The Comparison Between Different Approaches to Overcome the Multicollinearity Problem in Linear Regression Models

Hazim Mansoor Gorgees
Fatimah Assim Mahdi
Dept. of Mathematics/ College of Pure Science Ibn Al-Haithem/ Baghdad University
hazim5656@yahoo.com
fatima_assim92@yahoo.com

Received in:10 /September/2017, Accepted in:22/ October /2017

Abstract:

In the presence of multi-collinearity problem, the parameter estimation method based on the ordinary least squares procedure is unsatisfactory. In 1970, Hoerl and Kennard insert an alternative method labeled as estimator of ridge regression.

In such estimator, ridge parameter plays an important role in estimation. Various methods were proposed by many statisticians to select the biasing constant (ridge parameter). Another popular method that is used to deal with the multi-collinearity problem is the principal component method. In this paper, we employ the simulation technique to compare the performance of principal component estimator with some types of ordinary ridge regression estimators based on the value of the biasing constant (ridge parameter). The mean square error (MSE) is used as a criterion to assess the performance of such estimators.

Keywords: multi-collinearty, Ridge regression, Ridge parameter, condition number, Principal components.

https://doi.org/10.30526/31.1.1841

Introduction

Consider the linear regression model

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

where

$y$is$(n \times 1)$vector of response variable,

$X$is$(n \times p)$matrix of explanatory variablesand n > p,

$\beta$is$(p \times 1)$vector of unknown parameters,

$\varepsilon$is$(n \times 1)$vector of unobservable random errors and$E(\varepsilon) = 0$, $var(\varepsilon) = \sigma^2 I$

The aim of regression analysis is to estimate the numerical values of linear model parameters. Recently, biased estimators of regression parameters get attention of many researchers, because the ordinary least squares procedure is unable to provide reasonable point estimates when the matrix of explanatory variables if there exists the problem of multi-collineardata. Where we refer through the paper to the ridge regression estimators and principal component estimators as alternatives to the ordinary least square estimators with multi-collineardata. The estimators of each ridge regression and principal component allow a small amount of bias in order to achieve a major reduction in the variance in contrast to ordinary least squares.

The Case of Multi-collinearity

The problem of multi-collinearity occurs when there exists an exact linear relationship or an approximate linear relationship among two or more explanatory variables, two types of multi-collinearity may be faced in regression analysis, exact and near multi-collnearity. During regression calculations, the exact linear relationship causes a division by zero which in turn leads to the abortion of the calculations. In case of not exact relationship, the calculations will not be aborted and the division by zero does not occur. Nevertheless, the results will be distorted when the division is done by a very small quantity. Therefore, the determination whether the multi-collinearity is a problem is one of the first steps in regression-analysis.

Multi-collinearity can be thought of as a situation where two or more explanatory variables in the data set move together, as a consequence it is impossible to use this data set to decide which of the explanatory variables is producing the observed change in the response variable.

Some multi-collinearty is nearly always present, but the important point is whether it is serious enough to cause appreciable damage to the regression analysis. Indicators of multi-collineaity include a low determinant of the information matrix X'X, a very high correlation among two or more explanatory variables, very high correlation among two or more estimated coefficients, a very small

(near zero) eigenvalues of the correlation matrix of the explanatory variables and the too large condition number.

The Class of Shrinkage Estimators

Applying the singular value decomposition technique, we can decompose the matrix X as follows [1]

$$X = H\Lambda^{\frac{1}{2}}G' \qquad (2)$$

where H is ($n \times p$) matrix satisfying H'H =I$_p$, $\Lambda^{\frac{1}{2}}$ is a (p×p) diagonal matrix of ordered singular values of X.

$\lambda_1^{\frac{1}{2}} \geq \lambda_2^{\frac{1}{2}} \geq \ldots \geq \lambda_p^{\frac{1}{2}} > 0$, G is a (p×p)orthogonal matrix whose columns represent the normalized eigenvectors of X'X.

Consequently, the ordinary least squares estimator of the regression parameters vector β can be rewritten as:

https://doi.org/10.30526/31.1.1841

$$b_{OLS} = (X'X)^{-1}X'y$$

$$= (G\Lambda G')^{-1} G\Lambda^{\frac{1}{2}}H'y$$

$$= G\Lambda^{-\frac{1}{2}}H'y = GC$$

where $C = \Lambda^{-\frac{1}{2}}H'y = G'b_{OLS}$ is the vector of uncorrelated components of b_{OLS}.

