Hacettepe Journal of Mathematics and Statistics
Volume 31 (2002), 99–111

A COMPARISON OF PARTIAL LEAST SQUARES REGRESSION WITH OTHER PREDICTION METHODS

Özgür Yeniay* and Atilla Göktaş*

Received 25. 06. 2002

Abstract

The aim of this study is to compare popular regression methods with the partial least squares method. The paper presents a theoretical point of view, as well as an application on a real data set. It is found that partial least squares regression yields somewhat better results in terms of the predictive ability of models obtained when compared to the other prediction methods.

Key Words: Ordinary least squares, Ridge regression, Principal component regression, Partial least squares.

1. Introduction

In several linear regression and prediction problems, the independent variables may be many and highly collinear. This phenomenon is called multicollinearity and it is known that in this case the ordinary least squares (OLS) estimator for the regression coefficients or predictor based on these estimates may give very poor results [4]. Even for finite or moderate samples, collinearity problem may still exist. Plenty of methods have been developed to overcome this problem, such as principal component regression (PCR), ridge regression (RR) and partial least squares (PLS).

Two of the most used methods, namely PCR and RR, require a large amount of computation when the number of variables is large [5]. PCR handles the collinearity problem with few factors. However, PLS may even overcome the collinearity problem with fewer factors than PCR. Meanwhile simulations tend to show that PLS reaches its minimal mean square error (MSE) with a smaller number of factors than PCR. Hence, PLS gives a unique way of choosing factors, contrary to PCR, and it requires less computations than both PCR and RR. In the following sections theoretical aspects of those methods have been presented. Then, the methods mentioned above have been compared using real data and the results obtained have been discussed to stress the advantage of PLS.

*Hacettepe University, Faculty of Science, Department of Statistics, 06532 Beytepe, Ankara, Turkey.

2. OLS, PCR, RR and PLS

In this section OLS, PCR and RR are briefly outlined while PLS is presented in more detail. First of all, the vector of coefficients in the linear regression is given. The regression model used for these methods is defined by the equation,

$$\mathbf{y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where,

$\mathbf{y}$ is a $n \times 1$ vector of observations on the dependent variable,

β_0 is an unknown constant,

$\mathbf{X}$ is a $n \times p$ matrix consisting of n observations on p variables,

$\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression coefficients, and

$\boldsymbol{\varepsilon}$ is a $n \times 1$ vector of errors identically and independently distributed with mean zero and variance σ^2.

If the variables included in the matrix $\mathbf{X}$ and the vector $\mathbf{y}$ are mean centered, equation (1) can be simplified as follows;

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2}$$

When there is more than one dependent variable, the equation (2) can be written as,

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E} \tag{3}$$

where,

$\mathbf{Y}$ is a $n \times q$ matrix of observations on q dependent variables $y_1, y_2, \ldots, y_q$,

$\mathbf{E}$ is a $n \times q$ matrix of errors, whose rows are independently and identically distributed, and

$\mathbf{B}$ is a $p \times q$ matrix of parameters to be estimated.

2.1. Ordinary Least Squares

When the matrix X has a full rank of p, the OLS estimator $\hat{\boldsymbol{\beta}}_{\mathbf{OLS}}$ can be obtained by minimizing the sum of squared residuals,

$$\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \tag{4}$$

Hence,

$$\hat{\boldsymbol{\beta}}_{\mathbf{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \tag{5}$$

where $\hat{\boldsymbol{\beta}}_{\mathbf{OLS}}$ is a $p \times 1$ vector of estimated parameters. $\hat{\boldsymbol{\beta}}_{\mathbf{OLS}}$ provides unbiased estimates of the elements of $\boldsymbol{\beta}$, which have the minimum variance of any linear function of the observations. When there are q dependent variables, the OLS estimator in equation (5) can be generalized as follows;

$$\hat{\mathbf{B}}_{\mathbf{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \tag{6}$$

where $\hat{\mathbf{B}}_{\mathbf{OLS}}$ is the least square estimate of $\mathbf{B}$. When the independent variables are highly correlated, $\mathbf{X}'\mathbf{X}$ is ill-conditioned and the variance of the OLS estimator becomes large. With multicollinearity, the estimated OLS coefficients may be statistically insignificant (too large, too small and even have the wrong sign) even though the R-Square may be high. Therefore, a number of alternative estimation methods which settle into a category called *biased estimation methods* have been proposed and designed to combat multicollinearity. If we quit insisting on unbiassedness, biassed methods such as, PCR, RR and PLS can be used to tackle that problem of imprecise predictions.

