A Note on Partial Least Squares Regression for Multicollinearity (A Comparative Study)

Moawad El-Fallah Abd El-Salam
Department of Statistics & Mathematics and Insurance
Faculty of Commerce, Zagazig University, Egypt

Abstract

This paper presents and compares the partial least squares (PLS) regression as an alternative procedure for handling multicollinearity problem with two commonly used regression methods, which are ridge regression (RR) and principle component regression (PCR) .The performances of RR, PCR and PLS are compared to help and give future researchers a comprehensive view about the best procedure to handle multicollinearity problem. A Monte Carlo simulation study was used to evaluate the effectiveness of these three procedures. For comparison purposes, mean squared errors (MSE) were calculated the analysis including all simulations and calculations were done using statistical package S-Plus 2000 software. The results of this paper show that, the performances of RR are most efficient when the number of regressors is small, while the PLS is most efficient than others when the number of regressors is moderate and high.

Keywards: Multicollinearity; Ridge Regression; Principal Component Regression; Partial Least Squares Regression .

1. Introduction:

In the applications of regression analysis, multicollinearity is a problem that always occur when two or more predictor variables are correlated with each other. This problem makes the estimated regression coefficients by least squares method to be conditional upon the correlated predictor variables in the model. Multicollinearity is a condition in a set of regression data that have two or more regressors which are redundant and have the same information. Redundant information means, what one variable explains about Y is exactly what the other variable explains. In this case, the two or more redundant predictor variables would be completely unreliable since the β_j would measure the same effect of those x_i and the same goes for the other β. Furthermore, $(X'X)^{-1}$ would not exist because the denominator, $(1-r^2)$ is zero. As a result, the estimates of β cannot be found since the elements of the inverse matrix and coefficients become quite large.

The presence of multicollinearity in Least squares regression can cause larger variances of parameter estimates which means that the estimates of the parameters tend to be less precise. As a result, the model will have insignificant test and wide confidence interval. Therefore, the more the multicollinearity, the less interpretable are the parameters. In addition, the problem of multicollinearity in regression analysis can have effects on least squares estimated regression coefficients, computational accuracy, estimated standard deviation of least squares estimated regression coefficients, t-test, extra sum of squares, fitted values and predictions, and coefficients of partial determination.

There are a variety of methods that have been developed for detecting the presence of serious multicollinearity. One of the most commonly used method is the variance inflation factor that measures how much the variances of the estimated regression coefficients are inflated compared to when the independent variables are not linearly related.

The paper is organized as follows. Section (2) presents the three methods for handling multicollinearity. In section (3), we carry out the experimental study that compares the efficiency of considered methods. Section (4) concludes.

2. Methods of Multicollinearity:

Various estimation methods have been developed to overcome the multicollinearity problem, such as ridge regression (RR), principal component regression (PCR) and partial least squares regression (PLS).

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

$$y = I\beta_o + X\beta + \varepsilon \quad , \qquad (1)$$

where, *y* is a ($n \times 1$) vector of observations on the dependent variable, β_o is an unknown constant, X is a ($n \times p$)matrix consisting of n observations on p variables, β is a ($p \times 1$)vector of unknown regression coefficients, and $\in$ 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 *X* and the vector y are mean centered, equation (1) can be simplified as follows:

$$y = X\beta + \in \qquad (2)$$

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

$$Y = XB + E \quad , \qquad (3)$$

where, *Y* is a ($n \times q$) matrix of observations on q dependent variables $y_1, y_2, \ldots, y_q$, E is a ($n \times q$) matrix of errors, whose rows are independently and identically distributed, and B is a ($p \times q$) matrix of parameters to be estimated when the matrix *X* has a full rank of p, the ordinary least squares regression estimator $\hat{\beta}_{OLS}$ can be obtained by minimizing the sum of squared residuals,

$$\hat{\in}'\hat{\in} = (y - X\hat{\beta})'(y - X\hat{\beta}) \qquad (4)$$

Hence,

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'Y \quad , \qquad (5)$$

where $\hat{\beta}_{OLS}$ is a ($p \times 1$) vector of estimated parameters. $\hat{\beta}_{ols}$ provides unbiased estimates of the elements of β, 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:

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

where $\hat{B}_{OLS}$ is the least square estimate of B. When the independent variables are highly correlated, $X'X$ is ill-conditioned and the variance of the OLS estimator becomes barge. The problem of multicollinearity makes the estimated OLS coefficients 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 handle multicollinearity. If we quit insisting on unbiasedness, biased methods such as RR, PCR and PLS can be used to overcome multicollinearity.