This can be noticed by considering the variance-covariance matrix of C that can be easily shown to equal the diagonal matrix $\sigma^2\Lambda^{-1}$.

The generalized shrinkage estimators denoted by b_{SH} can be defined as

$$b_{SH} = G\Delta C = \sum_{j=1}^{p} \overrightarrow{g_j}\,\delta_j C_j \qquad (3)$$

Where

$\overrightarrow{g_j}$ is thej^th^column of the matrix G,

δ_j is the j^th^diagonal element of the shrinkage factors diagonal matrix Δ, $0 \le \delta_j \le 1, j = 1, 2, ..., p$ and

C_j is thej^th^element of the uncorrelated components vector C.

Ordinary Ridge Regression Estimators

The most popular method that has been proposed to deal with multi-collinearity problem is the ordinary ridge regression.

The ordinary ridge regression method is a modification of ordinary least squares method to allow the biased estimators of regression coefficients.

The ridge estimators depend crucially upon an exogenous parameter, say k, called the ridge parameter or the biasing parameter of the estimator. For any $k \ge 0$, the corresponding ordinary ridge estimator denoted byb_{RR} is defined as:

$$b_{RR} = (X'X + kI)^{-1}X'y \qquad (4)$$

where$k \ge 0$is a constant selected by the statistician according to some intuitively plausible criteria put forward by Hoerl and Kennard [2].

It can be shown that the ridge regression estimator given in equation (4) is a member of the class of shrinkage estimators as follows

By using matrix algebra and singular value decomposition approach we get

$$b_{RR} = (X'X + kI)^{-1}X'y$$

$$= [G(\Lambda + kI)G']^{-1} G\Lambda^{\frac{1}{2}}Hy$$

$$= G(\Lambda + kI)^{-1}G'G\Lambda^{\frac{1}{2}}H'y$$

$$= G(\Lambda + kI)^{-1}\Lambda^{\frac{1}{2}}H'y$$

$$= G[(\Lambda + kI)^{-1}\Lambda]\Lambda^{-\frac{1}{2}}H'y = G\Delta C \qquad (5)$$

Where$\Delta = (\Lambda + kI)^{-1}\Lambda$.

Equivalently, the shrinkage factors δ_j, j= 1,2,...,p of the ridge estimator has the form

$$\delta_j = \frac{\lambda_j}{\lambda_j + K} \qquad (6)$$

Where λ_j is thej^th^ element (eigenvalue) of the diagonal matrix Λ, and K is the ridge parameter.

The mean square error of ordinary ridge regression estimator can easily demonstrated to be[2]

https://doi.org/10.30526/31.1.1841

$$\mathrm{MSE}(b_{RR}) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + K)^2} + K^2 \beta'(X'X + kI)^{-2} \beta \qquad (7)$$

The first term can be shown to be the sum of variances(total variance) of the parameter estimates and the second term can be considered to be the square of the bias introduced when b_{RR} is used instead of b_{OLS}.

Choice of Ridge Parameter

The ordinary ridge regression estimators do not provide a unique solution to the multi-collinearity problem, but provide a family of solutions. These solutions depend upon the ridge parameter (the value of k). No explicit optimum value can be found for k. Yet, several stochastic choices have been proposed for this ridge parameter. Some of these choices may be summarized as follows

Hoerl and Kennard (1970).Suggested graphical method called ridge trace to select the value of the ridge parameter k. When viewing the ridge trace, the analyst picks the value of k for which the regression coefficients have stabilized.

Often, the regression coefficients will vary widely for small values of k and then stabilize. We have to select the smallest value of k (which introduced the smallest bias) after which the regression coefficients have seemed to remain constant.

Hoerl, Kennard and Baldwin in (1975), proposed another method to select a single value of K given as [3]

$$\hat{k}_{HKB} = \frac{pS^2}{b_{OLS}'b_{OLS}} \qquad (8)$$

Where p is the number of explanatory variables, S^2 is the OLS estimator of σ^2 and b_{OLS} is the OLS estimator of the vector of regression coefficients β.