2.2. Principal Component Regression

PCR [9] is one way to deal with the problem of ill-conditioned matrices [11]. What has been done basically is to obtain the number of principal components (PCs) providing the maximum variation of $\mathbf{X}$ which optimizes the predictive ability of the model. PCR is actually a linear regression method in which the response is regressed on the PCs. Consider $\mathbf{X}$ as mean centered and scaled (Mean-centering is achieved by subtracting the mean of the variable vector from all the columns of $\mathbf{X}$. Variable scaling is also used to remove differences in units between variables, which can be accomplished by dividing each element of the mean centered $\mathbf{X}$ by the root sum of squares of that variable), then

$$\mathbf{X}'\mathbf{X}\gamma_i = \lambda_i\gamma_i, \ i = 1, 2, \ldots, p, \tag{7}$$

where the λ_i's are the eigenvalues of the correlation matrix $\mathbf{X}'\mathbf{X}$ and the γ_i's are the unit-norm eigenvectors of $\mathbf{X}'\mathbf{X}$. The vector γ_i is used to re-express the X's in terms of PC Z's in the form,

$$Z_i = \gamma_{1i}X_1 + \gamma_{2i}X_2 + \cdots + \gamma_{pi}X_p \tag{8}$$

These Z_i's are orthogonal to each other and called the *artificial variables* [12]. Assume that the first m PCs optimize the predictive ability of the model. Then

$$y = Z_m\alpha_m + \epsilon, \tag{9}$$

where $\alpha_m = (Z_m'Z_m)^{-1}Z_m'y$ and m is the number of PCs retained in the model. Using estimates $\boldsymbol{\alpha}$, it is easy to get back to the estimates of $\boldsymbol{\beta}$ as,

$$\hat{\boldsymbol{\beta}}_{\mathbf{PCR}} = V_m\alpha_m, \tag{10}$$

where V_m is a matrix consisting of the first m unit-norm eigenvectors. PCR gives a biassed estimate of the parameters. If all of the PC's are used instead of using the first m PC's, then $\hat{\boldsymbol{\beta}}_{\mathbf{PCR}}$ becomes identical to $\hat{\boldsymbol{\beta}}_{\mathbf{OLS}}$ [9].

2.3. Ridge Regression

Another method to overcome the multicollinearity problem amongst the regressors is RR [8]. RR was first suggested by Hoerl [7]. When multicollinearity exists, the matrix $\mathbf{X}'\mathbf{X}$, where $\mathbf{X}$ consists of the original regressors, becomes nearly singular. Since $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, and the diagonal elements of $(\mathbf{X}'\mathbf{X})^{-1}$ become quite

large, this makes the variance of $\hat{\boldsymbol{\beta}}$ to be large. This leads to an unstable estimate of $\boldsymbol{\beta}$ when OLS is used. In RR, a standardized $\mathbf{X}$ is used and a small positive constant θ is added to the diagonal elements of $\mathbf{X}'\mathbf{X}$. The addition of a small positive number θ to the diagonal elements of $\mathbf{X}'\mathbf{X}$ causes $\mathbf{X}'\mathbf{X}$ to be non-singular. Thus,

$$\hat{\boldsymbol{\beta}}_{\mathbf{RIDGE}} = (\mathbf{X}'\mathbf{X} + \theta\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}, \tag{11}$$

