American Journal of Intelligent Systems 2014, 4(4): 142-147
DOI: 10.5923/j.ajis.20140404.03

Finding Optimal Value for the Shrinkage Parameter in Ridge Regression via Particle Swarm Optimization

Vedide Rezan Uslu[1], Erol Egrioglu[2,*], Eren Bas[3]

[1]Department of Statistics, University of Ondokuz Mayis, Samsun, 55139, Turkey
[2]Department of Statistics, Marmara University, Istanbul, 34722, Turkey
[3]Department of Statistics, Giresun University, Giresun, 28000, Turkey

Abstract A multiple regression model has got the standard assumptions. If the data can not satisfy these assumptions some problems which have some serious undesired effects on the parameter estimates arise. One of the problems is called multicollinearity which means that there is a nearly perfect linear relationship between explanatory variables used in a multiple regression model. This undesirable problem is generally solved by using methods such as Ridge regression which gives the biased parameter estimates. Ridge regression shrinks the ordinary least squares estimation vector of regression coefficients towards origin, allowing with a bias but providing a smaller variance. However, the choice of shrinkage parameter *k* in ridge regression is another serious issue. In this study, a new algorithm based on particle swarm optimization is proposed to find optimal shrinkage parameter.

Keywords Ridge regression, Optimal shrinkage parameter, Particle swarm optimization

1. Introduction

Linear regression method is a classic statistical method. Linear regression method has a lot of assumptions like other statistical techniques. These assumptions are not realistic in the real world application. These assumptions are checked by statisticians. If they are not suitable for data, advanced statistical techniques are applied to the data. Ridge regression is a kind of advanced statistical technique. When data has multicollinearity problem, ridge regression technique can give a solution for data. In this study, a new ridge regression method is introduced.

Consider a linear multiple

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

where Y is the $(n\times 1)$ vector of observations of the dependent variable, X is the $(n\times p')$ matrix of observations of explanatory variables with full rank p, β is the $(p'\times 1)$ vector of unknown parameters and $\in$ is the $(n\times 1)$ vector of random error, where $p' = p+1$ and p shows the number of explanatory variables in the model. It is assumed that each random error has zero mean and a constant variance σ^2 and that they are uncorrelated. Moreover it is assumed that the columns of X should not be in a linear dependency of each other.

Let us denote the columns of X as $X_1, X_2, \cdots, X_p$. If there is a relationship

$$\sum_{j=1}^{p} t_j X_j \cong 0 \tag{2}$$

for a set of numbers such as $t_1, t_2, \cdots, t_p$, not all zero, the relation is called the multicollinearity problem in multiple regression analysis.

The presence of multicollinearity has a number of potential serious effects on the ordinary least squares estimates of the unknown parameters. The most serious one is that it results in the large variances and covariance of the least squares estimates of the regression coefficients. Therefore it implies that different samples taken at the same level of X's could lead completely different estimates of the model parameters.

Multicollinearity can also cause to produce least squares estimates of β's which are too large in absolute value.

When the columns of X matrix are centered and scaled the matrix $X'X$ becomes the correlation matrix of the explanatory variables and $X'Y$ is the vector of the correlation coefficients of the dependent variable with each explanatory variable. If the columns X are orthogonal, $X'X$ matrix is a unit matrix. In the presence of multicollinearity $X'X$ becomes ill-conditioned which means that it is nearly singular and the determinant of it is

* Corresponding author:
erole@omu.edu.tr (Erol Egrioglu)
Published online at http://journal.sapub.org/ajis

nearly zero. Some of the eigenvalues of $X'X$ can also be very near to zero. Some prefer to examine

$$\phi = \frac{\lambda_{\max}}{\lambda_{\min}} \tag{3}$$

which is called as condition number. In this equation λ is the eigenvalues of $X'X$. Generally if the condition number is less than 100, there is no serious multicollinearity problem. Condition numbers between 100 and 1000 imply moderate to strong multicollinearity and if it exceeds 1000 it indicates that severe multicollinearity exists in the data. The variance-covariance matrix of β is determined by $\left(X'X\right)^{-1}$.

$$Var\left(\hat{\beta}\right) = \sigma^2\left(X'X\right)^{-1} \tag{4}$$

The diagonal elements of this matrix are called the variance inflation factors (*VIF*) and are given by

$$VIF_j = \left(\frac{1}{1-R_j^2}\right) \tag{5}$$

where R_j^2 is the determination coefficient obtained from the multiple regression of X_j on the remaining $(p-1)$ regressor variables in the model. If there is a strong collinearity between X_j and any subset of the remaining regressor variable the value of R_j^2 will be close to 1. Therefore VIF_j is going to be very large and it implies that the variance of β_j is to be large. Briefly speaking the following items can be considered as the multicollinearity diagnostics.