2.1. Ridge Regression:

Ridge regression is developed by Hoerl and Kennard (1970). When multicollinearity exists, the matrix $X'X$, where *X* consists of the original regressors, becomes nearly singular. Since, var $(\hat{\beta}) = \sigma^2 (X'X)^{-1}$, and the diagonal elements of $(X'X)^{-1}$ become quite large, so, the variance of $\hat{\beta}$ is to be large. This leads to an unstable estimate of β when OLS is used. In RR, a standardized *X* is used and a small constant θ is added to the diagonal elements of $X'X$. The addition of a small positive number θ to the diagonal elements of $X'X$ causes $X'X$ to be non-singular.

Thus,

$$\hat{\beta}_{RR} = (X'X + \theta I)^{-1} X'Y \ , \qquad (7)$$

where I is the ($p \times p$) identity matrix and $X'X$ is the correlation matrix of independent variables. Values of theta lie in the range (0,1). When $\theta = 0$, $\hat{\beta}_{RR} = \hat{\beta}_{OLS}$. 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 $X'X$ should be to maximized efficiency. There are many procedures in the literature for determining the best value. The simplest way is to plot the values of each $\hat{\beta}_{RR}$ versus θ. The smallest value for which each ridge trace plot shows stability in the coefficient is adopted (Mayers, 1990).

2.2. Principal Component Regression:

PCR is one method to deal with the problem of ill-conditional matrices. What has been done basically is to obtain the number of principal components (PCs) Providing the maximum variation of *X* which optimizes the efficiency of the model. PCR is actually a linear regression method in which the response is regressed on the PCs. Consider *X* as mean centered and scaled (Mean-Centering is achieved by subtracting the mean of the variable vector from all the columns of *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 *X* by the root sum of squares of that variable) then :

$$X'X \ J_i = \lambda_i \ J_i \ \ , i = 1, 2, \ldots, p \ \ , \qquad (8)$$

where the λ_i^s are the eigenvalues of the correlation matrix $X'X$ and the J_i^s are the unit-norm eigenvectors of $X'X$. The vector J_i is used to re-express the $X's$ in terms of PCZ's in the form:

$$Z_i = J_{1i} \ x_1 + J_{2i} \ x_2 + \ldots + J_{pi} \ x_p \ \ , \qquad (9)$$

These Z_i^s are orthogonal to each other and called the artificial variables. Assume that the first m PCs optimize the efficiency of the model. Then,

$$Y = Z_m \ \alpha_m + \in \qquad , \qquad (10)$$

where $\alpha_m = (z'_m \ z_m)^{-1} \ z'_m$ y and m is the number of PCs retained in the model. Using α estimates, it is easy to get back to the estimates of β as:

$$\hat{\beta}_{PCR} = v_m \ \alpha_m \qquad , \qquad (11)$$

where v_m is a matrix consisting of the first m unit-norm eigenvectors. PCR gives a biased estimate of the parameters. If all of the PC's , are used instead of using the first m PC's , then $\hat{\beta}_{PCR}$ becomes identical to $\hat{\beta}_{OLS}$ (Hwang and Nettleton , 2000).

2.3. Partial Least Squares Regression :

PLS is a reasonably alternative method developed by Helland (1990) as a method for constructing predictive models when the explanatory variables are many and highly collinear. It may be used with any number of explanatory variables, even for more than the umber of observations. Although PLS is heavily promoted, it is largely unknown to statisticians (Frank and Friedman (1993)).

To regress the *Y* variables with the explanatory variables $x_i, \ldots, x_p$, PLS attempts to find new factors that will play the same role as the $X's$. These new factors often called latent variables or components.

Each component is a linear combination of $x_i,\ldots,x_p$. There are some similarities with the PCR. In both methods, some attempts have been made to fined 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 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 *X* variables, that is useful for reducing the dimensionality of the regression problem by using fewer components than the number of *X* variables (Garthwaite, 1994).