where $\mathbf{I}$ is the $p \times p$ identity matrix and $\mathbf{X}'\mathbf{X}$ is the correlation matrix of independent variables. Values of *theta* lie in the range $(0, 1)$. When $\theta = 0$, $\hat{\boldsymbol{\beta}}_{\mathbf{RIDGE}}$ becomes $\hat{\boldsymbol{\beta}}_{\mathbf{OLS}}$. Obviously, a key aspect of ridge regression is determining what the best value of the constant that is added to the main diagonal of the matrix $\mathbf{X}'\mathbf{X}$ should be to maximize prediction. There are many procedures in the literature for determining the best value. The simplest way is to plot the values of each $\hat{\boldsymbol{\beta}}_{\mathbf{RIDGE}}$ versus θ. The smallest value for which each ridge trace plot shows stability in the coefficient is adopted [10].

2.4. Partial Least Squares

PLS is a reasonably new method developed by Herman Wold in the 1960s as a method for constructing predictive models when the explanatory variables are many and highly collinear [5-6,12]. It may be used with any number of explanatory variables, even for more than the number of observation. Although PLS is heavily promoted and used by chemometricians, it is largely unknown to statisticians [1].

To regress the $\mathbf{Y}$ variables with the explanatory variables $X_1, \ldots, X_p$, PLS attempts to find new factors that will play the same role as the $\mathbf{X}$'s. These new factors often called *latent variables* or *components*. Each component is a linear combination of $X_1, \ldots, X_p$. 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 $\mathbf{Y}$ variables. The major difference is, while PCR uses only the variation of $\mathbf{X}$ to construct new factors, PLS uses both the variation of $\mathbf{X}$ and $\mathbf{Y}$ to construct new factors that will play the role of explanatory variables.

The intension of PLS is to form components that capture most of the information in the $\mathbf{X}$ variables, that is useful for predicting $y_1, \ldots, y_q$, while reducing the dimensionality of the regression problem by using fewer components than the number of $\mathbf{X}$ variables [2].

Now we are going to derive the PLS estimators of $\boldsymbol{\beta}$ and $\mathbf{B}$. The matrix $\mathbf{X}$ has a bilinear decomposition in the following form;

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1' + \mathbf{t}_2\mathbf{p}_2' + \cdots + \mathbf{t}_p\mathbf{p}_p' = \sum_{i=1}^{\mathbf{p}} \mathbf{t}_i\mathbf{p}_i' = \mathbf{T}\mathbf{P}'. \tag{12}$$

Here the $\mathbf{t_i}$ are linear combinations of $\mathbf{X}$, which we will write as $\mathbf{X}\mathbf{r_i}$. The $p \times 1$ vectors p_i are often called loadings. Unlike the weights in PCR (i.e. the eigenvectors γ_i), the $\mathbf{r_i}$ are not orthonormal. The $\mathbf{t_i}$, however, like the principal components Z_i, are orthogonal. There are two popular algorithms for obtaining the PLS estimators. One is called NIPALS and the other one is called the SIMPLS algorithm.

In the first one this orthogonality is imposed by computing the $\mathbf{t_i}$ as linear combinations of residual matrices $\mathbf{E_i}$, in other words, as

$$\mathbf{t}_i = \mathbf{E}_{i-1}\mathbf{w}_i, \quad \mathbf{E}_i = \mathbf{X} - \sum_{j=1}^{i} \mathbf{t}_j \mathbf{p}_j', \quad \mathbf{E}_0 = \mathbf{X}, \tag{13}$$

where the $\mathbf{w}_i$ are orthonormal. Then two sets of weight vectors $\mathbf{w}_i$ and $\mathbf{r}_i$, $i = 1, 2, \ldots, m$, span the same space [12]. In most algorithms for both multivariate and univariate PLS, the first step is to derive either $\mathbf{w}_i$ or $\mathbf{r}_i$, $i = 1, \ldots, m$, in order to be able to calculate the linear combination of the $\mathbf{t}_i$. Then $\mathbf{p}_i$ are calculated by regressing $\mathbf{X}$ onto $\mathbf{t}_i$. When m factors are to be taken into account, the following relationship can be written,

