A Comparative Study of Some Estimation Methods for Multicollinear Data

Okeke Evelyn Nkiruka, Okeke Joseph Uchenna

***Abstract*— This article compares different estimation methods specially designed to combat the problem induced by multicollinearity using real life data of different specifications and distributions. From the mean squared error of the samples studied we observed that Partial least square came up as the best estimator among the methods we studied. Stepwise regression performs better when the predictor variables are highly correlated. Under the ridge regression study the smallest eigenvalue of the predictor variables of the original data was used in determining the ridge parameter of ridge regression since the variances of some of our samples cannot be estimated by ordinary least squares regression. From our results we found that among all the methods we studied PLSR estimator stands the "best", followed by the stepwise regression and then the PCR estimator in predicting the response variable. We are not surprise that RR estimator stands the least among the methods since it is known as biasing estimator and more useful in estimating the parameters of the model. We also wish to state that PLSR is efficient in prediction when the sample size is very small.**

***Index Terms*— Multicollinearity, Principal component regression, Eigen value, Partial least squares, Ridged regression, Nonorthogonal data and Stepwise regression.**

I. INTRODUCTION

The term multicollinearity is used to denote the existence of a perfect or exact, linear relationships (or near perfect relationships) among some or all explanatory variables of regression model [1]. If the explanatory variables are perfectly correlated, that is, if the correlation coefficient for these variables is equal to unity, the parameters become indeterminate: it is impossible to obtain numerical values for each parameter separately and the method of least squares breaks down. Multicollinearity may also be induced by the choice of model, for instance, the addition of polynomial terms to a regression model may cause ill-conditioning in $X'X$. Furthermore if the range of X is small, adding an X^2 term can result in severe multicollinearity and also if the number of explanatory variables are more than the sample size LS method may produce misleading result.

Several techniques have been proposed for dealing with the problems caused by multicollinearity. The general approach include the collection of additional information, model re-specification and the use of estimation methods specially designed to combat the problem induced by multicollinearity.

Okeke, Evelyn Nkiruka, Department of Mathematics & Statistics , Federal University Wukari, Wukari, Nigeria, +2348181278549.

Okeke, Joseph Uchenna, Department of Mathematics & Statistics, Federal University Wukari, Wukari, Nigeria, +2348036026806.

Collecting additional information is not always possible because of economic constraint or because the process being studied is no longer available for sampling. Even when the additional data are available, it may be inappropriate to use if the new data extends the range of interest. Furthermore, if the new data points are unusual or atypical of the process being studied, their presence in the sample could be highly influential on the fitted model. Finally, it is good to note that the addition of more data is not a valid solution to the problem of multicollinearity especially when the multicollinearity is due to constraint on the model or in the population.

Some re specification of the regression equation may lessen the impact of multicollinearity especially when it is caused by the choice of model. Model respecification done by either redefining the regressors or by variable elimination may not provide a satisfactory solution if the new model does not preserve the information contained in the original data and or if the regressors dropped from the model have significant explanatory power relative to the response variable.

With the impending dangers of the two scenarios discussed, this paper aimed at discussing and comparing different estimation methods designed to solving the problems of multicollinearity. The methods include Principal component regression, Partial least squares, ridged regression, and stepwise regression. Three different types of multicollinear data (when the sample size is smaller than or equal to the number of the predictor variables, where the predictor variables are highly correlated, and when the polynomials terms are added to the model) were studied with intention of finding the best method for each data type.

II. ESTIMATION METHODS

A. Partial Least Squares

Partial least squares (PLS) is a method for constructing predictive models when the factors are many and highly collinear [2]. Emphasis of PLS is on predicting the responses and not necessarily on trying to study the underlying relationship between the variables. For example, PLS is not usually appropriate for screening out factors that have a negligible effect on the response. However, when prediction is the goal and there is no practical need to limit the number of measured factors, PLS can be useful tool.

PLS can be applied in monitoring industrial processes; a large process can easily have hundreds of controlling variables and dozens of outputs.

