

SHRINKAGE REGRESSION METHODS APPLIED TO AGRICULTURE

Suna Akkol

Yuzuncu Yıl University, Faculty of Agriculture, Department of Animal Science, Van-Turkey

Abstract

In Multiple Linear Regression a common goal is to determine which independent variables contribute significantly to explaining the variability in the dependent variable. Multicollinearity occurs when there is a linear relationship among one or more of the independent variables. It is a statistical term for a problem that is common in technical analysis. When there is multicollinearity among the explanatory variables in the multiple regression model, the properties of BLUE is no longer effective. Shrinkage regression that refers to shrinkage methods of estimation or prediction in multiple linear regression can be used to obtain more robust results of the data than the ordinary least square (OLS). In this study it will be presented Ridge Regression, LASSO and Elastic-net that are shrinkage methods.

Key words: Ordinary Least square, LASSO, Ridge Regression, Elastic-net multicollinearity

INTRODUCTION

The general purpose of multiple linear regression is to learn more about the relationship between several independent or input variables, ($X^T = X_1, X_2, \dots, X_p$) and a dependent or output variable, Y. Typically we have a set of data $(x_1, y_1) \dots (x_n, y_n)$. The multiple linear regression has the general form with the n independent values.

$$Y = X\beta + e.$$

Here, Y is an $n \times 1$ vector of responses, X is an $n \times p$ matrix of the predictor variables, β is a $p \times 1$ vector of unknown constants, e is an $n \times 1$ vector of random errors, with $e_i \sim NIID(0, \sigma_e^2)$. The Ordinary Least Square (OLS) estimates are obtained by minimizing residuals sum of square (RSS) and are most favorable estimation method in multiple linear regression models (MLP). But in some cases, OLS solutions will not be reliable or unique. For instance, OLS estimation in MLR, the number of independent variables should be much smaller than the number of observation. But the data sets collected different field of study are generally very large in terms of number of variables (p) and often small in terms of number of the samples (n). This problem is often termed as the “large p and small n problem” ($p \gg n$) (Acharjee et al, 2013). In this case the OLS solution will not be unique and regression coefficients tend to be unstable. Some of the different independent variables can be significantly correlated to each other in many data sets. This problem is called collinearity or multicollinearity. Especially in wide data sets there will be collinearity because of $p \gg n$.

To handle the multicollinearity is problem some shrinkage methods such as LASSO (Least Absolute Shrinkage and Selection Operator) Regression (Tibshirani, 1996), Ridge regression (Hoerl and Kennard, 1970) and Elastic net (Zou and Hastie 2005) could be use rather than OLS (Acharjee, 2013). A discrete process exhibits high variance and it doesn't reduce the prediction error of the full model. However, shrinkage methods are more continuous and don't suffer as much from high variability (Hastie et al, 2009).

1. MULTIPLE LENEAR REGRESSION and ORDINARY LEAST SQUARE (OLS)

The multiple linear regression model has the following form with the n observed values

$$Y = X\beta + e \tag{1}$$

In the above equation, the response variables are $Y = (y_1, y_1, \dots, y_n)^T$, the error terms are $e = (e_1, e_2, \dots, e_n)^T$, the regression coefficients are $\beta = (\beta_0, \beta_1, \dots, \beta_n)^T$ and the design matrix is X (Filzmoser, 2008).

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{i1} & x_{i2} & \cdots & x_{ip} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} 1, & x_1 \\ 1, & x_2 \\ \vdots & \vdots \\ 1, & x_i \\ \vdots & \vdots \\ 1, & x_n \end{bmatrix}$$

In OLS the coefficients β 's are estimated by minimizing the residual sum of squares (RSS)

$$\begin{aligned} RSS(\beta) &= \sum_{i=1}^n (y_i - f(x_i))^2 \\ &= \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij}\beta_j \right)^2 \end{aligned} \quad (2)$$

We can write it in matrix notation as below

$$RSS(\beta) = (y - X\beta)^T (y - X\beta) \quad (3)$$

Optimization problem is

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|Y - X\beta\|^2 \quad (4.a)$$

In other words it can be written as

$$\hat{\beta}^{ols} = \underset{\beta}{\operatorname{argmin}} RSS(\beta) \quad (4.b)$$

Therefor the OLS estimation of β ,

$$\hat{\beta} = (X^T X)^{-1} X^T y \quad (5)$$

and the estimated y-values are respectively given by,

$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y \quad (6)$$

Hence, the OLS estimates of the parameters β , have the smallest variance among all linear unbiased estimates. If the data set has the large p and small n problem, parameters cannot be uniquely defined using OLS. In the case of many correlated variables, the OLS estimators can have a small bias but a very large variance. Prediction accuracy could sometimes be improved by shrinking or setting some coefficients to zero (Hastie et al, 2009). At the same time, we often would like to determine a smaller subset that exhibits the strongest effects, with a large number of predictors. It can be used as shrinkage methods as a variable selection methods.