$$\mathbf{T}_m = \mathbf{X}\mathbf{R}_m \tag{14}$$

$$\mathbf{P}_m = \mathbf{X}'\mathbf{T}_m(\mathbf{T}_m'\mathbf{T}_m)^{-1} \tag{15}$$

$$\mathbf{R}_m = \mathbf{W}_m(\mathbf{P}_m'\mathbf{W}_m)^{-1} \tag{16}$$

where the first m dominant factors, which capture most of the variance in $\mathbf{X}$, have maximum ability for predictive models. Equation (16) connects two sets of weight vectors by a linear transformation. From equations (14) and (15), $\mathbf{P}_m'\mathbf{R}_m$ equals I_m, since such a transformation exists. Also $\mathbf{R}_m'\mathbf{P}_m$ equals $\mathbf{I}_m$,

$$\mathbf{R}_m'\mathbf{P}_m = \mathbf{R}_m'\mathbf{X}'\mathbf{T}_m(\mathbf{T}_m'\mathbf{T}_m)^{-1} = \mathbf{T}_m'\mathbf{T}_m(\mathbf{T}_m'\mathbf{T}_m)^{-1} = \mathbf{I}_m. \tag{17}$$

After m dimensions have been extracted, the vector of fitted values from PLS can be represented by the first m PLS linear combinations $\mathbf{T}_m$. Thus the following equation is obtained;

$$\hat{\mathbf{y}}_{\mathbf{PLS}}^m = \mathbf{T}_m(\mathbf{T}_m'\mathbf{T}_m)^{-1}\mathbf{T}_m'\mathbf{y}. \tag{18}$$

Notice that this is the derivation only for the univariate case. The multivariate case is identical to the univariate case except that the vector $\hat{\mathbf{y}}_{\mathbf{PLS}}^m$ should be replaced by the matrix $\hat{\mathbf{Y}}_{\mathbf{PLS}}^m$[12]. Substituting $\mathbf{X}\mathbf{R}_m$ for $\mathbf{T}_m$ and $\hat{\boldsymbol{\beta}}_{\mathbf{OLS}}$for $\mathbf{y}$ results in

$$\hat{\mathbf{y}}_{\mathbf{PLS}}^m = \mathbf{X}\mathbf{R}_m(\mathbf{R}_m'\mathbf{X}'\mathbf{X}\mathbf{R}_m)^{-1}\mathbf{R}_m'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{OLS}}. \tag{19}$$

Then it is clear that

$$\hat{\boldsymbol{\beta}}_{\mathbf{PLS}}^m = \mathbf{R}_m(\mathbf{R}_m'\mathbf{X}'\mathbf{X}\mathbf{R}_m)^{-1}\mathbf{R}_m'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{OLS}}. \tag{20}$$

A somewhat a simpler expression for $\hat{\boldsymbol{\beta}}_{\mathbf{OLS}}$ can be obtained by first substituting equation (14) into (15), which yields $\mathbf{P}_m = \mathbf{X}'\mathbf{X}\mathbf{R}_m(\mathbf{R}_m'\mathbf{X}'\mathbf{X}\mathbf{R}_m)^{-1}$. Then using this result in equation (20) gives

$$\hat{\boldsymbol{\beta}}_{\mathbf{PLS}}^m = \mathbf{R}_m\mathbf{P}_m'\hat{\boldsymbol{\beta}}_{\mathbf{OLS}} = \mathbf{W}_m(\mathbf{P}_m'\mathbf{W}_m)^{-1}\mathbf{P}_m'\hat{\boldsymbol{\beta}}_{\mathbf{OLS}}. \tag{21}$$