Multiple linear regression can be used with very many factors. However, if the number of factors gets too large (for example, greater than the number of observations), you are likely to get a model that fits the sampled data perfectly but that will fail to predict new data well. This phenomenon is called over-fitting. In such cases, although there are many manifest factors, there

may be only a few underlying or latent factors that account for most of the variation in the response. The general idea of PLS is to try to extract these latent factors, accounting for so much of the manifest factor variation as possible while modeling the responses well

The aim of partial least squares is to predict the response by a model that is based on linear transformation of the explanatory variables. Partial least squares (PLS) is a method of constructing regression models of type

$$\bar{Y} = \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \cdots + \beta_p T_p \qquad 1$$

Where the T_i are linear combination of the explanatory variables $X_1, X_2, \ldots X_k$ such that the sample correlation for any pair T_i, T_j (i, j) is 0. Following the procedures given in [3], all the data are first centered. Let $\bar{y}, \bar{x}_1, \ldots, \bar{x}_k$ denote the sample means of the $T \times (K+1)$ -data matrix $(y, X) = (y_1, x_1, \ldots, x_k)$ and denote the variables

$$U_1 = Y - \bar{y}_i \qquad 2$$

$$V_{i1} = X_i - \bar{x}_i \quad (for\ i = 1, \ldots, k) \qquad 3$$

then the data values are the T-vectors

$$u_1 = y - \bar{y}_1, \qquad (\bar{u}_1 = 0) \qquad 4$$

$$v_{i1} = x_i - \bar{x}_{i1}, \qquad (\bar{v}_{1i} = 0) \qquad 5$$

The linear combination T_j called factors, latent variables, or components, are then determined sequentially. The procedure is as follows:

i. U_1 is first regressed against V_{11}, then regressed against V_{12},…, then regressed against V_{1k}. Then univariate regression equations are

$$\hat{U}_{1i} = b_{1i} V_{1i} \quad (i = 1, \ldots, k), \qquad *6$$

where $b_{1i} = \frac{v'_{1i} u_1}{v'_{1i} v_{1i}}$

Then each of the k equations in *(6) provides an estimate of U_1. To have one resulting estimate, one may use a simple average $\sum_{i=1}^{k} w_1 b_{1i} V_{1i} / k$ or the weighted average like

$$T_1 = \sum_{i=1}^{K} w_1 b_{1i} V_{1i} \qquad 7$$

with the data value

$$t_1 = \sum_{i=1}^{K} w_1 b_{1i} v_{1i} \qquad 8$$

ii. The variable T_1 should be a useful predictor of U_1 and hence of Y. The information in the variable X_i that is not in T_1 may be estimated by the residuals from a regression of X_i on T_1 which are identical to the residuals, say V_{2i}, if V_{1i} is regressed on T_1, that is

$$V_{2i} = V_{1i} - \frac{t'_1 v_{1i}}{t'_1 t_1} T_1 \qquad 9$$

To estimate the amount of variability in Y that is not explained by the predictor T_1, one may regress U_1 on T_1 and take the residuals, say U_2.

iii. Define now the individual predictors

$$\hat{U}_{2i} = b_{2i} V_{2i} \qquad (i = 1, \ldots, K) \qquad 10$$

where

$$b_{2i} = \frac{v'_{2i} u_2}{v'_{2i} v_{2i}}$$

and the weighted average

$$T_2 = \sum_{i=1}^{K} w_{2i} b_{2i} V_{2i} \qquad 11$$

iv. General iteration step

Having performed this algorithm k times, the remaining residual variability in Y is U_{k+1} and the residual information in X_i is $V_{(k+1)i}$, where

$$U_{k+1} = U_k - \frac{t'_k u_k}{t'_k t_k} T_k \qquad 12$$

and

$$V_{(k+1)i} = V_{ki} - \frac{t'_k v_{ki}}{t'_k t_k} T_k. \qquad 13$$

Regressing U_{k+1} against $V_{(k+1)i}$ for $I = 1, \ldots, K$ gives the individual predictors

$$\hat{U}_{(k+1)i} = b_{(k+1)i} V_{(k+1)i} \qquad 14$$

with

$$b_{(k+1)i} = \frac{v'_{(k+1)i} u_{k+1}}{v'_{(k+1)i} v_{(k+1)i}}$$

and the (k+1)th component

$$T_{k+1} = \sum_{i=1}^{K} w_{(k+1)i} b_{(k+1)i} V_{(k+1)i} \qquad 15$$

v.

