JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION
2020, VOL. 90, NO. 16, 3009–3024
https://doi.org/10.1080/00949655.2020.1793342

Taylor & Francis
Taylor & Francis Group

Comparison of partial least squares with other prediction methods via generated data

Atila Göktaş and Özge Akkuş

Department of Statistics, Muğla Sıtkı Koçman University, Muğla, Turkey

ABSTRACT

The purpose of this study is to compare the Partial Least Squares (PLS), Ridge Regression (RR) and Principal Components Regression (PCR) methods, used to fit regressors with severe multicollinearity against a dependent variable. To realize this, a great number of varying groups of datasets are generated from standard normal distribution allowing for the inclusion of different degrees of collinearities for 10000 replications. The design of the study is based on a simulation work that has been performed for six different degrees of multicollinearity levels and sample sizes. From the generated data, a comparison is made using the value of mean squares error of the regression parameters. The findings show that each prediction method is affected by the sample size, number of regressors or multicollinearity level. However, in contrast to literature (say $n <= 200$), whatever the number of regressors is, PCR had significantly better results compared to the other two.

ARTICLE HISTORY
Received 3 February 2020
Accepted 5 July 2020

KEYWORDS
Partial least squares; ridge regression; principal components regression; multicollinearity

1. Introduction

A significant relationship among the explanatory variables in a linear regression model is called multicollinearity. When multicollinearity exists in a linear regression model, using t test statistics for testing the coefficients of the independent variables becomes problematic. To overcome the problem there are a great number of prediction methods that can be used to appropriately fit the respective linear regression model.

Regression analysis is the most common statistical method used to estimate the quantitative relationship between a dependent variable (Y) and one or more explanatory variables ($\boldsymbol{X}$). The most common method used for model estimation is Ordinary Least Square (OLS) method, provided that it satisfies some assumptions required in regression analysis. This method is based on the idea of minimizing the sum of the squares of the differences between the Y values and the predicted values obtained from the measurements. The reliability of the model obtained depends on satisfaction of the OLS assumptions. The existence of multicollinearity between the examined explanatory variables may lead to misinterpretation of the coefficients belonging to the regression parameters estimated using OLS. Estimation of results obtained in fields such as agriculture, socioeconomics,

CONTACT Atila Göktaş gatilla@mu.edu.tr Department of Statistics, Muğla Sıtkı Koçman University, Muğla 48000, Turkey

medicine and biology via OLS method, without considering the required assumptions, may be misleading.

Ridge Regression (RR), Principle Component Regression (PCR), and Partial Least Square (PLS) regression are biased prediction methods developed to eliminate the negative effect that will occur on parameter estimates in case of multicollinearity.

The main purpose of the study is to compare the three biased prediction methods (PLS, PCR and RR) used in the presence of multicollinearity between explanatory variables included in a linear regression model. Since this comparison could not be made analytically, a simulation study was conducted and the obtained results were interpreted.

2. Literature review

The studies done in this field are chronologically listed below.

Hoerl and Kennard [1] suggested in their extended study that a separate k value could be selected for each regression. However, they also stated that there is no guarantee that this will give better results than the k trace in any case. Hoerl and Kennard [2] stated that there is no single value of k that is the ridge parameter estimator and that the results would be better than OLS if the optimal k could be determined. They suggested the ridge trace for the selection of k. Hoerl et al. [3] suggested an algorithm for selecting the parameter k with superior features than OLS. Kidwell and Brown [4] used generated data for applying RR method in their study on multicollinearity. Based on the obtained results, it was observed that the RR model yielded different outcomes compared to the OLS method when predictors were not orthogonal. In a study performed by Yeniay and Göktaş [5], OLS, PLS, RR and PCR methods were compared using a real and a single data set. The PLS method was found to yield better results for the data than the other methods used. Kibria [6] proposed a few new ridge parameter estimators based on a generalized ridge regression approach. Graham [7] noted that in multiple-regression analysis applied to ecological data, a multicollinearity problem was encountered. Graham showed the use of different statistical techniques in response to the multicollinearity problem using real ecological data. As a result, he stated that ecological data could enhance the explanation of multicollinearities in multiple regression models and the reliability of the model through the use of different methods.

In a study conducted by Albayrak [8], researchers investigated whether the RR and PCR methods were more effective than the OLS method in predicting body weight. The RR, PCR and OLS methods were applied to explanatory variables having multicollinearity. The RR and PCR methods were observed to yield more stable predictions with lower standard error. Karadavut et al. [9] compared the parameter estimates obtained for some characters affecting the yield of the chick pea plant using M-regression methods, one of the RR, OLS and Robust Regression methods. In the regression model that impacts the variables affecting the grain weight of the chickpea plant, the parameters were first estimated using the OLS method and, by detecting the multicollinearity between the exploratory variables, parameter estimates were also obtained via the RR method. Mansson et al. [10] conducted a simulation study to compare the performance of some ridge estimators based on both MSE values and Prediction Sum of Square (PRESS) values. Topal et al. [11] aimed to develop a model that estimates carcass weights using various body measurements of 91 carp at different ages. They used OLS regression, RR and PCR methods to eliminate the multicollinearity problem emerging between the obtained body measurements. According to