2. SHRINKAGE METHODS

Shrinkage methods have been proposed to improve OLS. The regression coefficients in shrinkage methods are shrunk by imposing a penalty on their size. The used penalty term is called as regularization parameter. So shrinkage methods are also called penalization methods. Ridge regression, LASSO and Elastic net among others are continuous penalization methods. They use and shrinkage factors and these factors can take any value between zero and infinity. The value of the shrinkage parameter decides the amount of penalization applied to the regression coefficients (Acharjee et al, 2013).

2.1. Ridge Regression (RR)

In ridge regression it is imposed a penalty on the sum of squares of the regression coefficients. The ridge coefficients minimize a penalized residual sum of square (Hastie et al, 2009),

$$\hat{\beta}^{ridge} = \arg \min_{\beta} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda_2 \sum_{j=1}^p \beta_j^2 \right\} \quad (7)$$

An equivalent form is written

$$\hat{\beta}^{ridge} = \arg \min_{\beta} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 \right\}$$

$$\text{subject to } \sum_{j=1}^p \beta_j^2 \leq t \quad (8)$$

Here λ_2 or t being similar to each other one is regularization parameter for ridge regression and controls the amount of shrinkage. $\sum_{j=1}^p \beta_j^2$ is known as L_2 norm. The value of λ_2 is defined the amount of shrinkage. The coefficients are shrunk towards zero (and towards each other). When $\lambda_2 = 0$, we get the linear regression estimate, $\lambda_2 = 1$, $\lambda_2 = \infty$ we get $\hat{\beta}^{ridge} = 0$, for λ_2 in between, we are balancing two ideas (Tibshirani, 1996).

So the bias introduces but variance of estimate reduces. After the centering has been done, the input matrix X has p columns. The penalized residual sum of square in matrix form is

$$RSS(\lambda) = (y - X\beta)^T (y - X\beta) + \lambda_2 \beta^T \beta \quad (9)$$

and the ridge regression estimate is defined as the value of β that minimizes

$$\hat{\beta}^{ridge} = (X^T X + \lambda_2 I)^{-1} X^T y \quad (10)$$

Here, I is the $(p \times p)$ identity matrix. Notice that with the choice of quadratic penalty $\beta^T \beta$ in (9) the ridge regression solution is again a linear function of y (Hastie et al, 2009).

2.2 The LASSO

Another shrinkage method is LASSO (Least Absolute Shrinkage and Selection Operator) works like ridge regression (Tibshirani, 1996). In LASSO, the penalty is applied to the sum of the absolute values of the regression coefficients. LASSO can be written in the *Lagrangian form* as,

$$\hat{\beta}^{lasso} = \arg \min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda_1 \sum_{j=1}^p |\beta_j| \right\} \quad (11)$$

In the equation (11), LASSO penalty, $\sum_{j=1}^p |\beta_j|$ is called L_1 norm and λ_1 being regularization parameter controls the amount of shrinkage. It is different than L_2 norm and makes the solutions nonlinear. So a quadratic programming is used compute the LASSO solution. Penalize by absolute value of parameter

$$\hat{\beta}^{lasso} = \arg \min_{\beta} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 \right\}$$

$$\text{subject to } \sum_{j=1}^p |\beta_j| \leq t$$
(12)

If t described in equation (12) is chosen small enough then the LASSO will set some coefficients exactly to zero. As a result of this, LASSO is a kind of continuous model selection. Like in ridge regression When $\lambda_2 = 0$, we get the linear regression estimate, $\lambda_2 = 1$, $\lambda_2 = \infty$ we get $\hat{\beta}^{LASSO} = 0$, for λ_2 in between, we are balancing two ideas (Tibshirani, 1996).

The LASSO has an increasingly popular method. LASSO method developed can be summarized as follows: Zou (2006) and Wang et al. (2007) proposed to use adaptive LASSO (ALASSO) that gives consistent model selection. Yuan and Lin (2005) studied an empirical Bayes method. Park and Casella (2008) studied Bayesian LASSO (BLASSO) to exploit model inference via posterior distributions. Hans (2010) considers a formal Bayesian approach to exploring model uncertainty with LASSO type priors on parameters in submodels. BOLASSO, is a bootstrapped version and gives faster convergence was studied by Bach (2008). Group-LASSO activated the variables by group is proposed by Yuan and Lin (2006). Bondell and Reich (2008) proposed the OSCAR (octagonal shrinkage and clustering algorithm for regression) approach. These methods mentioned above will not be presented in this study.