1. The correlation matrix constructed by*X*
2. The determinant of the matrix of $X'X$
3. The eigenvalues of $X'X$
4. VIFvalues

(Montgomery and Peck [1])

To overcome multicollinearity problem, the ridge regression has been suggested in the literature (Hoerl and Kennard [2], Hoerl et al. [3]). But there is another problem for applying ridge regression such as finding the optimal biasing parameter (*k*) value. Several methods have been proposed for finding it. These are; Hoerl and Kennard [2], Hoerl et al. [3], Mc Donald and Galarneau [4], Lawless and Wang [5], Hocking [6], Wichern and Curchill [7], Nordberg [8], Praga-Alejo et al. [9], Al Hassan [10], Ahn et al. [11]. And also, Siray et al. [12] proposed an approach to examine multicollinearity and autocorrelation problems.

In this study, a new algorithm of estimating *k* value by using particle swarm optimization was introduced.

In addition to these studies, there are some studies in the literature about ridge estimation and its estimators. For example, Sakallıoglu and Kacıranlar [14] proposed a new biased estimator for the vector of parameters in a linear regression model based on ridge estimation. Firinguetti and Bobadilla [14] proposed an approach to develop asymptotic confidence intervals for the model parameters based on ridge regression estimates. Tabakan and Akdeniz [15] proposed a new difference – based ridge estimator of parameters in partial linear model. Duran and Akdeniz [16] proposed an estimator named modified jackknifed Liu-type estimator to show its efficiency in ridge regression. Uemukai [17] showed the small sample properties of a ridge regression estimator when there exists omitted variables by inspiring the study of Huang [18]. Akdeniz [19] proposed new biased estimators under the LINEX loss function. And also, there are some combining methods about new estimators such as Alkhamisi [20].

The rest part of the paper can be outlined as below: The second section of the article is about Ridge regression. The methodology of the paper was given in Section 3. The implementation of our proposed method was given in Section 4 and finally, discussions were presented in Section 5.

2. Ridge Regression

In presence of multicollinearity, there are several remedies recommended for avoiding from its undesirable effects on the estimates. Ridge regression is one of the remedies mostly employed. It was firstly proposed by Hoerl and Kennard [2]. In this method the estimates of the regression coefficients are obtained with a little bias guaranteed a smaller variance by adding a very small positive number in the diagonal elements of $X'X$. While the least squares estimates of regression coefficients are

$$\hat{\beta} = \left(X'X\right)^{-1}X'Y \tag{6}$$

the ridge estimates are introduced as

$$\hat{\beta}_R = \left(X'X + kI\right)^{-1}X'Y \tag{7}$$

where *k* is a very small constant determined by the researcher Hoerl and Kennard [2]. Gauss Markov theorem states that under the standard assumptions about errors the ordinary least squares estimators (OLS) of the parameters of the model in (1) are unbiased and have the minimum variances. But there is no guarantee that the variance of $\hat{\beta}$ will be small. For this purpose the ridge estimator estimates β with a bias but has a smaller variance than the ordinary least squares estimators' one. When we look at the mean squared error of $\hat{\beta}_R$ we can easily see that

$$MSE\left(\hat{\beta}_R\right) = E\left(\hat{\beta}_R - \beta\right)^2 = Var\left(\hat{\beta}_R\right) + \left[E\left(\hat{\beta}_R\right) - \beta\right]^2 \tag{8}$$

can be made small than the mean squared error of $\hat{\beta}$ which is equal to variance of $\hat{\beta}$ since there is no bias in it.