Suppose that this process has stopped in the pth step, resulting in the PLS regression model given in (1). The parameters $\beta_0, \beta_1, \ldots, \beta_p$ are estimated by univariate OLS. This can be proved as follows.

In matrix notation we may define

$$V_{(k)} = (V_{k1}, \ldots, V_{kK}) \quad (k = 1, \ldots, p), \qquad 16$$

$$\hat{U}_{(k)} = (b_{k1} V_{k1}, \ldots, b_{kK} V_{kK}) \quad (k = 1, \ldots p) \qquad 17$$

$$w_{(k)} = (w_{k1}, \ldots, w_{kK})' \quad (k = 1, \ldots, p) \qquad 18$$

$$T_{(k)} = \hat{U}_{(k)} w_{(k)} \quad (k = 1, \ldots, p) \qquad 19$$

$$V_{(k)} = V_{(k-1)} - \frac{v'_{(k-1)} t_{(k-1)}}{t'_{k-1} t_{k-1}} T_{k-1} \qquad 20$$

By construction the sample residual $v_{(k+1)i}$ are orthogonal to $v_{ki}, v_{(k-1)i}, \ldots, v_{1i}$, implying that $v'_{(k)} v_{(j)} = 0$ for $k \neq j$, hence, $\hat{u}'_{(k)} \hat{u}_{(j)} = 0$ for $k \neq j$, and finally'

$t'_k t_j = 0 \quad k \neq j$

The well know feature of the PLS is that the sample components t_i are pairwise uncorrelated. The simple consequence is that parameters β_k may be estimated by simple univariate regression of Y against T_k. Furthermore, the preceding estimates $\hat{\beta}_k$ stay unchanged if a new component is added.

B. Principal Component Regressio

Principal component regression is a regression procedure used in the presence of multicollinearity among the k

dimensional random vector of the predictor variables X. That is, where the matrix of X is not of full rank ($rank\,(X) < k$) or when the number of predictor variables are more than the sample size.

The model of principal component regression is of the form

$$y = XPP'\beta + e \qquad 21$$

This can as well be written as

$$y = \breve{X}\breve{\beta} + e \qquad 22$$

where $\breve{X} = XP$, and $\breve{\beta} = P'\beta$

Let the column of the orthogonal matrix $P = (p_1, \dots, p_k)$ of the eigenvectors of $X'X$ be numbered according to the magnitude of the eigenvalues $\lambda_1 \geq \lambda_2 \geq, \dots, \lambda_k$. Then the $\breve{x}_i = Xp_i$ is the ith principal component and we get

$$\breve{x}'\breve{x} = p_i X'X p_i = \lambda \qquad 23$$

We now assume exact multicollinearity. Hence $rank\,(X) = k - j$ with $j \geq 1$. We get

$$\lambda_{K-j+1} = \cdots = \lambda_k = 0 \qquad 24$$

According to the subdivision of the eigenvalues into the groups $\lambda_1 \geq \cdots \geq \lambda_{K-j} > 0$ and $\lambda_{K-j+1} = \cdots = \lambda_k = 0$, we define the subdivision

$$P = (P_1, \quad P_2) \quad \Lambda = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \breve{X} = (\breve{X}_1, \breve{X}_2) = (XP_1, XP_2)$$

$$\breve{\beta} = \begin{pmatrix} \breve{\beta}_1 \\ \breve{\beta}_2 \end{pmatrix} = \begin{pmatrix} P_1'\beta \\ P_2'\beta \end{pmatrix}$$

with $\breve{X}_2 = 0$ as in (24). We now define

$$y = \breve{X}_1\breve{\beta}_1 + \breve{X}_2\breve{\beta}_2 + e$$
$$= \breve{X}_1\breve{\beta}_1 + e$$