2.3 Elastic Net

If there is high correlation among the independent variables, the LASSO tends to only select one variable among a group of highly correlated variables; and its prediction performance may not be as good as the ridge regression (Bondell and Reich, 2008). To overcome these limitations, Zou and Hastie (2005) proposed the Elastic net method using a weighted combination of L_1 and L_2 penalties. For any fixed non-negative λ_1 and λ_2 , the naïve elastic net criterion is defined (Zou and Hastie 2005).

$$L(\lambda_1, \lambda_2, \beta) = |y - X\beta|^2 + \lambda_2 |\beta|_2^2 + \lambda_1 |\beta|_1$$
(13)

$$|\beta|_2^2 = \sum_{j=1}^p \beta_j^2,$$

$$|\beta|_1 = \sum_{j=1}^p |\beta_j|,$$
(14)

The naïve elastic net estimators $\hat{\beta}$ is obtained by minimizing equation (13)

$$\hat{\beta}^{enet} = \arg \min_{\beta} |y - X\beta|^2 + \lambda_2 |\beta|_2^2 + \lambda_1 |\beta|_1$$
(15.a)

$$\hat{\beta}^{enet} = \arg \min_{\beta} (y - X\beta)^T (y - X\beta) + \lambda J(\beta)$$
(15.b)

The equation (15.a) is equivalent equation (15.b). Where J is nonnegative penalty function and $\lambda \geq 0$ regularization parameter. Let regularization parameter $\lambda = \lambda_1 + \lambda_2$ that is sum of two nonnegative penalties and let $\alpha = \lambda_2 / (\lambda_1 + \lambda_2)$. If $J(\beta) = \alpha |\beta|_2^2 + (1 - \alpha) |\beta|_1$ in equation (15.b), estimation obtained is a naïve estimate, $\hat{\beta}_N$. The naïve elastic net estimator finds the ridge regression coefficient for each λ_2 and then does shrinkage along the lasso coefficient solution. This procedure is a double amount of shrinkage that does not reduce variance and introduces unnecessary extra bias. Zou and Hastie (2005) introduced a rescaled version of the naïve estimate and called elastic net. The elastic net estimation is given by,

$$\hat{\beta}^{enet} = (1 + \lambda_2)\hat{\beta}_N \quad (16)$$

By rescaling the naïve elastic net estimator by $(1 + \lambda_2)$, the ridge shrinkage effect may be eliminated, then the LASSO shrinkage achieves a good prediction performance and sparsity. So elastic-net is a different compromise between ridge and LASSO. That is elastic-net selects variables like the LASSO, and shrinks together the coefficients of correlated predictors like ridge (Hastie et al., 2008)

3. DISCUSSION

The three approaches discussed so far are ridge regression, the LASSO and elastic net. All of them are shrinkage methods. If the input matrix X is orthogonal they use simply modifications. Ridge regression and LASSO apply a simple transformation to the least squares estimate. Ridge multiplies least squares coefficients by a constant < 1 , LASSO converts them towards zero by a constant, truncating at zero (Hastie et al, 2009; Zhou, 2013). In other words, there is a relationship between OLS with LASSO and ridge regression in the orthogonal case such as:

$$\hat{\beta}_j^{ridge} = \frac{1}{1 + \lambda_2} \hat{\beta}_j^{ols} \quad (17)$$

$$\hat{\beta}_j^{lasso} = \text{sign}(\hat{\beta}_j^{ols})(|\hat{\beta}_j^{ols}| - \lambda_1/2)_+$$

Where “+” denotes positive part and $\hat{\beta}_j^{ols} = x_j^T y$. That is the OLS solution j^{th} regression coefficient. When there are only two parameters in the nonorthogonal case it is given the picture of the LASSO and ridge regression. The Lasso is at the left of the picture and ridge regression at the right. The constraint region for ridge regression is the disk $|\beta_1| + |\beta_2| \leq t$, but the constraint region for lasso is the diamond $\beta_1^2 + \beta_2^2 \leq t^2$ (Hastie et al, 2009)

