İstatistikçiler Dergisi: İstatistik & Aktüerya

Journal of Statisticians: Statistics and Actuarial Sciences

IDIA 9, 2016, 2, 47-53

Geliş/Received:24.03.2016, Kabul/Accepted: 09.11.2016

www.istatistikciler.org

Araştırma Makalesi / Research Article

A Comparative Study on Regression Methods in the presence of Multicollinearity

Onur Toka
Hacettepe University, Department of Statistics,
06800, Beytepe, Çankaya, Ankara, Turkey
onur.toka@hacettepe.edu.tr

Abstract

This study aims to compare some regression methods in the presence of multicollinearity problem. Multicollinearity problem causes faults such as insignificant variable selection, biased variances in linear regression model. Therefore, some regression methods that handle with multicollinearity such as partial least square regression (PLSR), ridge regression (RR) and principal component regression (PCR) had reported. In this paper, the methods were compared by the simulation study. All results were compared with each other through MSE of their estimated beta values for different methods. The results show that PLRS is better methods with large numbers of independent variable. Further, RR is better method when observation number and number of multicollinearity is large enough. PCR cannot be better method in simulation study scenarios.
Keywords: Ridge Regression, Partial Least Square Regression, Multicollinearity, Principal Component Regression

Öz

Çoklu Bağlantı Durumunda Regresyon Yöntemlerinin Karşılaştırılması

Bu çalışmada çoklu bağlantı sorununa çözüm getirebilmek için birkaç regresyon modelinin karşılaştırılması amaçlanmıştır. Doğrusal regresyonda çoklu bağlantının olması durumunda yanlış kestirim değerleri, kestirim için yanlı varyanslar gibi yanlışlıklar ortaya çıkar. Bu yüzden, kısmi en küçük kareler regresyonu (PLSR), ridge regresyon (RR), temel bileşenler regresyonu (PCR) gibi yöntemler, çoklu bağlantı sorununu aşmak için önerilmiştir. Bu çalışmada yöntemler, farklı derecelerde çoklu bağlantıya sahip veri kümeleri için bir benzetim çalışmasıyla karşılaştırılmıştır. Regresyon model kestirimleri için bütün sonuçlar hata kareler ortalaması bakımından birbirleriyle karşılaştırılmıştır. PLSR yönteminin bağımsız değişken sayısının çok olduğu durumda iyi bir yöntem olduğu ve RR yönteminin ise gözlem sayısı ve çoklu bağlantı sayısının çok olduğu durumda iyi bir yöntem olduğu sonuçları elde edilmiştir. PCR yönteminin benzetim çalışması senaryolarında tercih edilebilir bir yöntem olmadığı görülmüştür.
Anahtar Sözcükler: *Ridge regresyon, Kısmi en küçük kareler regresyonu, Çoklu bağlantı, Temel bileşenler regresyonu.*

1. Introduction

Linear regression is a famous analysis to get relationship between dependent variable and independent variables in a simplified mathematical form:

$$y = X\beta + \varepsilon \quad (1)$$

y is a $nx1$ vector of observations of dependent variable, $X = [1_n, x_{i1}, x_{i2}, \dots, x_{ik}]$ is matrix that consists of n observations on p columns for k variables and unknown constant. $\beta = [\beta_0, \beta_1, \dots, \beta_k]$ is $px1$ vector for coefficients and ε is a vector of identically and independent distributed (iid) errors. Minimizing the sum of square errors is a way to get Ordinal Least Squares (OLS) estimator:

$$\hat{\beta}_{OLS} = (X^T X)^{-1}(X^T Y) \quad (2)$$

$\hat{\beta}_{OLS}$ is unbiased estimator of β. OLS estimator runs with some assumptions such as linearity, iid errors with zero mean and constant variance, homoscedasticity, no multicollinearity and no auto-correlation. If the assumptions are not provided, model cannot be enough better. Therefore, alternative methods has been proposed to handle with assumption distortions.

In regression analysis, when two or more independent variables are correlated with each other's, it is called multicollinearity problem. Multicollinearity problem increases variance of coefficient estimates, causes wrong sign coefficients and makes to specify the correct model more difficult. Therefore, multicollinearity problem has to be revealed in the process of regression modelling. Specifying best model is the important task to get better prediction. Fortunately, there are lots of proposed methods such as RR, PLSE, PCR to overcome multicollinearity problem. The most important issue is which method will be the best answer for application data sets. There are some studies to compare similar methods with real dataset (see [5-6-11-18]). Accordingly, simulated data sets are generated for different scenarios. It is clarified that which model is better in created scenarios. The rest of the study is organized as follows. The regression methods are given in Section 2. Simulation results are stated in Section 3 and finally, discussions are explained in Section 4.