In the the multivariate case $\hat{\mathbf{B}}_{\mathbf{PLS}}^m$ has a similar form. The only difference is that $\hat{\boldsymbol{\beta}}_{\mathbf{OLS}}$ is replaced by $\hat{\mathbf{B}}_{\mathbf{OLS}}$ [12] i.e.,

$$\hat{\mathbf{B}}_{\mathbf{PLS}}^m = \mathbf{W}_m(\mathbf{P}_m'\mathbf{W}_m)^{-1}\mathbf{P}_m'\hat{\mathbf{B}}_{\mathbf{OLS}} \tag{22}$$

3. A Comparative Study of OLS, RR, PCR and PLS on Real Data

Before comparing the predictive ability of the models, it is useful to introduce several measures of a model's fit to the data and of predictive power used in our study. In all of the measures considered, we are attempting to estimate the average deviation of the model from the data. The root-mean square error (RMSE) tells us about the fit of the model to the data set used. It is defined as,

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{23}$$

where the $\hat{y}_i$ are the values of the predicted variable when all samples are included in the model formation, and n is the number of observations. RMSE is a measure of how well the model fits the data. This is in contrast to the root-mean-square error of cross-validation (RMSECV), which is a measure of a model's ability to predict new samples. The RMSECV is defined as in Equation (23), except the y_i are predictions for samples not included in the model formulation. RMSECV is related to the PRESS value for the number of PC's or latent variables (LV's) included in the model, i.e.

$$\text{RMSECV} = \sqrt{\frac{\text{PRESS}}{n}} \tag{24}$$

where PRESS is the sum of squares prediction error. PRESS is calculated via a leave one out cross-validation, i.e. where each sample is left out of the model formulation and predicted once. It is also possible to calculate a root-mean-square error of prediction (RMSEP) when the model is applied to new data provided that the reference values for the new data are known. RMSEP is calculated exactly as in Equation (23) except that the estimates y_i are based on a previously developed model, not one in which the samples to be "predicted" are included in the model building. In our case new data are not provided. Therefore we will not be interested in calculating RMSEP.

3.1. Model Fitting by All Prediction Methods

In this section, we will compare the OLS, RR, PCR and PLS prediction methods on a real data set. The data with 80 observations consists of the gross domestic product per capita (GDPPC) and various variables affecting GDPPC in Turkey. Each observation represents one of the 80 provinces in Turkey. The data are taken from The State Planning Organization [14]. The variables used are described in Table 1.

Table 1. Description of the Variables

Variables	**Description**
Y	Gross domestic product per capita (GDPPC)
X_1	Number of private cars
X_2	Population according to provinces
X_3	Number of schools
X_4	Number of students
X_5	Number of teachers
X_6	Number of doctors
X_7	Number of dentists
X_8	Number of other health staff
X_9	Retirement salaried population from pension fund
X_{10}	Electricity consumption (dwellings)
X_{11}	Electricity consumption (commercial)
X_{12}	Electricity consumption (industrial)
X_{13}	Consolidated budget tax income
X_{14}	Public investment expenditure
X_{15}	Number of motor vehicles
X_{16}	Amount of private sector investment incentives
X_{17}	Number of employees in private sector investment incentives
X_{18}	Number of insured population connected with SII
X_{19}	Number of retirement salaried population connected with SII
X_{20}	Number of tax payers
X_{21}	Total bank deposits
X_{22}	Total bank credits per capita
X_{23}	Crops production value
X_{24}	Livestock production value
X_{25}	Animal products production value
X_{26}	Number of rural settlements with adequate drinking water supply

Using the ordinary least squares method, coefficients and their related statistics have been calculated and are presented in Table 2. Those calculations have been performed using Matlab via PLS Toolbox 2.1 [13].