Now we are going to derive the PLS estimators of β and B. The matrix *X* has a bilinear decomposition in the following form:

$$X = t_1 P_1' + t_2 p_2' + \ldots, t_p p_p' = \sum_{i=1}^{p} t_i P_i' = TP' \qquad (12)$$

Here the t_i are linear combinations of *X*, which we will write as X_{ri}. The $p \times 1$ vectors p_i are often called Loadings. Unlike the weights in PCR (i.e. the eigenvectors J_i), the r_i are not orthogonal. The 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 SIMPLS algorithm.

In the first one, this orthogonality is imposed by computing the t_i as linear combination of residual matrices E_i, in other words, as :

$$t_i = E_{i-1} w_i, E_i = X - \sum_{j=1}^{i} t_j p_j', E_o = X, \qquad (13)$$

where the w_i are orthogonal. Then two sets of weight vectors w_i and r_i, i = 1, 2, … , m. In most algorithms for both multivariate and univariate PLS, the first step is to derive either w_i or r_i, i = 1, 2, … , m, in order to be able to calculate the linear combination of the t_i. Then p_i are calculated by regressing *X* onto t_i. When *m* factors are to be taken into consideration, the following relationship can be obtained :

$$T_m = X R_m \qquad (14)$$

$$P_m = X' T_m (T'_m T_m)^{-1} \qquad (15)$$

$$R_m = W_m (P'_m w_m)^{-1}, \qquad (16)$$

where the first *m* dominant factors, which capture most of the variance in *X*, have maximum ability for the efficiency. Equation (16) connects two sets of weight vectors by a linear transformation. From equations (14) and (15), $P'_m R_m$ equals I_m, since such a transformation exists. Also $R'_m P_m$ equals I_m as follows:

$$R'_m P_m = R'_m X' T_m (T'_m T_m)^{-1} = T'_m T_m (T'_m T_m)^{-1} = I_m \qquad (17)$$

After *m* dimensions have been extracted, the vector of fitted values from PLS can be represented by the first *m* PLS linear combinations T_m. Thus the following equation is obtained:

$$\hat{Y}^m_{PLS} = T_m (T'_m T_m)^{-1} T'_m y \qquad (18)$$

Notice that this is the derivation only for the univariate case.

The multivariate case is identical to the univariate case except that the vector $\hat{y}^m_{PLS}$ should be replaced by the matrix ` $\hat{y}^m_{PLS}$ (Huber et al. 2005). Substituting XR_m for T_m and $\hat{\beta}_{OLS}$ for *Y* results in:

$$\hat{y}^m_{PLS} = XR_m\,(R'_m\,X'\,XR_m)^{-1}\,R'_m\,X'\,X\hat{\beta}_{OLS} \qquad (19)$$

Then it is clear that:

$$\hat{\beta}^m_{PLS} = R_m\,(R'_m\,X'\,XR_m)^{-1}\,R'_m\,X'\,X\hat{\beta}^m_{OLS} \qquad (20)$$

Some what a simpler form for $\hat{\beta}_{OLS}$ can be obtained by first substituting equation (14) into (15), which yields $P_m = X'XR_m(R'_mX'XR_m)^{-1}$. Then using this result in equation (20) gives:

$$\hat{\beta}^m_{PLS} = R_mP'_m\hat{\beta}_{OLS} = w_m\,(P'_m\,w_m)^{-1}\,P'_m\hat{\beta}_{OLS} \qquad (21)$$

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

$$\hat{B}^m_{PLS} = w_m\,(P'_m\,w_m)^{-1}\,P'_m\hat{B}_{OLS}$$

3. Simulation Study:

In this section, we will compare the efficiency of the above three methods, RR, PCR and PLS by performing a simulation study on simulated data sets. We emphasis on the parameter estimation not on the predictive performance of the methods. The experiments consider the following regression model :

$$y_i = \beta_o + \sum_{j=1}^{P} \beta_j\,x_{ij} + e_i \qquad (22)$$

3.1 Comparing the performances :

The efficiency of the considered methods is evaluated by means of the mean squared errors (MSE) of the estimated regression parameters $\hat{\beta}$, which are defined by :

$$MSE(\hat{\beta}) = \frac{1}{m}\sum_{L=1}^{m} \left\| \hat{\beta}^{(L)} - \beta \right\|^2 , \qquad (23)$$

where $\hat{\beta}^{(L)}$ denotes the parameter estimated in the L-th simulation. The MSE indicates to what extent the slope and intercept are correctly estimated. Therefore, the main objective in this study is to obtain an MSE value close to zero.