The OLS estimate f the $(K - j)$ -vector $\breve{\beta}_1$ is $b_1 = (\breve{X}_1'\breve{X}_1)^{-1}\breve{X}_1'y$. The OLS estimate of the full vector $\breve{\beta}$ is

$$\begin{pmatrix} b_1 \\ 0 \end{pmatrix} = (X'X)^{-1}X_1'y$$

.

$$= (P\Lambda^{-1}P)X'y \qquad 25$$

with

$$\Lambda^{-1} = \begin{pmatrix} \Lambda_1^{-1} & 0 \\ 0 & 0 \end{pmatrix}$$

being a generalized inverse of Λ

C. *Ridge Regression*

When the method of least squares is applied to nonorthogonal data, very poor estimate of the regression coefficients are usually obtained, the variance of the Least square (LS) estimates of the regression coefficient may be considerably inflated, and the length of the vector of least squares parameter estimates is too long on the average [4]. This implies that the absolute value of the least squares estimates are too large and that they are very unstable, that their magnitude and signs may change considerably given a different sample,

The problem with the method of LS is the requirement that the estimator of β should be unbiased. The Gauss-Markov property of regression parameter assures us that the LS estimator $\hat{\beta}$ of β has minimum variance in the class of unbiased linear estimators without guarantee that the variance will be small. If the variance of $\hat{\beta}$ is large, it implies that confidence interval on β would be wide and the point estimate $\hat{\beta}$ is very unstable.

One way to alleviate this problem is to drop the requirement that the estimator of β be unbiased. [5] and [6] proposed a biased estimator (ridge estimator) of β,

$$\tilde{\beta} = (X'X + kI)^{-1}X'y \qquad 26$$

that has a smaller variance than the unbiased estimator $\hat{\beta}$. This ridge estimator is a linear transformation of the LS estimator since

$$\tilde{\beta} = (X'X + kI)^{-1}X'y$$
$$= (X'X + kI)^{-1}(X'X)\hat{\beta}$$
$$= Z_k\hat{\beta} \qquad 27$$

$k \leq \frac{2\sigma^2}{\hat{\beta}\hat{\beta}'}$ where $\hat{\beta}$ and σ^2 are found from the least square solution.

[7] stated that both the mean squared error and the smallest eigenvalue of the predictor variables of the original data play vital role in determining the biased parameter (k) of ridge regression. [9] showed through simulation that the resulting ridge estimator had significant improvement in mean squares error (MSE) over LS.

The mean square error of the estimator $\tilde{\beta}$ is defined as

$$MSE(\tilde{\beta}) = E(\tilde{\beta} - \beta)^2$$
$$= V(\tilde{\beta}) + [E(\tilde{\beta}) - \beta]^2$$
$$MSE(\tilde{\beta}) = var(\tilde{\beta}) + (bias\ in\ \tilde{\beta})^2 \qquad 28$$

Note that the MSE is just the expected squared distance from $\tilde{\beta}$ to β. By allowing a small amount of bias in $\tilde{\beta}$, the variance of $\tilde{\beta}$ can be made small such that the MSE of $\tilde{\beta}$ is less than the variance of the unbiased estimator $\hat{\beta}$. Consequently confidence interval on β would be much narrower using the biased estimator. The small variance for the biased estimator also implies the $\tilde{\beta}$ is a more stable estimator of β than the unbiased estimator $\hat{\beta}$.

D. *Stepwise Regression*

In deciding on the "best" set of explanatory variable for a regression model, researchers often follow the method of stepwise regression. In this method the ordinary least square (OLS) regression of the variables are performed by introducing the X variables one at a time (stepwise forward regression) or by including all the possible X variables in one multiple regression and rejecting them one at a time (stepwise backward regression).The decision to add or drop a variable is usually made on the basis of the contribution of the variable to the error sum of squares (ESS) of the F test.