The ridge estimator can be expressed as a linear transformation of the ordinary least squares estimator as below.

$$\begin{aligned}\hat{\beta}_R &= \left(X'X + kI\right)^{-1} X'Y \\ &= \left(X'X + kI\right)^{-1}(X'X)\hat{\beta} = Z\hat{\beta}\end{aligned} \tag{9}$$

When we look at the expected value of ridge estimator, we can easily see that it is a biased estimator of β.

$$E\left(\hat{\beta}_R\right) = E\left(Z\hat{\beta}\right) = Z\beta \tag{10}$$

The variance-covariance matrix of $\hat{\beta}_R$ is

$$\begin{aligned}&Var\left(\hat{\beta}_R\right) = Var\left(Z\hat{\beta}\right) = ZVar\left(\hat{\beta}\right)Z' \\ &= Z\left(\sigma^2\left(X'X\right)^{-1}\right)Z' \\ &= \sigma^2\left(X'X + kI\right)^{-1}\left(X'X\right)\left(X'X + kI\right)^{-1}\end{aligned} \tag{11}$$

Let us look at the mean squared error of two estimators to compare them. Since the ordinary least squares estimator is unbiased, the mean squared error will be the variance of the estimator.

$$\begin{aligned}MSE\left(\hat{\beta}\right) &= E\left(\hat{\beta} - \beta\right)'\left(\hat{\beta} - \beta\right) = TrVar\left(\hat{\beta}\right) \\ &= \sigma^2 Tr\left(X'X\right)^{-1} = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}\end{aligned} \tag{12}$$

where λ_j is the j^{th} eigenvalues of $X'X$. Contrarily the mean squared error of ridge estimator is

$$\begin{aligned}&MSE\left(\hat{\beta}_R\right) = Var\left(\hat{\beta}_R\right) + Bias^2 \\ &= \sigma^2 Tr\left[\left(X'X + kI\right)^{-1}\left(X'X\right)\left(X'X + kI\right)^{-1}\right] + \\ &\qquad k^2\beta\left(X'X + kI\right)^{-2}\beta \\ &= \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{\left(\lambda_j + k\right)^2} + k^2\beta\left(X'X + kI\right)^{-2}\beta\end{aligned} \tag{13}$$

Notice from (13) $Var\left(\hat{\beta}\right) > Var\left(\hat{\beta}_R\right)$ can be made by choosing an optimal *k* value. Hoerl and Kennard [2] proved that there is nonzero *k* value for which $MSE\left(\hat{\beta}_R\right)$ is less than $MSE\left(\hat{\beta}\right)$ provided that $\beta'\beta$ is bounded. On the other hand the mean squared error based on the ridge estimator is also compound of two parts; one part which is the first term of the right-hand side of (13) decreases and the other increases when *k* increases. The residual sum of squares based on the ridge estimator can be expressed as below.

$$\begin{aligned}SSE\left(\hat{\beta}_R\right) &= \left(Y - X\hat{\beta}_R\right)'\left(Y - X\hat{\beta}_R\right) \\ &= \left(Y - X\hat{\beta}\right)'\left(Y - X\hat{\beta}\right) + \left(\hat{\beta}_R - \hat{\beta}\right)'\left(\hat{\beta}_R - \hat{\beta}\right)\end{aligned} \tag{14}$$

This expression implies that as *k* increases the sum of squares of residual increases and consequently R^2 decreases. Therefore the ridge estimate will not give the best fit to the data necessarily when we are more interested in obtaining a stable set of parameter estimates.

From this point, we face the question how we can find an appropriate value for biasing parameter *k*. Ridge trace is one of the methods which are used for it. It is a plot of the elements of the ridge estimator versus *k* usually in the interval (0, 1) (Hoerl and Kennard [21]). Marquardt and Snee [22] suggested using only 25 values of *k*, spaced approximately logarithmically over that interval. From the ridge trace, the researchers can see that at a reasonable *k* value the estimates become stable. In this paper for the purpose of comparing the results we just consider the methods of which a brief introduction is given as below.

Hoerl et al. [3] suggested another method for finding *k* value which is given as

$$k = \frac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}} \tag{15}$$

where $\hat{\sigma}^2$ and $\hat{\beta}$ are the ordinary least squares estimates. This method is referred as *fixed point ridge regression method*. For ease of use we will symbolize this method as FPRRM.