3.2 Simulation Settings:

The simulated data used in this study consist of $P = 2,4,6$ and 25 predictors variables for $n = 20,30,40,50$ and *60*. The goal is to develop a linear equation that relates all the predictor variables to a response variable. For the purpose of comparing the three methods for multicollinearity, the analysis was done using S-plus software.

The data were constructed as follows:

$$x_1 = N(0,1)$$
$$x_{p-1} = N(0,1) + x_1$$

$$Y = x_1 + ... + x_p + N(0,1), \quad (24)$$

where $P = 2,4,6$ and 25 that represent low $(p = 2)$, medium $(P = 4,6)$ and high number of predictor variables $(p = 25)$. For each simulation, $m = 100$ data sets were generated.

3.3 Simulation Results:

To determine whether multicollinearity exists or not, variance inflation factor (VIF) for each predictor for all cases are computed. VIF is the measure of the speed with which variances and covariances increase and it is the most commonly used method for detecting multicollinearity problem. VIF is a measure of multicollinearity in a regression design matrix (that is, independent variables) in a scaled version of the multiple correlation coefficient between the independent variable, and the rest of the dependent variables. The measure shows the number of times that the variances of the corresponding parameter estimate is increased due to multicollinearity as compared to as what it would be if there were no multicollinearity. Therefore, this diagonostic is designed to indicate the strength of the linear dependencies and how much the variances of each regression coefficient is inflated. The formula of VIF is:

$$(VIF)_j = \frac{1}{1 - R_j^2}, \quad (25)$$

where R_j^2 is the multiple correlation coefficient and measures the coefficient of correlation between two variables with $-1 < R_j < 1$.

There is no formal cut off value to use with the VIF for determining the presence of multicollinearity but Neter et al. (1990) recommended looking at the largest VIF value. A value greater than 10 is often used as an indication of potential multicollinearity problem. The VIF values for each predictor for all given cases in this study are greater than 50. This shows that all the regression coefficients $\beta_1,...,\beta_p$ appear to be affected by collinearity. The efficiency test of the considered methods is evaluated by means of the estimated regression parameters $\hat{\beta}$. These values indicates to what extent the slope and intercept are correctly estimated. According to the value of MSE that is close to zero, the slope and intercept are correctly estimated. The results of the simulations are listed in Tables 1-4.

Table (1) : The efficiency tests for low-number of regressions data sets, $P = 2$

N	20	30	40	50	60
RR	4.52	4.06	2.25	1.36	0.75
PCR	15.34	8.68	5.43	4.55	2.23
PLS	15.28	8.64	5.40	4.52	2.22

Table (2) : The efficiency tests for medium -number of regressions data sets, $P = 4$

N	20	30	40	50	60
RR	24.07	13.47	9.31	6.27	3.51
PCR	25.15	17.46	13.12	8.89	4.86
PLS	2.63	1.39	0.99	0.49	0.13

Table (3) : The efficiency tests for medium -number of regressions data sets, $P = 6$

N	20	30	40	50	60
RR	53.36	26.61	21.92	13.13	6.37
PCR	65.71	28.91	21.59	13.69	7.13
PLS	5.36	2.11	0.68	0.37	0.13

Table (4): The efficiency tests for high -number of regressions data sets, $P = 25$

N	40	50	60
RR	13.14	9.20	6.37
PCR	13.69	10.32	7.13
PLS	0.37	0.17	0.13

From the results of Table (1) where $P = 2$ and the specified n observations, ridge regression performed best compared to the other two methods which gives $MSE = 4.52$ for $n = 20$ and $MSE = 0.75$ for $n = 60$, followed by PLS regression with $MSE = 15.28$ for $n = 20$ and $MSE = 2.22$ for $n = 60$ and PC regression with $MSE = 15.43$ for $n = 20$ and $MSE = 2.23$ for $n = 60$, respectively. The ridge regression method is considered the best since it has the lowest MSE values for all specified n observations and the differences from the other two methods are quite big. On the other hand, there is a slight difference in the MSE for PLS and PC regressions which are chosen at the results are consistent for each n specified cases. The results also show that, for a low number of regressors as $p = 2$, MSE decreases as the number of observations increases.