the results, it was stated that it is more accurate to use the RR and PCR methods in place of OLS to eliminate the multicollinearity problem. Li [12] conducted the first simulation work to compare the bias regression methods with few and unacceptable replication (100) for different scenarios. She was unwilling to present a definite priority in favour of any of the three bias regression methods. Rathert et al. [13] compared the OLS and PCR methods for estimating egg internal quality characteristics in Japanese quail. As a result, it was stated that the multicollinearity problem had been eliminated via the PCR model. Acharjee et al. [14] applied OLS regression and some other regression methods such as RR, PCR, and PLS regression to the omic data set and analytically compared the results on a single data set. Mahesh et al. [15] compared the estimates of protein content and hardness values of Canadian wheat obtained at different sites and sowing times. OLS and PCR methods were used for these comparisons. In the end, it was reached to a conclusion that the OLS model performed better than the PCR model in estimating protein content and wheat hardness. In an empirical study conducted by Simeon et al. [16], the classic OLS regression, RR and PCR results were compared over gynecological data with multicollinearity. As a result, RR was found to be the most appropriate statistical method. In an empirical study conducted by Goharnejad, Zarei and Tahmasebi [17], multiple regression method, PCR, PLS regression and RR results were compared over pasture biomass data. The results showed that the best estimates were given by the PLS regression and RR. Polat and Günay [18] conducted a study where PLS regression, PCR, RR and OLS regression results were compared over a real data set related to air pollution. Brito, Almeida and Matos [19] performed an application on water flow rate in urban areas by applying the PLS method. Firinguetti, Kibria and Araya [20] studied the OLS, PLS and RR methods used in case of a multicollinearity problem. Results from the simulation study comparing the performances demonstrated that RR performed better in cases where the error variance was large and that the PLS method achieved the best results when the model included more variables. Kibria and Banik [21] have performed a simulation study to present the power and nominal size of the test used for the linear regression model for RR method only.

When these studies are examined, it becomes clear that comparisons have generally been made on a single data set. As it is not possible to generalize results on the basis of a single data set, the current study conducts a simulation under specific scenarios to present more general conclusions. In the literature, the work done by Firinguetti, Kibria and Araya [20] in this area is also a simulation study. However, the current study differs from theirs in that the PCR method is also included and a more sophisticated simulation design has been used. The findings obtained from our simulation study have thus been compared with the results obtained by Firinguetti, Kibria and Araya [20].

3. Methodology

Multicollinearity is generally observed due to the strong correlation between explanatory variables in regression models. The presence of multicollinearity increases the variance of parameter estimates. While models with particularly small and medium sample sizes are found to be strongly significant, the explanatory variables are individually less significant. Multicollinearity can also lead to imprecise results about the relationship between dependent and explanatory variables. Because of all these negativities, biased predictors used in the case of multicollinearity are introduced below.

3.1. Partial least square method (PLS)

PLS method is a method developed by Wold [22], which is especially useful when the number of explanatory variables is large and the number of samples is small. PLS method aims to find the number of components that maximizes the covariance between the dependent and explanatory variables in a data set. In this method, centralization and rescaling are performed to ensure that all variables are at the same measurement and make comparisons possible before conducting analyses.

In order to rotate the Y-dependent variable with $X_1, \cdots, X_p$ explanatory variables via the PLS method, new components are obtained which have a similar role to the X-explanatory variables and are generally identified as latent variables. The latent variables obtained by reducing the size of the explanatory variables are linear combinations of the explanatory variables. PLS pretends similar to PCR in terms of working with less factors in place of all explanatory variables. In the PLS and PCR methods, the dimensionality of the regression problem is reduced by using fewer components than the number of X variables. Each component is a linear combination of $X_1, \cdots, X_p$. The main difference between PLS and PCR (given in the Sub-section 3.3) is that while determining the basic components in PLS the dependent variable has an important role; whereas the basic components in the PCR method do not use the dependent variable as a reference.

The PLS method gives linear decompositions of X by following a path similar to Principal Component Analysis (PCA). In Equation (3.1), t_j's are the linear combination of X. $p \times 1$ dimensional p_j's are defined as loads. Algorithms called NIPALS and SIMPLS are used to find PLS estimators. In the conventional NIPALS algorithm, the t_j's given by Equation (3.2) are obtained as linear combinations of E_j residual matrices.

$$X = t_1 p_1^{'} + t_2 p_2^{'} + \cdots + t_p p_p^{'} = \sum_{j=1}^{p} t_j p_j^{'} = TP' \tag{3.1}$$

$$t_j = E_{j-1} w_j, \quad E_j = X - \sum_{i=1}^{j} t_i p_i^{'} \, E_0 = X \tag{3.2}$$

w_j and r_j $(j = 1, 2, \cdots, \mathrm{h})$ weight clusters are in the same space. w_j's here are orthonormal. For univariate and multivariate PLS, it is necessary to first obtain w_j or r_j $(j = 1, 2, \cdots, \mathrm{h})$ in order to calculate the linear combination of t_j in many algorithms. Then by rotating $\boldsymbol{X}$ matr t_j, p_j is found. The reduced dimension h is obtained and the following equations can be written [5, 23].

$$T_h = XR_h \tag{3.3}$$

$$P_h = X' T_h (T_h^{'} T_h)^{-1} \tag{3.4}$$

$$R_h = W_h (P_h^{'} W_h)^{-1} \tag{3.5}$$

The subset h here consists of the first h sequences of the corresponding vectors of the matrix. Equation (3.5) shows that a set of two weight vectors is combined to a linear

transformation. Equations (3.3) and (3.4) show that $P_h'R_h = I_h$ and $R_h'P_h$ is equal to I_h.

$$R_h^{'}P_h = R_h^{'}X'T_h(T_h^{'}T_h)^{-1} = T_h^{'}T_h(T_h^{'}T_h)^{-1} = I_h \tag{3.6}$$

After determining the dimension h, the vector of fitting values obtained with PLS can be shown as in Equation (3.7).

$$\hat{y}_{PLS}^{h} = T_h(T_h^{'}T_h)^{-1}T_h^{'}y \tag{3.7}$$

When XR_h is substituted in place of T_h in Equation (3.7) and $X\hat{\beta}_{OLS}$ is substituted in place of y,

$$\hat{y}_{PLS}^{h} = XR_h(R_h^{'}X'XR_h)^{-1}R_h^{'}X'X\hat{\beta}_{OLS} \tag{3.8}$$

is obtained. On the basis of Equation (3.8), it is expressed as

$$\hat{\beta}_{PLS}^{h} = R_h(R_h^{'}X'XR_h)^{-1}R_h^{'}X'X\hat{\beta}_{OLS} \tag{3.9}$$