Hoerl and Kennard [23] introduced an iterative method for finding the optimal *k* value. In this method *k* is calculated as in below;

$$k_t = \frac{p\hat{\sigma}^2(t-1)}{\hat{\beta}(t-1)'\hat{\beta}(t-1)} \tag{16}$$

where $\hat{\sigma}^2(t-1)$ and $\hat{\beta}(t-1)$ are the corresponding residual mean square and the estimate vector of regression coefficients at *(t-1)th* iteration, respectively. Generally, the initials are chosen the results from the least squares method. The method will be presented here as *iterative ridge regression method* (IRRM) for abbreviation.

3. Methodology

Finding optimal *k* value has always been problematic. In recently, genetic algorithm has been used for this purpose such as Praga-Alejo et al. [9] and Ahn et al. [11] did. Praga-Alejo et al. [9] found the optimal *k* value by minimizing a distance based on *VIFs*. Ahn et al.[11], differently from Praga-Alejo et al. [9], used SSE as fitness function. Praga-Alejo et al. [9] found the optimal *k* value as nearly 1 because of regarding the minimizing of only *VIF* values. On the other hand Ahn et al. [11] found *k* as nearly zero because they minimize *SSE*. Consequently when *k* value is near to 1 the magnitude of the bias of the estimator, therefore *SSE* is becoming very large as *VIF* values are less than 10, which means that there is no multicollinearity problem. If *k* value is very near to zero then SSE is almost

near to the result from the least squares but we cannot get any improvement for *VIF* values. In order to overcome these deficiencies a method which takes into consideration simultaneously both criteria, for finding the optimal *k*was proposed. In this study the proposed method firstly finds the *k* value which make the *VIF's* smaller, that is, less than 10 and *SSE* minimum, at the same time. The optimization problem in the proposed method can be constructed as below.

$$\text{Objective function: } \min_k MAPE(k)+\phi(k) \quad (17)$$

with subject to: $0 \le k \le 1$

where *MAPE(k)* and $\phi(k)$ can be defined as below:

$$\varphi(k)=\begin{cases}0 & \forall VIF_j<10,\ j=1,2,\cdots,p\\ \sum_{j=1}^{p} VIF_j & otherwise\end{cases} \quad (18)$$

$$MAPE(k)=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right| \quad (19)$$

The optimization problem defined as in (17) is solved via particle swarm optimization in the proposed method. Particle swarm optimization, which is an artificial intelligence technique, was firstly proposed by Kennedy and Eberhart [24]. The algorithm of the proposed method is introduced as in the following steps.

Algorithm.

Step 1 The parameters such as *pn*, c_1, c_2 etc., are determined. These parameters are as follows:

pn: particle number of swarm

c_1 : Cognitive coefficient

c_2 : Social coefficient interval

maxt: Maximum iteration number

w: Inertia weight

Step 2 Generate a random initial positions and velocities. The initial positions and velocities are generated by uniform distribution with (*0,1*) parameters. Each particle has one velocity and one position which represent *k* value. x_m^t represents the position of particle *m* at iteration *t* and v_m^t represents the velocity of the particle *m* at iteration *t*.

Step 3 The Fitness function is defined as in (17). Fitness values of the particles are calculated.

Step 4 According to fitness values, *Pbest* and *Gbest* particles given in (20) and (21), respectively, are determined.

$$Pbest_m^t=(pm),\ m=1,2,\cdots,pn \quad (20)$$

$$Gbest^t=(pg) \quad (21)$$

Pbest is constructed by the best results obtained in the related positions at iteration *t*. *Gbest* is the best result in the swarm at iteration *t*.

Step 5 New velocities and positions of the particles are calculated by using the formulas given in (22) and (23).

$$v_m^{t+1}=\begin{bmatrix} w\times v_m^t+c_1\times rand_1\times\left(Pbest_m^t-x_m^t\right)+\\ c_2\times rand_2\times\left(Gbest^t-x_m^t\right)\end{bmatrix} \quad (22)$$

$$x_m^{t+1}=x_m^t+v_m^{t+1} \quad (23)$$

where $rand_1$ and $rand_2$ are random numbers which are generated from *U(0,1).*

Step 6 Repeat from Step 3 to Step 6 until *t<maxt*.