From the results of Table (2), the PLS regression performed better than RR and PC regressions when $p = 4$ for all the specified n observations. The MSE values for PLS regression differ a lot from the other two methods, where $MSE = 2.63$ for $n = 20$ and $MSE = 0.13$ for $n = 60$, $MSE = 25.15$ for $n = 20$ and $MSE = 4.86$ for $n = 60$ respectively. The optimal number of components for PLS and PC regressions should be chosen at the smallest value of MSE. These show that both methods performed well with optimal number of components in handling multicollinearity for $p = 4$ regressors.

The results of Table (3) show that, the PLS regression performed best followed by RR and PC methods when $p = 6$ for all specified n observations. The MSE values for PLS are 5.36 for $n = 20$ and 0.13 for $n = 60$, while for RR, the MSE values for the same n are 53.36 and 6.37, respectively, and for PC, the MSE values are 65.71 and 7.13. The results also show that MSE values decrease as n increases from 20 to 60. This shows that, as the number of observations becomes higher, the MSE values become smaller compared to a small number observations. The results are also consistent where PLS performed better than RR followed by PC regression for every specified number of observations.

The results of Table (4) show that, the RR method performs best followed by PLS and PC respectively when the number of regressors is high, $p = 25$. RR gives MSE values 13.14, 9.20 and 6.37 for $n = 40$, 50 and 60, respectively, followed by PLS which gives MSE values of 0.37, 0.17 and 0.13 , and PC which gives MSE values of 13.69, 10.32 and 7.13.

4. Conclusions:

In this paper, the number of dimensions is less than the number of observations. The Numerical results show that, RR and PLS methods are generally effective in handling multicollinearity problem in the specified observations with $p = 2,4,6$ (for Low and moderate number of regressors) and 25 (for high number of regressors). The performances of RR are most efficient than others when $p = 2$, while PLS is most efficient when $p = 4,6$ and 25.

The results also show that, both PLS and RR performed better than PCR in all the cases. However, the differences between the PCR performance from PLSR and RR are only slight. These confirmed result that there is no one method that dominates the other, and that the difference between the methods is typically small when the number of observations is large.

In all cases, it is obvious that, the superior method performed well when the number of observations, n are larger than the number of regressors. It also shows that, the results are consistent for every specified number of observations , n that were included in the analysis. Generally, RR is approximate effective and efficient for a small p, while PLS is efficient for a large number of regressors p.

References

Abdi, H. (2010). "Partial least squares regression and projection on latent structure regression (PLS-Regression)". *Wiley Interdisciplinary Reviews: Computational Statistics* **2**: 97–106

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

Garthwaite, P.H. (1994) "An interpretation of partial least squares" Journal of the American statistical Association, 89, 122 – 127.

Gunst, R.F. and R.L. Mason(1979)" Some considerations in the evaluation of alternate prediction equations " Technometrics, 21, 55 – 63.

Helland, Inge S. (1990) "PLS regression and statistical models". *Scandinavian Journal of Statistics* **17**, 2, 97–114

Hoerl, A. E. and R.W. Kennard (1970) " Ridge regression biased estimation for nonorthogonal problems " Technometrics 8, 27- 51.

Huber, M., Rousseeuw, P.J. and K. Vanden Branden (2005) ' A New approach to robust principal components analysis " Technometerics, 47, 64 – 79.

Hwang, J. T. and D. Nettleton (2000) " Principle Commponents regression with data chosen components and related methods " Technometrics , 45, 70 – 79.

Joliffe, I.T. (1986) " Principal Commponent Analysis " Springer, New York.

Mayers, R. H. (1990) " Classical and modern regression with applications" 2nd edition, Duxbury press.

Michael H. Kutner, John Neter, Christopher J. Nachtsheim, William Li, (1990) "*Applied Linear Statistical Models*", McGraw-Hill College.

Tenenhaus M. (2008)" Structural Equation Modelling for small samples "Working paper No. 885. HEC Paris, Jouy-en-Josas.

Tenenhaus M., Esposito Vinzi V., Chatelin Y.-M., Lauro C. (2005)" PLS path modeling" Computational Statistics & Data Analysis, 48, 159-205