In the multivariate case, instead of $\hat{y}_{PLS}^{h}$ vector $\hat{Y}_{PLS}^{h}$ matrix is used. By means of a non-singular rotation, R_h in Equations (3.8) and (3.9) can be changed without changing the outcome. W_h can be used in place of R_h. In order to reach a more simplified form of $\hat{\beta}_{PLS}^{h}$, when the corresponding term in Equation (3.4) is written in place of T_h in Equation (3.5);

$$P_h = X'XR_h(R_h^{'}X'XR_h)^{-1} \tag{3.10}$$

is obtained. From here, with the inequality in Equation (3.9);

$$\hat{\beta}_{PLS}^{h} = R_hP_h^{'}\hat{\beta}_{OLS} = W_h(P_h^{'}W_h)^{-1}P_h^{'}\,\hat{\beta}_{OLS} \tag{3.11}$$

is obtained. For the multivariate case;

$$\hat{\beta}_{PLS}^{h} = R_hP_h^{'}\hat{B}_{OLS} = W_h(P_h^{'}W_h)^{-1}P_h^{'}\,\hat{B}_{OLS} \tag{3.12}$$

is obtained where matrix $W_h(P_h'W_h)^{-1}$ is a projection matrix. However, since this matrix is not symmetrical, it is referred to as an oblique projector. The PLS regression predictor for the h component shown in Equation (3.11) is the oblique reflection of W_h vertical to P_h along P_h space on $\hat{\beta}_{OLS}$. The degree of bias in the PLS regression can be controlled with the dimension of space corresponding to vertical reflection of $\hat{\beta}_{OLS}$. The smaller the dimension is, the larger the bias is.

3.2. Ridge regression method (RR)

The RR method is a method developed by Hoerl and Kennard (1970) to eliminate the multicollinearity problem. The RR aims to reduce the degree of collinearity by adding positive and small k values to the diagonal elements of the $X'X$ correlation matrix. Estimates obtained using the RR method, are more reliable than the ones obtained with the OLS method.

The use of the RR method is recommended for the following situations;

(i) For finding estimates with a variance smaller than the OLS estimate in cases where the explanatory variables are related to each other in the multiple linear regression models,
(ii) In the case of strong multicollinearity, for the graphical representation of the instabilities in coefficients,
(iii) For reducing the Mean Square Error (MSE) value by changing the variance and the bias square in regression;
(iv) For eliminating the multicollinearity found in explanatory variables,

In the RR method, when calculating regression coefficient estimates, a small and positive constant has been added to the diagonal elements of the $X'X$ matrix. Thus, RR solution is,

$$\hat{\beta} = (X'X + kI)^{-1}X'Y \tag{3.13}$$

where I is the pxp dimensional unit matrix and $X'X$ is the correlation matrix of the explanatory variables. Much work has been done on the selection of k value. However, there is no work yet to provide a definite solution on the optimum k-value.

3.2.1. The relationship of the ridge estimator with the OLS estimator

The OLS estimator is determined to be,

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{3.14}$$

It is clear from Equation (3.14) that the following equivalency can be obtained,

$$(X'X)\hat{\beta} = X'Y \tag{3.15}$$

Remember that the ridge estimator was given as;

$$\hat{\beta}^* = (X'X + kI)^{-1}X'Y \tag{3.16}$$

When its equivalent given in Equation (3.15) is written in place of $X'Y$,

$$\hat{\beta}^* = (X'X + kI)^{-1}X'X\hat{\beta} \tag{3.17}$$

is obtained. As the inverse of the inverse of $(X'X)$ matrix is equal to itself, Equation (3.17) may be rewritten as follows,

$$\hat{\beta}^* = (X'X + kI)^{-1}[(X'X)^{-1}]^{-1}\hat{\beta} \tag{3.18}$$

As both of the matrices are not singular,

$$\hat{\beta}^* = [(X'X)^{-1}(X'X + kI)]^{-1}\hat{\beta} \tag{3.19}$$

From here,

$$\hat{\beta}^* = [(X'X)^{-1}(X'X) + k(X'X)^{-1}]^{-1}\hat{\beta} \tag{3.20}$$

is obtained. After the necessary operations,

$$\hat{\beta}^* = [I + k(X'X)^{-1}]^{-1}\hat{\beta} \tag{3.21}$$

is obtained.

If

$$Z = [I + k(X'X)^{-1}]^{-1} \tag{3.22}$$

Then

$$\hat{\beta}^* = Z\hat{\beta} \tag{3.23}$$

This equation shows that the ridge estimator is a transformation of the OLS estimator. At the same time, the length of $\hat{\beta}^*$ is shorter for k $\neq 0$ than $\hat{\beta}$ and shown as follows:

$$\hat{\beta}^{*'}\hat{\beta}^* < \hat{\beta}'\hat{\beta} \tag{3.24}$$

3.2.2. Characteristics of Z and W matrices

The Ridge estimator was given as $\hat{\beta}^* = (X'X + kI)^{-1}X'Y$. When W matrix is defined as,

$$W = (X'X + kI)^{-1} \tag{3.25}$$

$$\hat{\beta}^* = WX'Y \tag{3.26}$$

Matrix Z was defined as,

$$Z = [I + k(X'X)^{-1}]^{-1} \tag{3.27}$$

From here, eigenvalues of W matrices;

$$\tau_i(w) = \frac{1}{\lambda_i + k} \tag{3.28}$$

Eigenvalues of Z matrix:

$$\tau_i(z) = \frac{\lambda_i}{\lambda_i + k} \tag{3.29}$$

is at the same time given with the following equation,

$$Z = I - k(X'X + kI)^{-1} = I - kW \tag{3.30}$$

In our study, the ridge parameter estimate proposed by Hoerl and Kennard (1970) has been used as follows,

$$k = \frac{\hat{\sigma}^2_{OLS}}{max(\hat{\beta}_{OLS})} \tag{3.31}$$

3.2.3. Some characteristics of the ridge estimator

(i) **(i)** $\hat{\beta}(k)$ minimizes residual sum of squares on the origin-centered sphere whose radius is in the length of $\hat{\beta}(k)$.
(ii) Residual sum of squares is the increasing function of k.
(iii) **(iii)** $\hat{\beta}(k)'\hat{\beta}(k) < \hat{\beta}'\hat{\beta}$ and $k \to \infty$ for $\hat{\beta}(k)'\hat{\beta}(k) \to 0$.
(iv) **(iv)** $\lambda_1 \geq \cdots \geq \lambda_p$, $X'X$ being the eigenvalues, $X'X + kI$ is the ratio of the biggest eigenvalue to the smallest eigenvalue $(\lambda_1 + k)/(\lambda_n + I)$ and k is a decreasing function. The square root of this ratio is called X condition number [24].