Table 2. Coefficients and Collinearity Statistics

Model	Unstandardized Coefficients		Coeff.	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
Const.	-1714.5	528591.059		-0.003	0.997		
X_1	-55.524	23.592	-7.914	-2.354	0.022	0.000	2393.922
X_2	-1.043	0.821	-1.427	-1.270	0.210	0.004	267.175
X_3	88.113	12543.582	0.005	0.007	0.994	0.011	94.617
X_4	72.613	36.871	2.043	1.969	0.054	0.004	227.950
X_5	-1239.6	664.815	-2.315	-1.865	0.068	0.003	326.386
X_6	222.033	476.390	0.590	0.466	0.643	0.003	339.256
X_7	328.163	3439.452	0.184	0.095	0.924	0.001	789.592
X_8	-293.169	328.627	-1.056	-0.892	0.376	0.003	296.590
X_9	88.736	44.571	3.429	1.991	0.052	0.002	628.148
X_{10}	-2.826	3.240	-1.824	-0.872	0.387	0.001	925.461
X_{11}	0.477	2.655	0.155	0.180	0.858	0.006	158.090
X_{12}	-0.748	0.442	-1.012	-1.695	0.096	0.013	75.535
X_{13}	7.026	1.884	0.824	3.729	0.000	0.097	10.341
X_{14}	3.0E-02	0.013	0.405	2.302	0.025	0.153	6.538
X_{15}	11.941	10.442	2.229	1.144	0.258	0.001	804.182
X_{16}	-5.0E-03	0.005	-0.502	-1.002	0.321	0.019	53.240
X_{17}	62.252	40.616	0.559	1.533	0.131	0.036	28.116
X_{18}	22.224	13.366	4.584	1.633	0.102	0.001	1608.990
X_{19}	6.558	10.431	0.606	0.629	0.532	0.005	196.703
X_{20}	10.222	19.705	0.836	0.519	0.606	0.002	549.737
X_{21}	6.7E-03	0.003	0.722	1.959	0.055	0.035	28.730
X_{22}	-8.0E-04	0.004	-0.076	-0.233	0.817	0.044	22.659
X_{23}	1.7E-02	0.008	0.543	2.142	0.037	0.073	13.611
X_{24}	-8.0E-03	0.021	-0.075	-0.396	0.694	0.131	7.611
X_{25}	-2.0E-02	0.024	-0.192	-0.972	0.336	0.121	8.256
X_{26}	8387.467	6489.710	0.135	1.292	0.202	0.431	2.319

Looking at both the tolerance and VIF columns, only four of the independent variables (x_{14}, x_{24}, x_{25}, and x_{26}) are significantly non collinear, but the rest are highly correlated. This leads to imprecise predictions.

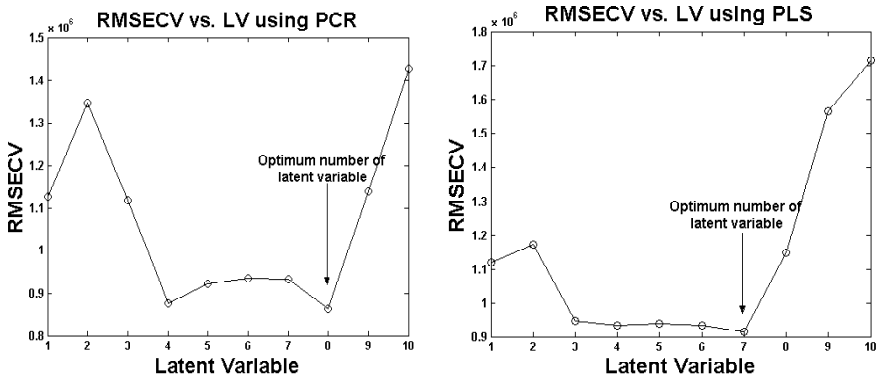
Using the ordinary least squares method, an ANOVA table has been performed in order to test whether the model, in which the regressors may be a linear combination of the predicted variable, is significant.

Table 3. Analysis of Variance Results for GDPPC Data.

Model	Sum of Squares	df.	Mean Square	F	Sig.
Regression	4.07E+13	26	1.5657E+12	6.105	0.000
Residual	1.36E+13	53	2.5647E+11		
Total	5.43E+13	79			

From Table 3, it is clear that the model is significant with a probability of 95%. Even though the OLS model fits the data well, multicollinearity may severely prohibit quality prediction. RMSECV values for both PCR and PLS are calculated and plotted as a function of the number of latent variables in Figure 1.