Step 7. The optimal *k* value is obtained as *Gbest*.

4. Implementation

The proposed algorithm has been experienced on two different and well known real data sets in order to investigate the progress provided by the algorithm. These two data sets are known as “Import Data” and “Longley Data”. Import data has been analyzed by Samprit and Hadi [25]. The variables are imports (IMPORT-Y), domestic production (DOPROD-X1), stock formation (STOCK-X2) and domestic consumption (CONSUM-X3), all measured in billions of French francs for the years 1949 through 1959. Longley’s data set is a classic example of multicollinear data (Samprit and Hadi [25]).

Import data and Longley data have been solved by using fixed point method (FPMRRM) (Hoerl et al. [3]), iterative method (IPRRM) (Hoerl and Kennard [23]) and the algorithm proposed in this paper. In the algorithm PSO parameters were chosen as pn=30, w=0.9, $c_1=c_2=2$ and *max t* =100. In the iterative ridge method the stopping criteria has been chosen as $\varepsilon=10^{-6}$. The results were presented in the Table 1 and Table 2, respectively.

In the tables given below; PRM represents the proposed ridge method and SC represents the Standardized Coefficients.

As it can be seen from these tables, the *k* value obtained from our algorithm has provided most optimal VIF values which are smaller than 10. It implies there is no more multicollinearity problem in the data. *k* values from the other techniques have made VIF values smaller than those from ordinary least squares method, but they are not sufficiently small, still. Moreover the value of MAPE has been reduced by the proposed algorithm, comparing with MAPE from OLS.

Table 1. The Coefficient Estimates, VIF Values and SSE and MAPE Obtained From OLS, FPRRM, IRRM and PRM for Import Data

	OLS (k=0)		**FPRRM (k=0.0016)**		**IRRM (k=0.0042)**		**PRM (k=0.0090)**	
Variable	S.C.	VIF	S.C.	VIF	S.C.	VIF	S.C.	VIF
X1	-0.34	186.11	-0.03	72.09	0.16	27.99	0.29	9.99
X2	0.21	1.02	0.22	1.00	0.22	1.00	0.22	0.98
X3	1.30	186.00	0.99	72.13	0.80	28.01	0.67	10.0
SSE	0.0810		0.0086		0.0095		0.0103	
MAPE	0.1196		0.1097		0.1139		0.1185	

Table 2. The Coefficient Estimates, VIF Values and SSE and MAPE Obtained from OLS, FPRRM, IRRM and PRM for Longley Data

	OLS (k=0)		**FPRRM (k=0.00036)**		**IRRM (k=0.0014)**		**PRM (k=0.0172)**	
Variable	S.C.	VIF	S.C.	VIF	S.C.	VIF	S.C.	VIF
X1	0.05	135.53	-0.01	87.31	0.04	56.89	0.02	9.99
X2	-1.01	1788.5	-0.25	472.15	0.17	88.98	0.10	1.54
X3	-0.54	33.60	-0.43	10.94	-0.36	4.36	-0.37	2.55
X4	-0.20	3.59	-0.18	2.87	-0.16	2.55	-0.16	1.95
X5	-0.10	399.15	-0.28	180.21	-0.25	82.21	-0.27	7.32
X6	2.48	758.98	1.88	309.82	1.32	119.4	1.44	4.07
SSE	0.0045		0.0050		0.0064		0.0123	
MAPE	0.0887		0.0753		0.0839		0.0138	

5. Discussion

In the regression analyze, the variances of the estimated parameters and the residual sum of squares as a goodness of fit measure are desired to be very small. When there exists multicollinearity problem unfortunately the property of being minimum variances of the ordinary least squares estimates does not satisfy anymore. The ridge regression is one of the remedy of multicolinearty problem and it can be employed very often in the literature.

However finding k is another problem while implying ridge regression. The existing methods for finding k value in the literature are based on either reducing *VIF* values or minimizing the residual sum of squares. In this study the proposed algorithm for finding k value is based on both reducing *VIF* values and minimizing the residual sum of squares at a time. Since the objective function introduced in the paper is a piecewise function, classical optimization techniques are not suitable.