(v) The Ridge estimator gives the OLS estimator for $k = 0$. Moreover, the ridge estimator can also be written as a linear transformation of the OLS estimator:

$$\hat{\beta}(k) = (X'X + kI)^{-1}X'y$$

Hoerl and Kennard (1970) have shown that the total variance of the ridge estimator is a continuous monotonically decreasing function of k and the square of the bias is a continuous monotonically increasing function of k. Therefore, the ridge estimator is said to be a good technique as long as the variance reduction is greater than the increase in the square of the bias.

3.3. Principal component regression (PCR)

Another biased predictor used to eliminate multicollinearity is PCR. It was first used by Hotelling [25]. PCR is applied to explanatory variables having high correlation between them. It is also considered a dimension reduction method reserving the highest amount of variance from the explanatory variables and assuring a lower number of uncorrelated explanatory variables. These new explanatory variables are called components. The score values of these components are used in the regression model established to explain the outcome variable. Factor analysis is sometimes performed instead of PCR for dimension reduction. The factor scores obtained as a result of factor analysis are used as explanatory variables in regression analysis.

In PCR, the data matrix constituted by n (the number of observations) and p (the number of variables) can be expressed as a population formed by a large number of points according to the state of X in p-dimensional space. If raw data is used in this matrix, the variance-covariance matrix is utilized. However, if standardized data is used, the correlation matrix is utilized. Which of these two ways yielding different results is used depends on the unit of measure of the data. If the units of measure are the same, the variance-covariance matrix should be used, otherwise, the correlation matrix is preferable.

For reference, PCR is a multivariate statistical method that explains the variance-covariance structure of a set of variables with linear combinations of these variables and allows dimension reduction and interpretation.

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X_{11}} & \boldsymbol{X_{12}} & \boldsymbol{X_{13}} & \cdots & \boldsymbol{X_{1j}} & \cdots & \boldsymbol{X_{1p}} \\ \boldsymbol{X_{21}} & \boldsymbol{X_{22}} & \boldsymbol{X_{23}} & \cdots & \boldsymbol{X_{2j}} & \cdots & \boldsymbol{X_{2p}} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \boldsymbol{X_{i1}} & \boldsymbol{X_{i2}} & \boldsymbol{X_{i3}} & \cdots & \boldsymbol{X_{ij}} & \cdots & \boldsymbol{X_{ip}} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \boldsymbol{X_{n1}} & \boldsymbol{X_{n2}} & \boldsymbol{X_{n3}} & \cdots & \boldsymbol{X_{nj}} & \cdots & \boldsymbol{X_{np}} \end{bmatrix}$$

p different number of random variables can be specified as follows;

$$X_1 = \begin{bmatrix} X_{11} \\ X_{21} \\ \vdots \\ X_{i1} \\ \vdots \\ X_{n1} \end{bmatrix}, X_2 = \begin{bmatrix} X_{12} \\ X_{22} \\ \vdots \\ X_{i2} \\ \vdots \\ X_{n2} \end{bmatrix}, \cdots, X_j = \begin{bmatrix} X_{1j} \\ X_{2j} \\ \vdots \\ X_{ij} \\ \vdots \\ X_{nj} \end{bmatrix}, \cdots, X_p = \begin{bmatrix} X_{1p} \\ X_{2p} \\ \vdots \\ X_{ip} \\ \vdots \\ X_{np} \end{bmatrix}$$

In a mathematical sense, the basic components are linear combinations of $X_1, X_2, \cdots, X_p$ variables. Geometrically, these linear combinations aim at the generation of new independent coordinate systems by rotating the original systems whose interrelated coordinate axes are $X_1, X_2, \cdots, X_p$. Newly obtained axes show directions with maximum variance, but at the same time allow the structure of change to be explained with simpler and fewer numbers of variables. The equation for the first principal component of the X observation matrix with highest information is:

$$Z_1 = t_{11}X_1 + t_{21}X_2 + \cdots + t_{p1}X_p = t_1^{'} \tag{3.32}$$

Here, it can be written as;

$$t_1^{'} = (t_{11}, t_{21}, \cdots, t_{p1}) \text{ and } X' = (X_1, X_2, \cdots, X_p) \tag{3.33}$$

Eigenvalues and eigenvectors of the variance-covariance matrix are used to find the linear components of p different variables presented in matrix X.

4. Simulation study

In this section, the efficiencies of the above-mentioned PLS, RR and PCR methods were investigated via a simulation study. With the Minitab 16.0 programme, a great number of varying groups of datasets are generated from standard normal distribution allowing for the inclusion of different degrees of collinearities for 10000 replications. The design of the study is based on simulation work that has been performed for six different degrees of multicollinearity levels (0.0, 0.3, 0.5, 0.7, 0.9, 0.99), three different number of variables (4, 7 and 9) and six different sample sizes (30, 50, 100, 200, 500 and 1000). The three proposed prediction regression methods are applied to the generated data. The Mean Square Error (MSE) value of the parameter estimates for each of these models is calculated and their means are computed. β^* and β in Equation (4.1) denote parameter estimation and the real parameter value, respectively.

$$MSE(\beta^*) = \frac{1}{10000}\sum_{i=1}^{10000}\sum_{j=1}^{p}(\beta^* - \beta)'(\beta^* - \beta) \tag{4}$$

Taking these MSE values into consideration, an attempt is made to determine which of the PLS, RR and PCR methods is more preferable under which condition. The mean MSE values obtained from the generated data are shown in Table 1; graphical representations of the results are given in App.1. Explanatory variables ($\boldsymbol{X}$) were generated as follows,

$$x_{ij} = (1 - \rho_j^2)^{1/2}u_{ij} + \rho_j u_{ip}, j = 1, \cdots, p-1; i = 1, \cdots, n$$

$$u_{ij} \sim N(0, 1),\ j = 1, \cdots, p\,; i = 1, \cdots, n$$

Observations for the dependent variable were obtained with the following equation.