III. THE DATA SETS AND THEIR RESULTS

To compare the performance of the methods that we have considered seven different real data sets were studied to investigate their effectiveness at predicting response variable using their mean squares error (MSE). Attempt was also made to see how a biased regression estimator (ridge regression) competes with other estimators we studied. The data sets studied includes: a data sets that contains predictor variable that are highly correlated, this data set is from Nigeria Stock Exchange and is based on their transaction for the period of 1991-2007. The data is available at [8]; a data set from chemometric study where the number of predictors are far more than the sample size; and a data set with polynomials of

different degrees of predictor variable. A case of time series data was also considered where time is included as a predictor variable. This data is obtained from [1]. Three different data sets with very small sample sizes at varying number of predictor variables were studied in an effort to find the best estimator that is suitable when dealing with small sample size problem. One of the data is on CO_2 emission and its possible correlates of four countries, while the other two are sampled data from some data we found in [1]. The mean squares error of each of the estimators described in section two was computed for all the seven data sets. Partial least square regression offered an almost imperceptible improvement over other estimators. The specified results for the estimators are not reported here to save space and to focus more on the main objective of this
study. The MSE estimates for all the methods we considered are provided in Table I. For each of the seven data sets, the OLS regression estimators along with the other estimators discussed in section II were ranked according to their MSE values with the lowest ranks corresponding to lowest MSE. The median of the ranks for each method is given in Table II.

In this study we made effort to finding the best regression estimator in predicting the response variable. Four useful estimators for treating multicollinearityr were compared together with OLS in each data set. Sample correlations between pair of predictor variables of data set 2 in column 3 of Table I range from 0.421 to 1.000. For PCR estimator, the last eigenvalue of $X'X$ was used in determining the biasing parameter of ridge regression since we could not run OLSR of some of our data sets due to small nature of our sample size. Here efforts were made to select the principal components that account for 90% and above of the total variation in the original data. In most cases not all the components are considered in PLSR estimator. Only those components that provide the desired results were considered.

IV. Discussion

The OLSR and ridge regression estimators performed quite poorly in all the data sets we studied. This comes as no surprise because the selected data sets were chosen to study the behavior of the estimators when OLSR estimation is expected to be deficient. Also since the measure of error is used for comparison it is expected that ridge regression estimator as a biasing estimator will not be effective in predicting Y. In four of the seven data sets studied OLSR estimator produces no result. We are not surprise that the rank of OLSR is 1 in the first data set in column 2 of Table I because it has been said in the literature that if the sole purpose of regression analysis is prediction or forecasting, then multicollinearity is not a serious problem because the higher the R^2,the better the prediction.

We observed that method D, stepwise regression estimator (which is often used in deciding the "best" set of predictor variables for a regression model) produced fairly good result with median rank of 2 as can be seen in table II. This method is second to the best methods we studied.

The method of PLSR is the "best" among all the methods we studies with a median rank of 1. This method came first in five of the data sets we studied and took second and third position in the remaining two data sets.

PCR estimator, method B is next to stepwise regressing in predicting the response variable of regression analysis with the median rank of 3 as can be seen in Table II.

Among the weakest estimator in predicting Y variable we studied is RR estimator. This method took the last position among other methods we studied as can be seen in Table II. It is good to note that one may get different result when other

methods (e.g. variance of the distribution) are used in determining the biasing parameter of ridge regression. From Table I, it is clear that this method can produce misleading result when the sample size is very small and less than or equal to the number of the predictor variables. See tables below.

Table I: MSE of Different Estimators across Different types of Multicollinear Data

Estimator	Data with time as predictor variable	Highly correlated data	Data with size n=15; p=25	Data with Different polynomials of X	Data with size n=p= 6	Data with size n=4; p=6	Data with size n=5; p=6
OLSR	0.475	103085	-	0.054	-	-	-
Stepwise	0.57	102727	0.004	0.53	0.004	15620	0.33
PLSR	0.474	103088	0.002	0.033	0	0	0.001
PC	1.059	104150	0.015	2.72	0.431	22415	0.579
RR	8.774	294810	0.062	0.075	1nfinity	-1.665×10^{11}-82650	

Table II: Estimators and their Ranks

Estimators	Ranks	Median rank
OLSR	2,2,2	-
Stepwise	1,2,2,2,3,4	2
PLSR	1,1,1,1,1,2,3	1
PC	3,3,3,4,5	3
RR	3,4,5,5	4.5

Acknowledgment

We extend our sincere appreciation to everyone that contributed by way of questions or suggestions towards this research and most especially to all the authors, whose works are cited in the references, for their comments and contributions in the subject matter that guided our study. All your contributions are gratefully acknowledged.