2. Regression Methods for Multicollinearity Problem

Shrinkage methods and latent variables can combat to multicollinearity in linear regression modelling. There are lots of proposed methods to shrinkage or select subset of independent variables (see [9-15-16]) and there are also some methods to combine independent variables (see [1-3-12-13-17]) to eliminate multicollinearity in linear regression model. In this paper, OLS, RR, PCR and PLSR methods are included.

2.1. Ridge Regression

OLS estimator for regression parameter(β) is imposed large variance in the presence of multicollinearity problem. The problem often causes unstable point estimate and excessive wide confidence intervals. One of the preferable way is biased estimators. Hoerl and Kennard [7] proposed a biased but more stable estimator for multicollinearity problem:

$$\hat{\beta}_k = \left(X^T X + kI_p\right)^{-1} X^T y, \;\; k \geq 0. \quad (3)$$

The estimator is similar OLS. However, the addition of a positive number k to the diagonal elements copes with non-singular problem in $X^T X$ matrix. Determining k value is an important part of ridge regression. The goal is to find some k which is large enough to reduce the variance compared to the OLS estimator, but which is small enough to produce some acceptable low bias [16]. There are several ways to choice k value. Ridge trace is a subjective method by deciding with scatter plot of k versus ridge estimators of $\hat{\beta}_k$'s. Also, some objective selection methods were proposed such as Heoerl and Kennard [7], Theobald [14], Hoerl, Kennard and Baldwin [8], Lawless and Wang [10]. In this study, Lawless and Wang [10] k selection method was used.

2.2. Principal Component Regression

PCR is actually a linear regression method. However, the dependent variable is regressed on the principal components instead of dependent variables to cope with collinearity. The number of principal components are identified by obtaining maximum of variation of X. Assume that λ_i's are the eigenvalues of correlation matrix $X^T X$ and γ_i's are the unit-norm eigenvectors of $X^T X$.

$$X^T X\gamma_i = \lambda_i\gamma_i,\ i = 1,2,\dots,k \tag{4}$$

Then, the vector γ_i is used to express the X's in terms of PC Z's in the form,

$$Z_i = \gamma_{1i}X_1 + \gamma_{2i}X_2 + \cdots + \gamma_{ki}X_k. \tag{5}$$

All of Z_i's are orthogonal each other and PCR estimator is found as

$$\hat{\beta}_{PCR} = V_m\alpha_m \tag{6}$$

where $\alpha_m = (Z_m^T Z_m)^{-1}Z_m^T y$, m is the number of PCs retained in the model, V_m is a matrix consisting of the first m unit-norm eigenvectors [18].

2.3. Partial Least Squares Regression

In the linear regression model, it is really important to get best model to predict future observations. However, multicollinearity problem can prevent to predict well. The goal of PLSR extracts latent variables which are linear combinations of the independent variables. The highly relationship between independent variable are not occur after constructing latent variables [2-4].

PLSR finds components from X that are also relevant to Y. PLSR searches for a set of components, latent vectors, that performs a simultaneous decomposition X and Y with the constraint that these components explain as much as possible of the covariance between matrix X and Y [1].

There are some similarities with the PCR. In both methods, some attempts have been made to find some factors that will be regressed with the Y variables. The major difference is, while PCR uses only the variation of X to construct new factors, PLS uses both the variation of X and Y to obtain new factors that will play the role of explanatory variables [18].

For k independent variables X and dependent variable y, there is a sample size n of a $(1+k)$ dimensional vector $z = (y, X)^T$. Let S_Z be the sample covariance [5]:

$$S_Z = \begin{pmatrix} S_y^2 & S_{y,X}^T \\ S_{y,X} & S_X^2 \end{pmatrix} \tag{7}$$

X is nxk data matrix of independent variables and x_i^T is ith row of X matrix. Then the following equation holds when P is kxm matrix of the loadings of the vector $t_i = (t_{i1}, t_{i2}, \dots, t_{im})^T$ and q is a dimensional vector of the y loadings:

$$x_i = Pt_i + \epsilon_i \tag{8}$$

$$y_i = q^T t_i + \delta_i \tag{9}$$

ϵ_i and δ_i have zero mean and uncorrelated. $T = (t_1, t_2, \dots, t_m)^T$ score matrix should be estimated as $T = XW_m$ where $W_m = [w_1, w_2, \dots, w_m]$ is the loading matrix and the vectors are the solution of