Figure 1. Latent Variables Versus RMSECV for PCR and PLS



As can be seen from Figure 1, the optimal number of latent variable for PCR, which is determined by the minimum RMSECV value, is 8 whereas it is 7 for PLS.

Table 4 and Table 5 present the percent variance captured by the model. For the optimal number of latent variable in PCR, 100% of the variance is captured by the regressors. These 8 latent variables could explain 79.65% of the variation. This is equivalent to R-Square in OLS. For the optimal number of latent variable in PLS, 100% of the variance is captured by the regressors. These 7 latent variables could explain 79.59% of the variation.

Table 4. Percent Variance Captured by Regression Model Using PCR

	X block		Y block	
LV	**This LV**	**Total**	**This LV**	**Total**
1	86.05	86.12	53.69	53.69
2	5.15	94.49	2.06	55.75
3	6.25	97.59	19.63	75.38
4	2.09	99.55	3.60	78.98
5	0.28	99.83	0.11	79.09
6	0.15	99.97	0.05	79.14
7	0.04	100.00	0.04	79.18
8	**0.00**	**100.00**	**0.47**	**79.65**
9	0.00	100.00	1.98	81.63
10	0.00	100.00	1.51	83.14

Table 5. Percent Variance Captured by Regression Model Using PLS

	X block		Y block	
LV	**This LV**	**Total**	**This LV**	**Total**
1	86.05	86.05	55.63	55.63
2	5.15	91.19	19.54	75.17
3	6.25	97.44	3.60	78.77
4	2.09	99.53	0.28	79.05
5	0.28	99.81	0.12	79.16
6	0.15	99.96	0.07	79.24
7	**0.04**	**100.00**	**0.35**	**79.59**
8	0.00	100.00	3.23	82.82
9	0.00	100.00	0.44	83.26
10	0.00	100.00	3.97	87.23

It should be evident from the comparison of the models that fit and prediction are entirely different aspects of a model's performance. If prediction is the goal, a model that gives the minimum RMSECV value amongst the prediction models should be selected.

Before starting with comparisons of the models according to RMSE and RMSECV, it is useful to present some plots describing how well the model fits the data. These plots are presented in Figure 2 and they help us to visualize the model performance for all the prediction methods.