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots + \beta_p X_{ip} + \varepsilon_i, \quad i = 1, \cdots, n$$

In the preliminary study, models were created with different mass parameter values. However, it was observed that parameter selection did not change the end result. For this reason, all of the mass parameters in our study were taken as '1'.

Table 1. MSE Values.

		$n=30$				$n=50$				$n=100$		
	ρ	PLS	PCR	RR	ρ	PLS	PCR	RR	ρ	PLS	PCR	RR
$p=4$	**0**	0.0423	0.4303	0.0349	**0**	0.0222	0.4283	0.0227	**0**	0.0121	0.3952	0.0109
	0.3	0.0380	0.1824	0.0378	**0.3**	0.0211	0.1386	0.0216	**0.3**	0.0108	0.1328	0.0104
	0.5	0.0440	0.127	0.047	**0.5**	0.0238	0.1022	0.0264	**0.5**	0.0116	0.0967	0.0129
	0.7	0.0538	0.0748	0.0591	**0.7**	0.0365	0.0746	0.0385	**0.7**	0.0206	0.0622	0.0210
	0.9	0.1745	0.0662	0.1695	**0.9**	0.0786	0.0432	0.0843	**0.9**	0.0500	0.0340	0.0491
	0.99	1.2815	0.3121	0.5200	**0.99**	0.7709	0.1927	0.3895	**0.99**	0.3780	0.0988	0.2621
$p=7$	**0**	0.0673	0.6716	0.0477	**0**	0.0326	0.5923	0.0256	**0**	0.0125	0.5995	0.0113
	0.3	0.0431	0.2655	0.0421	**0.3**	0.0276	0.2055	0.0261	**0.3**	0.0137	0.1840	0.0107
	0.5	0.0407	0.1454	0.0556	**0.5**	0.0224	0.1225	0.0301	**0.5**	0.0127	0.1128	0.0157
	0.7	0.0534	0.0806	0.0835	**0.7**	0.0297	0.0693	0.0460	**0.7**	0.0187	0.0610	0.0233
	0.9	0.1237	0.0404	0.1767	**0.9**	0.0791	0.0339	0.1166	**0.9**	0.0426	0.0263	0.0543
	0.99	1.1879	0.1314	0.6007	**0.99**	0.7797	0.0868	0.4316	**0.99**	0.4084	0.0460	0.2597
$p=9$	**0**	0.0957	0.7329	0.0543	**0**	0.0432	0.7429	0.0247	**0**	0.0169	0.7265	0.0120
	0.3	0.0567	0.2868	0.0538	**0.3**	0.0342	0.2396	0.0250	**0.3**	0.0195	0.1974	0.0123
	0.5	0.0394	0.1451	0.0659	**0.5**	0.0254	0.1252	0.0347	**0.5**	0.0129	0.1093	0.0148
	0.7	0.0430	0.0760	0.0991	**0.7**	0.0300	0.0651	0.0504	**0.7**	0.0153	0.0561	0.0223
	0.9	0.1089	0.0304	0.2185	**0.9**	0.0783	0.0267	0.1132	**0.9**	0.0409	0.0204	0.0602
	0.99	1.0542	0.1044	0.7882	**0.99**	0.6874	0.0665	0.5328	**0.99**	0.3796	0.0382	0.3379
		$n=200$				$n=500$				$n=1000$		
	ρ	PLS	PCR	RR	ρ	PLS	PCR	RR	ρ	PLS	PCR	RR
$p=4$	**0**	0.0052	0.3988	0.0051	**0**	0.0018	0.3957	0.0018	**0**	0.0010	0.4055	0.0010
	0.3	0.0046	0.1259	0.0048	**0.3**	0.0023	0.1278	0.0020	**0.3**	0.0012	0.1266	0.0010
	0.5	0.0049	0.0898	0.0054	**0.5**	0.0022	0.0890	0.0024	**0.5**	0.0013	0.0890	0.0012
	0.7	0.0117	0.0587	0.0094	**0.7**	0.0060	0.0565	0.0036	**0.7**	0.0036	0.0555	0.0018
	0.9	0.0261	0.0288	0.0237	**0.9**	0.0136	0.0241	0.0111	**0.9**	0.0086	0.0223	0.0048
	0.99	0.1853	0.0503	0.1621	**0.99**	0.0762	0.0224	0.0826	**0.99**	0.0383	0.0125	0.0459
$p=7$	**0**	0.0056	0.6170	0.0050	**0**	0.0021	0.5835	0.0020	**0**	0.0010	0.6248	0.0010
	0.3	0.0093	0.1707	0.0054	**0.3**	0.0042	0.1578	0.0021	**0.3**	0.0010	0.0153	0.0010
	0.5	0.0059	0.1018	0.0062	**0.5**	0.0028	0.1008	0.0029	**0.5**	0.0012	0.0031	0.0012
	0.7	0.0091	0.0582	0.0101	**0.7**	0.0042	0.0558	0.0041	**0.7**	0.0017	0.0011	0.0016
	0.9	0.0234	0.0220	0.0288	**0.9**	0.0119	0.0200	0.0116	**0.9**	0.0045	0.0009	0.0041
	0.99	0.2108	0.0260	0.1508	**0.99**	0.0858	0.0110	0.0678	**0.99**	0.0438	0.0058	0.0362
$p=9$	**0**	0.0063	0.7193	0.0053	**0**	0.0021	0.7232	0.0020	**0**	0.0010	0.7007	0.0010
	0.3	0.0120	0.1759	0.0060	**0.3**	0.0069	0.1668	0.0022	**0.3**	0.0040	0.1612	0.0011
	0.5	0.0074	0.1019	0.0073	**0.5**	0.0037	0.0962	0.0027	**0.5**	0.0024	0.0977	0.0014
	0.7	0.0089	0.0530	0.0111	**0.7**	0.0038	0.0514	0.0044	**0.7**	0.0022	0.0519	0.0021
	0.9	0.0239	0.0183	0.0354	**0.9**	0.0109	0.0166	0.0131	**0.9**	0.0072	0.0168	0.0062
	0.99	0.1968	0.0221	0.2062	**0.99**	0.0817	0.0103	0.1029	**0.99**	0.0416	0.0063	0.0575