$w_i = \underset{w}{\arg max}\, cov^2(Xw, y)$ with the constraint that $w^T w = 1$ and $w^T S_X w_j = 0$ for $1 \leq j < i$ with $w_1 \propto S_{y,x}$. Consequently, it is concluded that factors $(t_1, t_2, \dots, t_m)$ are orthogonal. The vectors w_i are found as the eigenvectors linked to the largest eigenvalues of the matrix $\big(I - P_x(i)\big)S_{y,x}S_{y,x}^T$ where $P_x(i) = (S_X W_i)[(S_X W_i)^T (S_X W_i)]^{-1}(S_X W_i)^T$. From the results the vectors w_i can be computed recursively as

$$w_1 \propto S_{y,x} \tag{10}$$

$$w_{i+1} \propto S_{y,x} - S_x W_i (W_i^T S_x W_i)^{-1} W_i^T S_{y,x}\,, \quad 1 \leq i < m. \tag{11}$$

PLS regression coefficient estimators are found as:

$$\hat{\beta}_{PLSR} = W_m (W_m^T S_x W_m)^{-1} W_m^T S_{y,x} \tag{12}$$

The algorithm has two subsection. Firstly, the weights w_i are computed by using covariance matrix of observations and the regression coefficients q_i are computed by OLS with dependent y and independent latent variables t_i[5].

3. Simulation Study

In the simulation study, error vector for regression model is generated with four variables using multinomial normal distribution $MN(\underline{0}, I)$. The independent data matrix is generated with multinomial normal distribution $MN(\underline{0}, S)$ where S covariance matrix is accounted into with $Cor(X_t, X_{t-1}) = 0.5$. It is created 100 and 1000 observations respectively to compare the methods in terms of the observation numbers. To observe the effect of the number of multicollinearity, it was created one and two multicollinearity equations for both data sets. The number of independent variable is another impact factor on the methods. Therefore, 4 and 10 independent variables are used for regression equations respectively. Also, betas are formed by different Poisson distribution or subjective fixed constant values to observe the impact of randomness. Lawless and Wang [10] k selection method is selected to get shrinkage proportion for ridge regression. Number of PC is selected as number of unrelated covariates for all simulation cases. MSE scores for the regression methods are calculated after 1000 repeats. R (version: 3.0.2) program was used for simulation. All of simulation scenarios are given in Table 1.

Table 1. Simulation Scenarios

Errors	$MN(\underline{0}, I)$.
Independent Variables	$MN(\underline{0}, S)$ with $Cor(X_t, X_{t-1}) = 0.5$ for covariance matrix S
Multicollinearity	$x_t = x_v + rnorm(n) * .0001\,,\ t \neq v$
The number of Observations	100; 1000
Beta Types	Randomly Poisson Distribution(r), Fixed (f).
The number of multicollinearity	# of MC: 1,2
The number of independent variables	k: 4,10
Repeats	1000

Mean square errors (MSE) of betas are achieved and compared for all of methods. The graphics are only for PLSR, PCR and RR because of avoiding OLS's vision incomparable results. On the other hand, there are also simulation results table including OLS, too. MSE scores of OLS, PLRS, PCR and RR for 100 observation are given in Table 2. As it is expected, OLS results are quite worse than the biased estimators. Multicollinearity causes misleading for betas in OLS estimators so MSE score is influenced. When independent variable number is 4 and observation number is 100, then MSE scores for all biased estimators are similar. However, independent variable number is 10 and observation number is 100, then PLS is better than PCR and RR. The number of multicollinearity and betas' type are ineffective on MSE to specify the best methods. On the other hand, MSE scores are smaller when betas are fixed and the number of multicollinearity is one under the condition that observation number and the number of independent variables are fixed. It is clarified that when the number of independent variables increases, PLRS has appreciable smaller MSE scores. Further, PLRS has also smaller MSE scores for the other situations in Table 2, too. RR is the second best and PCR fails for the scenarios in Table 2.