Figure 2. Actual and Predicted Values for OLS, RR, PCR and PLS

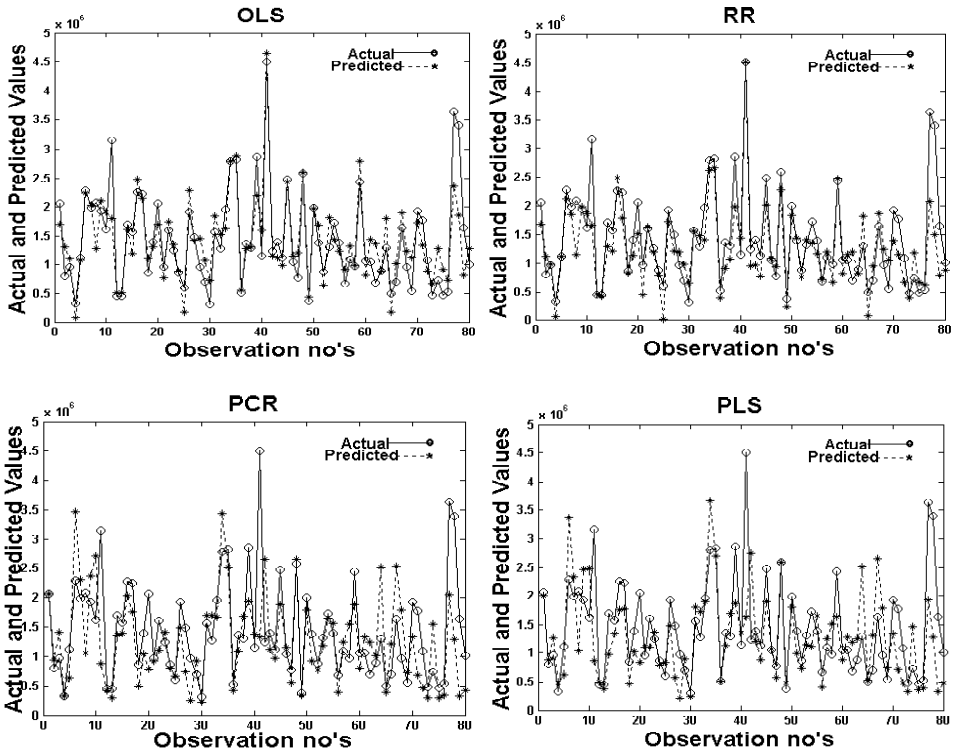


Table 7 presents RMSE and RMSECV values obtained for the OLS, PCR, PLS and RR models.

Table 7. RMSE and RMSECV Values for All Prediction Methods

	OLS	**PCR**	**PLS**	**RR**
RMSE	**506428.868***	616250	591592	**507285.736***
RMSECV	1974920.421	1337700	**1335270***	1377270.852

As can be seen from Table 7, the OLS model has the smallest RMSE value. The second smallest RMSE value belongs to the RR model. Under the condition of no collinearity in the independent variables, this indicates that, the RR and OLS models fit the data better than both PCR and PLS. Due to the existence of the collinearity in the data used, this interpretation would not be true at all. For comparison of models intended for prediction it is inadequate to look just at model fit. Note, however, from the RMSECV that the OLS does not predict as well as the other prediction methods, even for samples within the original data. Here the best models are PLS and PCR, respectively.

4. Conclusion

In this comparative study, OLS, RR, PCR and PLS have been applied to real data with high collinearity, and compared from the point of view of model fit and prediction. The results show that when the model fit is considerable, RR and OLS seem to fit the data best. However the existence of collinearity among the regressors prevents us making this comment. The data set used in this paper has also proved that the regression model constructed by PLS has the highest predictive ability with the smallest number of factors. This is advantageous in that there are fewer factors to interpret.

References

[1] Frank, I. E. and Friedman, J. H. A statistical view of some chemometrics regression tools (with discussion), Technometrics **35**, 109–135, 1993.

[2] Garthwaite, P. H. An interpretation of partial least squares, Journal of the American Statistical Association **89**, 122–127, 1994.

[3] Geladi, P. and Kowalski B. R. Partial least-squares regression: A Tutorial, Analytica Chimica Acta **185**, 1–17, 1986.

[4] Gunst, R. F. and Mason, R. L. Some considerations in the evaluation of alternate prediction equations, Technometrics **21**, 55–63, 1979.

[5] Helland, I. S. On the structure of partial least squares regression, Communications in Statistics, Simulations and Computation **17**, 581–607, 1988.

[6] Helland, I. S. Partial least squares regression and statistical methods, Scandinavian Journal of Statistics **17**, 97–114, 1990.

[7] Hoerl, A. E. Application of ridge analysis to regression problems, Chemical Engineering Progress **58**, 54–59, 1962.

[8] Hoerl, A. E. and Kennard, R. W. Ridge regression biased estimation for nonorthogonal problems, Technometrics **8**, 27–51, 1970.

[9] Massy, W. F. Principal component regression in exploratory statistical research, Journal of the American Statistical Association **60**, 234–246, 1965.

[10] Mayers, R. H. Classical and modern regression with applications, 2nd edition, Duxbury Press, 1990.

[11] Naes, T. and Martens, H. Principal component regression in NIR analysis: Viewpoints, Background Details and Selection of Components, Journal of Chemometrics **2**, 155–167, 1988.

[12] Phatak, A. and Jong, S. D. The geometry of partial least squares, Journal of Chemometrics **11**, 311–338, 1997.

[13] PLS-Toolbox Version 2.1, Eigenvector Research Inc. Manson, WA, 2000.

[14] The State Planning Organization (SPO), Various indicators related to Provinces and Regions, Ankara, 1999.