4.1. Results and discussion

When the results given in Table 1 and Figures 1–6 are generally evaluated,

(i) Regardless of the sample size and the number of variables, the MSE value obtained via PCR is always decreasing with increasing level of correlation.
(ii) Parallel to the study by Firinguetti, Kibria and Araya [20], regardless of the method used, with increasing sample size, the MSE values of the obtained models decrease.
(iii) Contrary to the findings reported by Firinguetti, Kibria and Araya [20], there was a significant decrease in the MSE values obtained by the RR method as the sample size increased regardless of the level of the relationship. This is because the ridge parameter estimation methods used in both studies are different. The ridge parameter we considered in our study was chosen as one of the best ridge parameters obtained

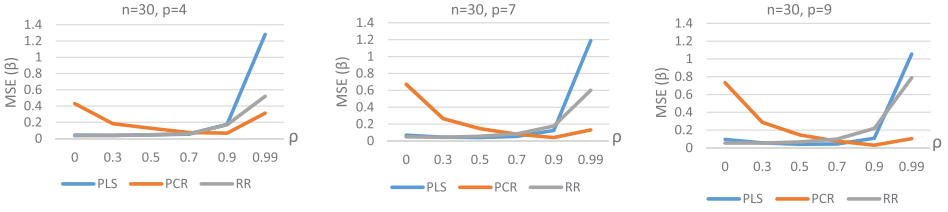

Figure 1. For $n = 30$, Line Graphs for MSE Values Obtained from PLS, PCR and RR Methods.

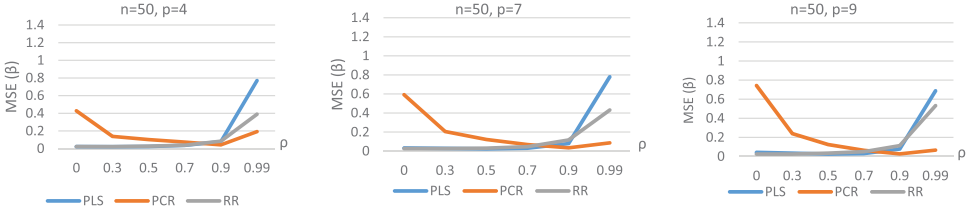

Figure 2. For $n = 50$, Line Graphs for MSE Values Obtained from PLS, PCR and RR Methods.

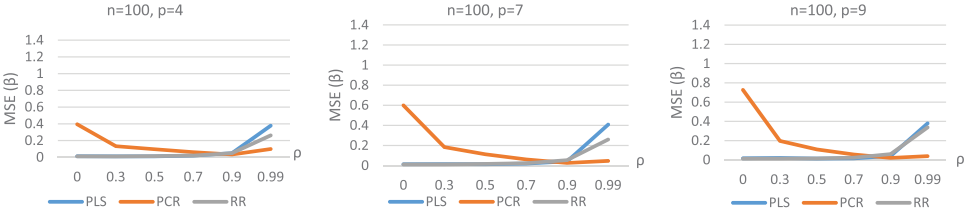

Figure 3. For $n = 100$, Line Graphs for MSE Values Obtained from PLS, PCR and RR Methods.

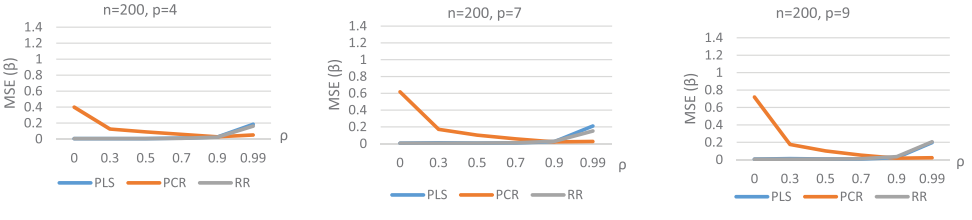

Figure 4. For $n = 200$, Line Graphs for MSE Values Obtained from PLS, PCR and RR Methods.

from the study by Göktaş and Sevinç [26], which compared thirty-seven different ridge parameters.

(iv) Unlike the study of Firinguetti, Kibria and Araya [20], the PCR method was included in the simulation design in addition to the others. In literature, it is generally stated that PLS method is better than PCR in many of the studies performed on a single dataset. Yet, although it is frequently stated that the PLS method is superior to the PCR method as it is a method that considers the dependent variable information, according to the simulation results, however it is determined that this interpretation is not correct for each sample or each level of correlation.

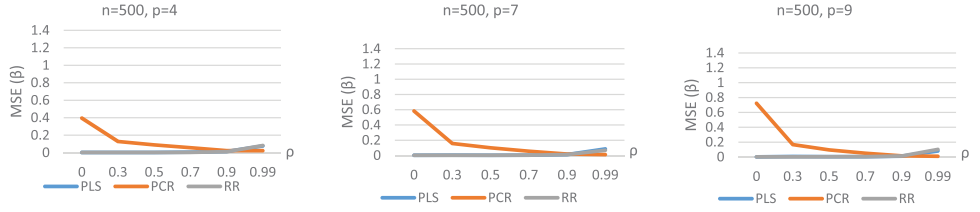

Figure 5. For $n = 500$, Line Graphs for MSE Values Obtained from PLS, PCR and RR Methods.