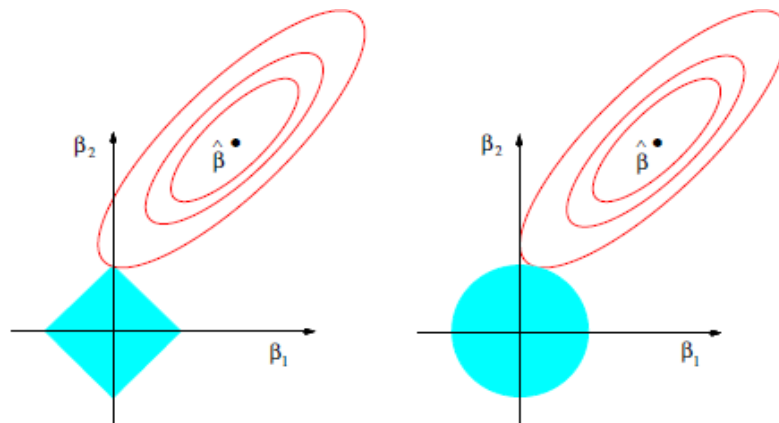


Figure1. Estimation picture for the LASSO (left) and ridge regression (right). Contours are shown of the error and constraint functions. The solid blue areas are the constraint regions The solid blue areas are the constraint region $|\beta_1| + |\beta_2| \leq t$ and $\beta_1^2 + \beta_2^2 \leq t^2$, respectively, while the red ellipses are the contours of the least squares error function (after Hastie et al, 2009) .

In ridge regression, all the input variables stay in the model since regression coefficients do not become exactly zero (that would be equivalent to variables dropping out of the regression model). Ridge gives equal weight to absolutely correlated variables in the data set (Hastie et al, 2009). LASSO penalizes the absolute values of the regression coefficients. By penalizing a number of the estimated coefficients will become exactly zero. Some of the input variables drop out of the regression model. In case of absolutely correlated variables, LASSO just selects one of these variables and ignores the rest in the group (Hastie et al, 2009; Acharjee, 2013).

Hastie et al (2009) and Zhou (2013) presented that all the linear regression techniques under a common framework can be given

$$\hat{\beta} = \arg \min_{\beta} \left\{ \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j|^q \right\} \quad (18)$$

In equation (18), λ includes bias, q indicates a prior distribution on β . The value $q = 0$ corresponds to variable subset selection, as the penalty simply counts the number of nonzero parameters; $q = 1$ corresponds to the LASSO, while $q = 2$ to ridge regression. Hastie et al (2009) showed that the contours of constant value of $\sum_j |\beta_j|^q$ for given the values of q . They are given in Figure 2.

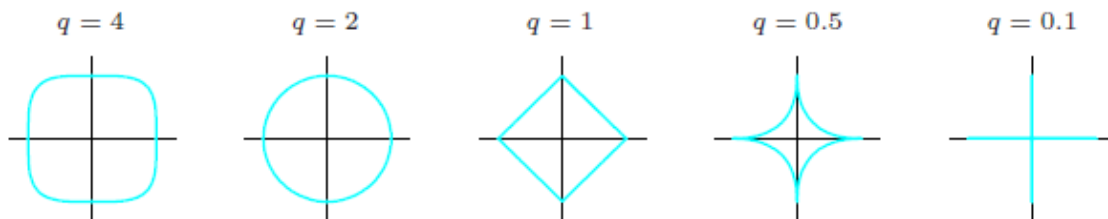


Figure 2. Contours of constant value of $\sum_j |\beta_j|^q$ for given values of q (after Hastie et al, 2009).

The regularization resulting from LASSO's L_1 penalty leads to sparse solutions. Sparse models are more interpretable and often preferred in the sciences and social sciences. LASSO provides a computationally feasible way for model selection. However, in order to use LASSO for model selection, it is necessary to assess how well the sparse model given by LASSO relates to the true model (Zhao and Yu, 2006).

Ridge regression minimizes RSS by using the L_2 norm of the coefficients and shrinks the coefficients continuously toward zero. However, in Ridge regression all the predictors has retained in the model. So it is not done the variable selection. The LASSO uses the L_1 norm of the coefficients to minimize RSS. It does both continuous shrinkage and variable selection at the same time. Therefore, the LASSO has some of the good features of Ridge regression. Especially when there is a multicollinearity problem in the data set, it can be made the variable selection with the LASSO successfully.

When there is the “large p and small n problem” ($p \gg n$) and the LASSO selects at most n variables. The reason for this is the LASSO has the nature of the convex optimization problem. This looks like to be a restrictive feature for a variable selection method. However, if a group variable is a very high correlation, then the LASSO inclines to select only one variable from the group. So the LASSO cannot perform “grouped selection” (Zou and Hastie 2005). Particularly in microarray data set where the number of the predictors is bigger than the number of observation, it is asked that is performed grouped selection. To overcome the limitations of the LASSO, elastic net was introduced by purposed by Zou and Hastie (2005).