Therefore the particle swarm optimization has been used for solving optimization problem in this study. It can be actually possible to use other artificial intelligence optimization techniques. Furthermore, we might think about finding different k values for each explanatory variable in future works.

REFERENCES

[1] Montgomery, D., and Peck, E.A., Introduction to linear regression Analysis, John Wiley &Sons New York, 1982.

[2] Hoerl, A.E., and Kennard, R.W., 1970a, Ridge regression: biased estimation for non-orthogonal problems, Technometrics, 12, 55-67.

[3] Hoerl, A.E., Kennard, R.W., and Baldwin, K.F., 1975, Ridge regression: some simulation, Communication in Statistics 4, 105-123.

[4] Mc Donald, G.C., and Galarneau, D.I., 1975, A Monte Carlo Evaluation of sum ridge-type estimators, Journal of the American Statistical Association, 70, 407-412.

[5] Lawless, J.F., and Wang, P., 1976, A simulation study of ridge and other regression estimators, Communication and Statistics, A5, 307-323.

[6] Hocking, R.R, 1976, The analysis and selection of variables in linear regression, Biometrics, 32, 1-49.

[7] Wichern, D., and Curchill, G., 1978, A comparison of ridge estimators, Technometrics, 20, 301-311.

[8] Nordberg, R., 1982, A procedure for determination of a good ridge parameter in linear regression, Communications in Statistics, A11, 285-309.

[9] Prago-Alejo, R.J., Torre-Trevino, L.M., and Pina-Monarrez M.R., Optimal determination of k constant of ridge regression using a simple genetic algorithm, Electronics robotics and Automotive Mechanics Conference, 2008.

[10] Al-Hassan, Y.M., 2010, Performance a new ridge regression estimator, Journal of the Association of Arab Universities for Basic and Applied Sciences, 9, 23-26.

[11] Ahn, J.J., Byun, H.W., Oh, K.J., and Kim, T.Y., 2012, Using ridge regression with genetic algorithm to enhance real estate appraisal forecasting, Expert Systems with Applications, 39, 8369–8379.

[12] Siray, G.U., Kacıranlar, S., and Sakallıoglu, S., 2012, $r-k$ Class estimator in the linear regression model with correlated errors, Statistical Papers (DOI 10.1007/s00362-012-0484-8)

[13] Sakallıoglu, S., and Kacıranlar, S., 2008, A new biased estimator based on ridge estimation, Statistical Papers, 49, 669-689.

[14] Firinguetti, L., and Bobadilla, G., 2011, Asymptotic confidence intervals in ridge regression based on the Edgeworth expansion, Statistical Papers, 52, 287-307.

[15] Tabakan, G., and Akdeniz, F., 2010, Difference-based ridge estimator of parameters in partial linear model, Statistical Papers, 51, 357-368.

[16] Duran, E.A., and Akdeniz, F., 2012, Efficiency of the modified jackknifed Liu-type estimator, Statistical Papers, 53, 265-280.

[17] Uemukai, R., 2011, Small sample properties of a ridge regression estimator when there exist omitted variables, Statistical Papers, 52, 953–969.

[18] Huang, J.C., 1999, Improving the estimation precision for a selected parameter in multiple regression analysis, Economic Letters, 62, 261–264.

[19] Akdeniz, F., 2004, New biased estimators under the LINEX loss function, Statistical Papers, 45, 175-190.

[20] Alkhamisi, M.A., 2010, Simulation study of new estimators combining the SUR ridge regression and the restricted least squares methodologies, Statistical Papers, 51, 651–672.

[21] Hoerl, A.E., and Kennard, R.W., 1970b, Ridge regression: applications to non-orthogonal problems, Technometrics, 12, 69-82.

[22] Marquardt, D.W., and Snee, R.D., 1975, Ridge regression in practice, The American Statisticians, 29, 3-20.

[23] Hoerl, A.E., and Kennard, R.W., 1976, Ridge regression: iterative estimation of the biasing parameter, Communication in Statistics, Part A5, 77-88.

[24] Kennedy, J., and Eberhart, R.C., In: Particle Swarm Optimization, IEEE International Conference on Neural Network, 1942-1948, 1995.

[25] Samprit, C., and Hadi, A.S., Regression Analysis by Example, John Wiley & Sons, Inc, 2006.