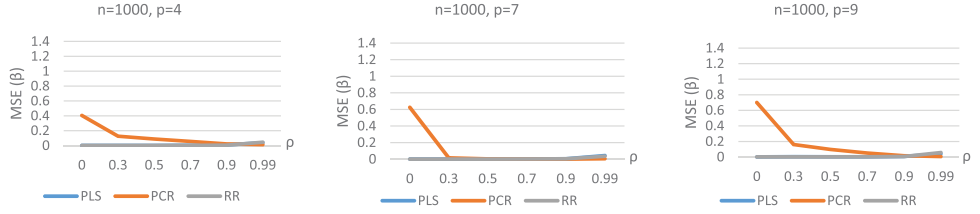

Figure 6. For $n = 1000$, Line Graphs for MSE Values Obtained from PLS, PCR and RR Methods.

(v) In the study by Firinguetti, Kibria and Araya [20], no definite conclusion was reached as to which of the RR and PLS methods was superior in each case. It was determined that RR gave better results than PLS only when the number of explanatory variables was smaller. In our study, when the level of correlation was very low ($< = 0.3$), better results were obtained with the RR method; whereas the PLS method gave better results when the correlation was moderate.

(vi) Generally, only at very high levels of correlation ($\rho \geq 0.9$) did PCR yield better results than the other two methods (n $< = 200$).

(vii) When the sample size was too large ($n \geq 500$), the results obtained with RR were found to be much better.

5. Application on real data set

In order to compare the use of the estimators investigated in the current study and their performance on real data, 'Wage' data of 11 variables and 534 individuals taken from the Current Population Survey (CPS) in 1985 were used.

The definitions of dependent and explanatory variables in the data set used are given below.

Y : Wage (dollars per hour)
X_1: Occupational category (1 = Management; 2 = Sales; 3 = Clerical; 4 = Service; 5 = Professional; 6 = Other)
X_2: Sector (0 = Other; 1 = Manufacturing; 2 = Construction)
X_3: Union (1 = Union member; 0 = Not union member)
X_4: Education (Number of years)
X_5: Experience (Number of years)
X_6: Age (Years)

Table 2. ANOVA table.

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	3948.391	10	394.839	20.388	0.000
Residual	10128.308	523	19.366		
Total	14076.699	533			

Table 3. Coefficients.

	Unstandardized coefficients				Collinearity Statistics
	B	Std. Error	t	Sig.	VIF
Constant	−2.040	6.879	−0.297	0.767	
X1	−0.153	0.131	−1.165	0.245	1.298
X2	0.719	0.388	1.855	0.064	1.199
X3	1.517	0.525	2.889	0.004	1.121
X4	**1.326**	**1.108**	**1.197**	0.232	**231.196**
X5	**0.525**	**1.109**	**0.473**	0.636	**5184.094**
X6	**−0.428**	**1.108**	**−0.386**	0.699	**4645.665**
X7	−2.144	0.399	−5.370	0.000	1.092
X8	0.425	0.420	1.013	0.311	1.096
X9	0.479	0.285	1.676	0.094	1.037
X10	−0.698	2.332	−1.628	0.104	1.047

Table 4. Correlation coefficients between the variables with high degrees of collinearity.

	X4	X5	X6
X4	1	−0.306 (0.000)	−0.107 (0.013)
X5		1	0.973 (0.000)
X6			1

Note: Terms in parentheses give *p* values.

X_7: Sex (1 = Female; 0 = Male)
X_8: Marital Status (0 = Unmarried; 1 = Married)
X_9: Race (1 = Other; 2 = Hispanic; 3 = White)
X_{10}: Southern Region (1 = Person lives in South; 0 = Person lives elsewhere)

OLS estimation and collinearity diagnostic results in Tables 2 and 3 show that the VIF values of X_4, X_5 and X_6 variables are significantly large (VIF $>>$ 10) and all p values that had previously been expected to be significant were found to be insignificant. These results suggest that correlations between variables and collinearity degree may be significant and high.

Since variables are quantitative but not continuous, Spearman rank correlation coefficients were calculated instead of Pearson correlations and are presented in the table below Table 4.

The results obtained when the PLS, PCR and RR biased methods that were used as alternatives to OLS were applied to these data are given in Table 5.

The MSE results show that even though the MSE result of the model predicted by OLS is smaller than those of the others, as the collinearity of the model is high, the model predicted by OLS is unusable. Therefore, if one must make a choice between the biased prediction

Table 5. Regression parameter estimates from OLS, RR, PCR and PLS.

	OLS_B	RR_B	PCR_B	PLS_B
(Constant)	−2.040	−0.482	−8.530	−4.251
X1	−0.153	−0.154	0.035	−0.161
X2	0.719	0.674	0.057	0.493
X3	1.517	1.334	0.432	1.705
X4	1.326	1.438	0.068	0.793
X5	0.525	0.665	−0.365	0.019
X6	−0.428	−0.571	−0.304	0.060
X7	−2.144	−2.006	−0.384	−2.159
X8	0.425	0.392	−0.075	0.606
X9	0.479	0.395	−0.038	0.570
X10	−0.698	−0.686	0.048	−1.072
MSE	18.338	19.389	21.060	19.170

methods, the model of the method that gives the smallest MSE value is desired. According to our results, the smallest MSE value was obtained by the PLS method. As there are 534 observations, 10 independent variables and a high collinearity degree in the data set used in the application, we can compare the results with $n = 500$, $p = 9$, and $\rho = 0.9$ in the simulation design. Accordingly, it can be said that the results obtained in the application are parallel to the simulation results.