The elastic net does automatic variable selection and continuous shrinkage like the LASSO. Let $\sum_j |\beta_j|^q = L_q$ as given by Figure.3 where compares the L_q penalty with $q = 1.2$ and the elastic-net penalty with $\alpha = 2$ (Hastie et.al. 2009).

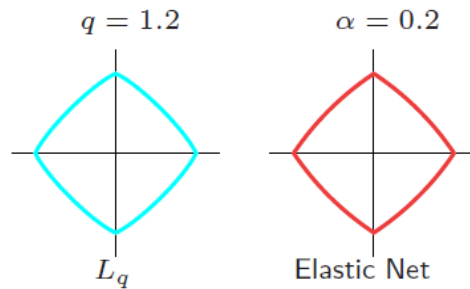


Figure 3. Contours of constant value $\sum_j |\beta_j|^q = L_q$ for $q = 1.2$ (left plot), and the elastic-net penalty $J(\beta) = \alpha |\beta|_2^2 + (1 - \alpha) |\beta|_1$ for $\alpha = 0.2$ (right plot). Although visually very similar, the elastic-net has sharp (non-differentiable) corners, while the $q = 1.2$ penalty does not (after Hastie et al, 2009).

The Elastic-net is a combination of the lasso penalty and the ridge penalty. It has ability to perform grouped selection and is proper for the $p \gg n$ problem. However Elastic net makes simultaneous regularization and variable selection. Zou and Hastie (2005) used the elastic net to do microarray classification and automatic gene selection, and announced they had achieved good results by using Elastic net.

REFERENCES

- Acharjee A, Finkers R, Visser RGF, Maliepaard C 2013, "Comparison of Regularized Regression Methods for ~Omics Data". *Metabolomics* vol.3, pp.126.
- Bach F R, 2008, "Bo lasso: Model Consistent Lasso Estimation through the Bootstrap Appearing in Proceedings of the 25th International Conference on Machine Learning", Helsinki, Finland, 2008. http://www.di.ens.fr/~fbach/fbach_bolasso_icml2008.pdf
- Bondell H D and Reich B J 2008, "Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR", *Biometrics*, vol. 64, pp.115–123.
- Efron, B, Hastie, T, Johnstone, I and Tibshirani, R, 2004, "Least angle regression". *Ann. Statist.*, vol.32, pp.407–499.
- Filzmoser, P, 2008, "Linear and nonlinear methods for regression and classification and applications in R", <http://www.statistik.tuwien.ac.at/forschung/CS/CS-2008-3complete.pdf>
- Hans, C, 2010, "Model uncertainty and variable selection in Bayesian LASSO regression", *Statistics and Computing*, vol.20, pp.221–229.
- Hastie, T, Tibshirani, R, Friedman, J, 2009, "The elements of statistical learning: Data mining, inference, and prediction" (2nd Edn), *Springer, New York, USA*.
- Hans, C, 2011. Elastic Net Regression Modeling With the Orthant Normal Prior. *Journal of the American Statistical Association*, Vol. 106, pp. 1383–1393
- Hoerl, AE, Kennard, RW, 1970, "Ridge regression: Biased estimation for nonorthogonal problems", *Technometrics* vol.12, pp. 55–67.
- Park, ., Casella, G, 2008, 'The Bayesian Lasso' *Journal of the American Statistical Association*, vol.103, pp. 681–686.
- Tibshirani, R, 1996, 'Regression shrinkage and selection via the LASSO' *Journal of the Royal Statistical Society*, vol. 58, pp. 267–288.
- Wang, H, Li, G, Tsai, CL, 2007, 'Regression coefficients and autoregressive order shrinkage and selection via the Lasso' *Journal of the Royal Statistical Society, Series B*, vol. 69, pp. 63–78.
- Yuan, M, Lin, Y, 2006, 'Model selection and estimation in regression with grouped variables', *Journal of the Royal Statistical Society, Series B*, vol.68, pp.49–67.

Zhao, P and Yu, B, 2006, 'On Model Selection Consistency of Lasso' *Journal of Machine Learning Research*, vol.7, pp. 2541-2563.

Zhou J, 2013, 'Linear Methods for Regression. Center for Evolutionary Medicine and Informatics' accessed on: January 21 st 2013, available on

<http://www.public.asu.edu/~jzhou29/slides/Ch3Regression.pdf>

Zou H and Hastie T (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, vol. 67, pp. 301–320.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101, pp.1418–1429.