Table 2. MSE of betas for OLS, PLRS, PCR and RR

n	k	# of MC	Beta	OLS	PLSR	PCR	RR
100	4	1	r	2102193.00	5.5471	5.9234	5.9164
100	4	1	f	10253.45	1.0323	1.4965	1.3223
100	4	2	r	4373768.00	36.2488	36.4239	36.4034
100	4	2	f	4407395.00	20.0141	20.1796	20.1749
100	10	1	r	73652.51	10.2087	30.3663	30.3584
100	10	1	f	66115.11	2.8195	17.4321	17.3969
100	10	2	r	135653.80	25.9577	48.4172	47.3738
100	10	2	f	131943.70	13.4428	29.7955	29.7873

MSE scores of OLS, PLRS, PCR and RR for 1000 observation are given in Table 3. OLS results are worse than the biased estimators. Multicollinearity causes misleading for betas in OLS estimators so MSE score is influenced. When independent variable number is 4, then MSE scores for all biased estimators are similar to Table 2. However, when the number of multicollinearity increases, RR has smaller results than the other methods. The number of multicollinearity alters the results for large sample size and small number of independent variables. The other scenarios are similar to small observation number. It is clarified that when the number of independent variables is large, PLRS has appreciable smaller MSE scores, same as smaller sample size. Betas' type are ineffective on MSE to compare the methods. On the other hand, MSE scores are smaller when betas are fixed under the condition that the other parameters are fixed.

Table 3. MSE of betas for OLS, PLRS, PCR and RR

n	k	# of MC	Beta	OLS	PLSR	PCR	RR
1000	4	1	r	197917.70	5.1747	5.2079	5.2234
1000	4	1	f	2047.177	2.0038	2.0282	2.5981
1000	4	2	r	394441.50	36.6944	36.7145	36.6639
1000	4	2	f	391368.10	20.0013	20.0193	19.9972
1000	10	1	r	5832.86	9.8090	28.8310	28.6776
1000	10	1	f	5778.62	2.6800	16.8740	16.8653
1000	10	2	r	12439.15	26.5912	48.6602	48.5178
1000	10	2	f	11768.49	13.3435	29.4203	29.3265

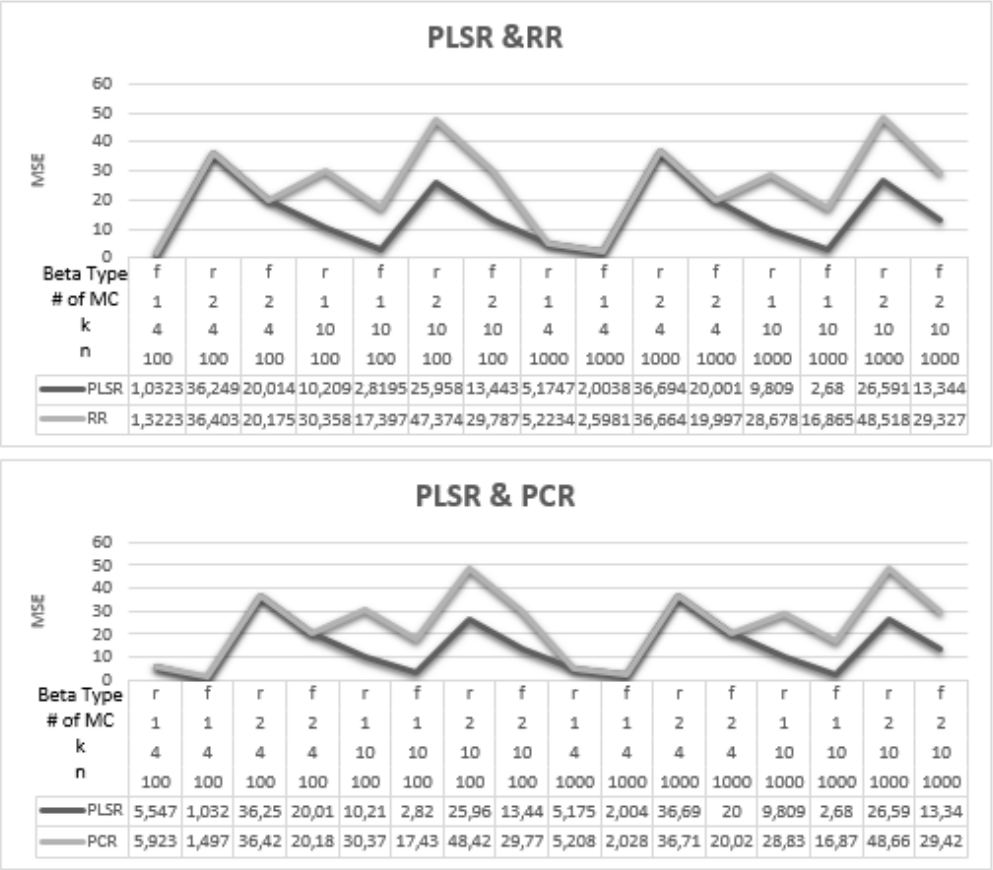

Figure 1. Comparison of MSE: PLSR&RR – PLSR&PCR

Figure 1 show that PLSR has significantly smaller MSE than the other methods when the number of independent variable is large.

