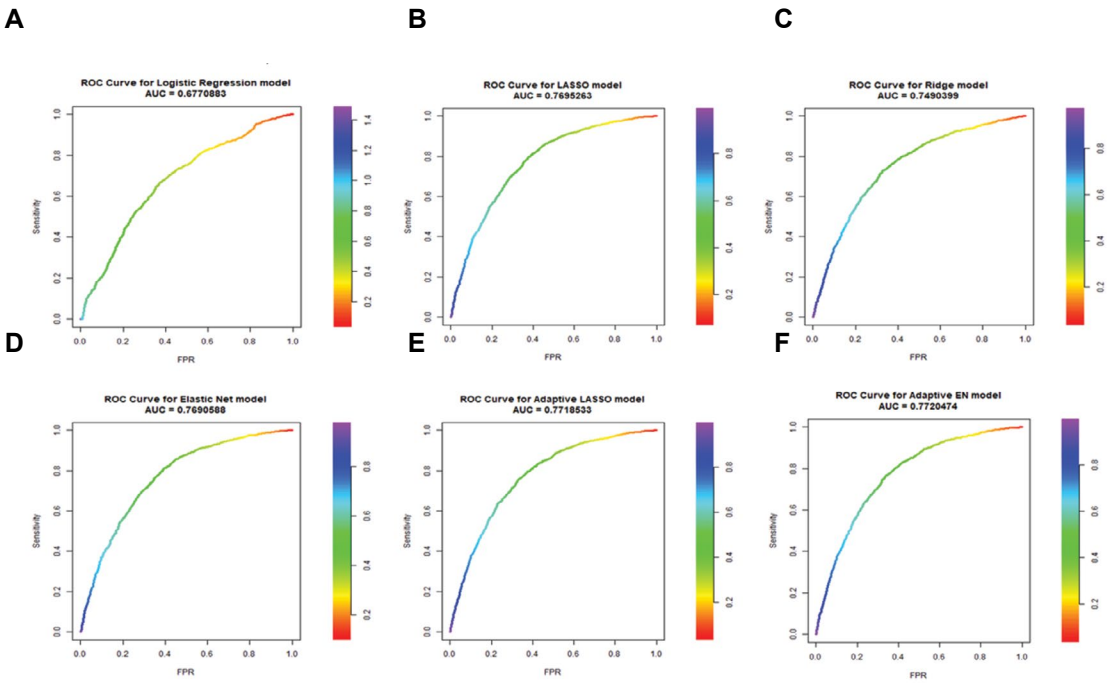

Fig.S1: ROC curves for penalized and logistic regression methods. **A.** ROC curve for logistic regression, **B.** ROC curve for LASSO regression, **C.** ROC curve for ridge regression, **D.** ROC curve for elastic net model, **E.** ROC curve for adaptive lasso model, **F.** ROC curve for adaptive elastic net model; ROC; Receiver operating characteristics, AUC-ROC; Area under the ROC curve, LASSO; Least absolute shrinkage and selection operator, EN; Elastic net, and FPR; False positive rate. The adaptive EN model shows a higher AUC-ROC curve.

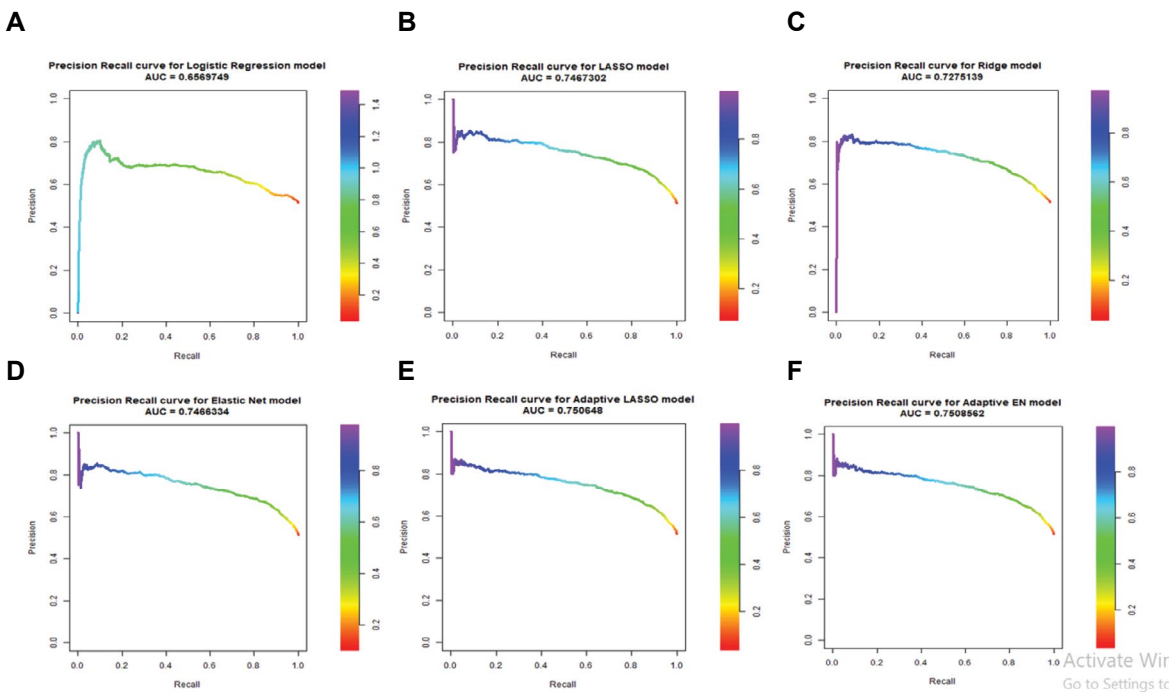

Fig.S2: Precision-recall (PR) curves for penalized and logistic regression methods. **A.** PR curve for logistic regression, **B.** PR curve for LASSO regression, **C.** PR curve for ridge regression, **D.** PR curve for elastic net model, **E.** PR curve for adaptive lasso model, **F.** PR curve for adaptive elastic net model. AUC-PR; Area under the precision-recall curve, LASSO; Least absolute shrinkage and selection operator, EN; Elastic net, and FPR; False positive rate. The adaptive EN model shows a higher AUC-PR curve.

References

1. Tibshirani R. Regression shrinkage and selection via the lasso. J R Statist Soc B. 1996; 58(1): 267-288.
2. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Statist Soc B. 2005; 67(2): 301-320.
3. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970; 12(1): 55-67.
4. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Statist Soc B. 2005; 67(2): 301-320.
5. Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc. 2006; 101(476): 1418-1429.
6. Zou H, Zhang HH. On the adaptive elastic-net with a diverging number of parameters. Ann Stat. 2009; 37(4): 1733-1751.