6. The concluding remarks

In the current study, the PLS, RR and PCR methods used to solve the problem of multicollinearity emerging in multiple linear regression models were compared. For the data set obtained through simulation, which of these methods give more effective results depending on the number of variables, sample size and correlation level has been shown. However, in a similar study by Firinguetti, Kibria and Araya [20], it was pointed out that there is no biased prediction method that demonstrated a superior performance for a particular case. Beside the inclusion of the PCR method, priorities for each method were determined for different cases in the current study. In particular, it was observed that at certain levels of correlation, the question of which method is more preferable could be easily answered. In cases where the level of correlation was very low ($\rho \leq 0.3$), RR methods have yielded better results; whereas in the case of moderate level correlation, the PLS method generally is better. Only at very high levels of correlation ($\rho \geq 0.9$), the PCR method yielded surprisingly much better results than both PLS and RR. Furthermore, when the sample size was large ($n \geq 500$), it was observed that PLS and RR generally were considerably improving. When the degree of collinearity increases the MSE values obtained from PLS and RR are increasing no matter what the sample size or the number of regressors is whereas the MSE values obtained from PCR method keeps decreasing which is interesting. In conclusion PCR method produces much better estimates when there exists severe multicollinearity among regressors.

For illustration an empirical study of the popular 'Wage' dataset from Current Population Survey (1985) where the sample can be treated as $n = 500$, $p = 9$, and $\rho = 0.9$ or $\rho = 0.99$ has been used to make the comparison among the bias prediction methods. The results (Tables 2–5) obtained are supporting the simulation results in favour of PLS method.

In addition this study offers guidance to researchers who have to use biased prediction methods, simplifying method choice by taking into account degree of multicollinearity, number of independent variables and sample size used.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970a;12(1):55–67.
[2] Hoerl AE, Kennard RW. Ridge regression: applications to non-orthogonal problems. Technometrics. 1970b;12(1):69–82.
[3] Hoerl AE, Kennard RW, Baldwin KF. Ridge regression: some simulations. Commun Stat. 1975;4(2):105–123.
[4] Kidwell JS, Brown LH. Ridge regression as a technique for analyzing models with multicollinearity. J. Marriage and Family. 1982;44(2):287–299.
[5] Yeniay Ö, Göktaş A. A comparison of partial least squares regression with other prediction methods. Hacettepe J Math Stats. 2002;31:99–111.
[6] Kibria BMG. Performance of some new ridge regression estimators. Commun Stat-Simul C. 2003;32(2):419–435.
[7] Graham MH. Confronting multicollinearity in ecological multiple regression. Ecology. 2003;84(11):2809–2815.
[8] Albayrak SA. Çoklu doğrusal bağlanti halinde en küçük kareler tekniğinin alternatifi yanli tahmin teknikleri ve bir uygulama. ZKÜ Sosyal Bilimler Dergisi. 2005;1(1):105–126.
[9] Karadavut U, Genç A, Tozluca A, et al. Nohut (cicer arietinum l.) bitkisinde verime etki eden bazi karakterlerin alternatif regresyon yöntemleriyle karşılaştırılması. Tarım Bilimleri Dergisi. 2005;11(3):328–333.
[10] Mansson K, Shukur G, Kibria BMG. A simulation study of some ridge regression estimators under different distributional assumptions. Commun Stat-Simul C. 2010;39(8):1639–1670.
[11] Topal M, Eyduran E, Yağanoğlu AM, et al. Çoklu doğrusal bağlanti durumunda ridge ve temel bileşenler regresyon analiz yöntemlerinin kullanımı. Atatürk Üniversitesi Ziraat Fakültesi Dergisi. 2010;41(1):53–57.
[12] Li Y. A comparison study of principle component regression, partial least squares regression and ridge regression with application to FTIR data. Master thesis in statistics Faculty of Social Sciences Uppsala University, Sweden; 2010.
[13] Rathert TÇ, Üçkardeş F, Narinç D, et al. Comparison of principal component regression with the least square method in prediction of internal egg quality characteristics in Japanese quails. Kafkas Üniversitesi Veterinerlik Fakültesi Dergisi. 2011;17(5):687–692.
[14] Acharjee A, Finkers R, Visser RGF, et al. Comparison of regularized regression methods for omics dat. Metabolomics. 2013;3(3):1–9.
[15] Mahesh S, Jayas DS, Paliwal J, et al. Comparison of partial least squares regression and principal components regression methods for protein and hardness predictions using the near-infrared hyperspectral images of bulk samples of Canadian wheat. Food Bioproc Tech. 2014;8(1):31–40.
[16] Simeon O, Timothy AO, Thompson OO, et al. Comparison of classical least squares, rigde and principal component methods of regression analysis using gynecological data. IOSR J Math. 2014;9(6):61–74.
[17] Goharnejad A, Zarei A, Tahmasebi P. Comparing multiple regression, principal component analysis, partial least square regression and ridge regression in predicting rangeland biomass in the semi steppe rangeland of Iran. Environ Nat Res J. 2014;12(1):1–21.

[18] Polat E, Günay S. The comparison of partial least squares regression, principal component regression and ridge regression with multiple linear regression for predicting PM10 concentration level based on meteorological parameters. J Data Sci. 2015;13:663–692.
[19] Brito RS, Almeida MC, Matos JS. Estimating flow data in urban drainage using partial least squares regression. Urban Water Journal. 2017;14(5):467–474.
[20] Firinguetti L, Kibria G, Araya R. Study of partial least squares and ridge regression methods. Commun Stat-Simul Comput. 2017;46(8):6631–6644.
[21] Kibria BMG, Banik S. A simulation study on the size and power Properties of some ridge regression Tests. Appl Appl Math Int J (AAM). 2019;14(2):741–761.
[22] Wold H. Estimation of Principle components and related models by Iterative Least squares. In: Krishnaiaah PR, editor. Multivariate analysis. New York: Academic Press; 1966. p. 391–420.
[23] Phatak A, De Jong S. The geometry of partial least squares. J Chemom. 1997;11:311–338.
[24] Judge GG, Griffiths WE, Hill RC, et al. The theory and practice of econometrics. Wiley&Sons; 1985. p. 1056. ISBN: 978-0-471-89530-5.
[25] Hotelling H. Analysis of complex of statistical variables into principle components. J Educ Psychol. 1993;24(7):498–520.
[26] Göktaş A, Sevinç V. Two new ridge parameters and a guide for selecting an appropriate ridge parameter in linear regression. Gazi University J Sci. 2016;29(1):201–211.