4. Conclusion

PLRS, PCR and RR are useful program for multicollinearity problem. There are studies to compare these methods by using real data sets. Especially, PLRS method is used in many fields of science such as chemo metrics, social sciences and marketing. RR is a well-known statistical methods in the presence of multicollinearity and PCR is also popular methods in multivariate statistics field. It is really important that which model gives the best estimation for unknown real parameters. Therefore, simulated data sets are constructed to investigate which regression method is useful concerned data structure.

To summarize all alternative scenarios for the methods, while number of independent variable increases, PLSR is better than PCR and RR. If the observation number and the number of multicollinearity are large enough and the number of independent variable is small, RR is the smallest MSE. PCR fails for all scenarios. This paper also had similar results El-Fallah and El-Salam's [3] simulation but also this study investigated impact of the number of multicollinearity in linear regression model. For the future studies, same methods and robust alternatives can be explored in the presence of outliers.

References

[1] H. Abdi, 2003, Partial least squares (PLS) regression. – In: Lewis-Beck M. *et al.* (eds), *Encyclopedia of social sciences research methods*, Sage, 792–795.

[2] E. Bulut and A. Alın, 2009, Kısmi En Küçük Kareler Regresyon Yöntemini Algoritmalarından Nipals Ve Pls-Kernel Algoritmalarının Karşılaştırılması Ve Bir Uygulama, *Dokuz Eylül Üniversitesi Iktisadi Ve Idari Bilimler Fakültesi Dergisi*, 24, 2, p. 127-138.

[3] M. El-Fallah and A. El-Salam, 2014, A Note on Partial Least Squares Regression for Multicollinearity (A Comparative Study), *International Journal of Applied Science and Technology*, Vol. 4 No. 1; January 2014, 163-171.

[4] P. Geladi and B. Kowalski, 1986, Partial Least-Squares Regression: A Tutorial, *Analytica Chimica Acta*, 185, 1–17.

[5] J. Gonzalez , D. Pena, R. Romera, 2009, A robust partial least squares regression method with applications, *J. Chemometr.*, 23, pp. 78–90.

[6] I. S. Helland, 1990, Partial Least Squares Regression and Statistical Models, *Scandinavian Journal of Statistics*, 17(2), 97–114.

[7] A. E. Hoerl and R. W. Kennard, 1970, Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Technometrics*, 12, 1: 55-67.

[8] A. E. Hoerl and R. W. Kennard and K. F. Baldwin, 1975. Ridge Regression: Some Simulation. *Communication in Statistics*, 4(2): 105-123.

[9] L. Kejian, 2004, More on Liu-Type Estimator in Linear Regression, *Communications in Statistics - Theory and Methods*, 33:11, 2723-2733.

[10] J. F. Lawless and P. Wang, 1976, A Simulation Study of Ridge and Other Regression Estimators, C*ommunications in Statistics - Theory and Methods*, A5 (4), 307-323.

[11] S. Maitra and J. Yan, 2008, Principal Component Analysis and Partial Least Squares: Two Dimension Reduction Techniques for Regression, *2008 Casualty Actuarial Society Discussion Paper Program*, Presented June 15-18, 2008 Fairmont Le Château Frontenac Québec City, Québec, Canada.

[12] W. F. Massy, 1965, Principal Component Regression in Exploratory Statistical Research, *Journal of the American Statistical Association*, 60, 234-256.

[13] M. Stone and R. J. Brooks, 1990, Continuum Regression: Cross-Validated Sequentially Constructed Prediction Embracing Ordinary Least Squares, Partial Least Squares and Principal Components Regression. *Journal of the Royal Statistical Society*. Series B (methodological), 52(2), 237–269.

[14] C. M. Theobald, 1974. Generalizations Of Mean Square Error Applied To Ridge Regression, *Journal Of The Royal Statistical Society*, Series B, 36 : 103-105.

[15] R. Tibshirani, 1996, Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society*, Series B (methodological), 58(1), 267-288.

[16] S. Toker, S. Kaçiranlar, 2013, On the Performance of Two Parameter Ridge Estimator under the Mean Square Error Criterion, *Applıed Mathematıcs And Computatıon*, vol.219, 4718-4728.

[17] S. Wold, A. Ruhe, H. Wold, and W. J. Dunn, III, 1984, The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses, *SIAM Journal on Scientific and Statistical Computing*, 1984, Vol. 5, No. 3: 735-743.

[18] O. Yeniay and A. Goktas, 2002, A comparison of partial least squares regression with other prediction methods, *Hacettepe Journal of Mathematics and Statistics*, Vol. 31, 99-111.