References

[1] D. N. Gujarati, "Basic Econometrics," 4th ed. New York :Tata McGraw-Hill Companies Inc, 2003, pp. 371-378.
[2] R. D. Tobias. (1995). An introduction to partial least squares regression. SUGI Proceedings, SAS Institute Inc.
[3] P. H. Garthwaite. (1994). An interpretation of partial least square. *Journal of American Statistical Association*. 89, pp. 122-127.
[4] L. Tauer. (2001). Efficiency and competitiveness of the small New York dairyfarm. *Journal of Dairy Science*. 84, 2573-2576.
[5] A. E. Hoerl and R.W. Kennard. (1970a). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*. 12, pp. 59-84.

[6] A. E. Hoerl and R. W. Kennard. (1970b). Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12, pp. 69-82.

[7] E. N. Okeke and J. U. Okeke. (2015). Bootstrapping in determination of ridge parameter. *International Journal of Applied Mathematics and Statistical Sciences*. 4(6), 15-34

[8] I. A. Iteire, "Transaction of Nigeria Stock Exchange from 1991-2007," B.SC Project, Nnamdi Azikiwe University, Awka, Nigeria, 33-35. 2004 (unpublished).

[9] *E. Hoerl, R. W. Kennard and K. F. Baldwin.(1975) Ridge regression: some simulation. Communications in Statistics, 4, pp. 105–123*

Okeke, Evelyn Nkiruka was born in Obeledu, Anambra State, Nigeria, on the 16th July, 1971. She holds: PhD Statistics (2011) from ABSU, Abia, Nigeria; M.Sc. Statistics (2002) from NAU, Anambra, Nigeria and B.Sc. Statistics (1997) from NAU, Anambra , Nigeria. Her major field of study is Multivariate statistics(Discrimination and classification) and also has interest in Econometric modeling.

She was a LECTURER at the Nnamdi Azikiwe University (NAU, Awka) 2001-2013. Presently, she lectures in the Department of Mathematics and Statistics of the Federal University Wukari, Taraba State, Nigeria.. She has published in both local and foreign reputable journals. Her research interests is in the area of Discriminant Analysis.

Dr.Mrs. Okeke is a member of the Nigerian Statistical Association, a consultant with the United Nations Development Program, the Head, Department of Mathematics and Statistics, Federal University Wukari, Nigeria. Chairperson Welfare committee, Faculty of Pure and Applied Sciences, Federal University Wukari, Nigeria.

Okeke, Joseph Uchenna was born in Asaba, Delta State, Nigeria , on the 14th May, 1971. He holds: PhD Statistics (2011) from ABSU, Abia, Nigeria; M.Sc. Statistics (2005) from NAU, Anambra, Nigeria and B.Sc. Statistics (1997). His major field of study is Econometric statistics with stint in multivariate statistics which was his area of research at his masters thesis.

He was a LECTURER at the Anambra State University now Chukwuemeka Odumegwu Ojukwu University (2007-2013). Presently, he lectures in the Department of Mathematics and Statistics of the Federal University Wukari, Taraba State, Nigeria.. He has published in both local and foreign reputable journals. His research interests are in the areas of Econometric dynamic modeling and Multivariate classification modeling.

Dr. Okeke is a member of the Nigerian Statistical Association, a consultant with the United Nations Development Program, the Secretary, Anambra West Elite Club (2012 to date), the seminar coordinator, Faculty of Pure and Applied Sciences, Federal University Wukari, Nigeria