Lawless and Wang (1976) proposed selecting the value of K by using the formula [4]

$$\hat{k}_{Lw} = \frac{pS^2}{b_{OLS}'X'X\, b_{OLs}} \qquad (9)$$

Assuming that the regression coefficients vector has certain prior distribution srivastava followed Bayesian approach to estimate the ridge parameter. He concluded that [5]

$$\hat{k}_{Bayes} = \mathrm{Max}[0, \frac{\mathrm{tr}(X'X)}{[\frac{n-p-3}{n-p-1}(\frac{b_{OLS}'X'X\, b_{OLS}}{S^2}) - p]} \qquad (10)$$

Where tr (X'X) denote the trace of the matrix X'X.

Hazim Mansoor Gorgees and Fatimh Assim Mahdi (2017) proposed a new method for selecting the ridge parameter by employing the concept of condition number [6].

The suggested estimator denoted as $\hat{k}_{CN}$ is defined as

$$\hat{k}_{CN} = Max\,[0, \frac{\rho S^2}{b_{OLS}\,'b_{OLS}} - \frac{1}{CN}] \qquad (11)$$

Where CN reffered to the condition number which is the ratio of the largest to the smallest singular value of the matrix of explanatory variables X.

Principal Components Regression

Ridge regression was offered as a technique which attempted to overcome the multi-collinearity problem. An alternative procedure known as principal components approach, was first proposed by Harold Hoteling (1933).

In order to obtain a good realization of this approach let us proceed our discussion with the case of two predictors $x_1 and x_2$.If these predictors are correlated then the matrix X will not be orthogonal consequently, this will complicate the interpretation of the effects of X_1 and X_2 on the response variable y.

From the geometric point of view, suppose we rotate the coordinate axis so that in the new system, the predictors are orthogonal. Moreover, let us make the rotation so that the first axis lies in the direction of the greatest variation in the data, the second axis lies in the direction of the second greatest variation in the data.

These rotated directions (Z_1 and Z_2 say in our two predictors' case) are simply linear combinations of the original predictors.

We now illustrate how these directions can be calculated. Using singular value decomposition then

$X = H\,\Lambda^{\frac{1}{2}}\,G'$ where each of H, Λ , G is defined earlier

$X'X = G\,\Lambda^{\frac{1}{2}}\,H'H\,\Lambda^{\frac{1}{2}}\,G' = G\,\Lambda G'$

Since G is orthogonal matrix then the general linear regression model $y = X\beta + \varepsilon$ can be rewritten as

$$y = X\,GG'\,\beta + \varepsilon = Z\alpha + \varepsilon \qquad (12)$$

Where $Z= XG$ and $\alpha = G'\beta$

Hence:

$Z'Z = G'X'XG = G'\,(G\,\Lambda G')\,G = \Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_P)$

Where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_P > 0$ are the eigenvalues of X'X. The columns of G are the eigenvectors of X'X and the columns of Z are the principal components of X and these are orthogonal to each other.

Thus, the procedure creates a set of artificial variables $z_j^{'s}$ from the original $x_j{}^{'s}$ via a linear transformation Z = XG in such a way that the Z vectors are orthogonal to each other. The Z_j Corresponding to the largest λ_j value is called the principal component and it explains the largest proportion of the variation in the standardized dataset.

Further, $z_j^{'s}$ explain smaller and smaller until all variation is explained. Typically, one does not use all the $z_j^{'s}$ but follows some type of selection rule. No universal rule is presented for selecting the components. Some statisticians use the rule that only eigenvalues greater than 1 are of interest. Other statisticians suggested that the components might be computed until some arbitrarily large proportion

(maybe 0.75 or more) of the variances has been explained the OLS estimator of α is given as:

$$\hat{\alpha} = (Z'Z)^{-1} Z'y = \Lambda^{-1} G'X'y \qquad (13)$$

Assuming that the first q (q< p) principal components are selected, then the reduced estimator can be written as

$$\hat{\alpha}_q = (Z_q' Z_q)^{-1} Z'_q y = \Lambda_q^{-1} G'_q X'y \qquad (14)$$

Where $Z_q = XG_q$, G_q denote as the first q eigenvectors of X'X matrix and Λ_q is the diagonal matrix contains the first q eigenvalues of X'X.

To find the principal component estimator of the regression coefficients in terms of the original variables we can solve $\alpha = G'\beta$ for β to get $\beta = G\alpha$ since G is orthogonal matrix. Let b_{PC} denote the principal component estimator of β then

$$b_{PC} = G\,\hat{\alpha}$$

If q principal components are selected then

$$b_{(PC)q} = G_q\,\hat{\alpha}_q = G_q\,\Lambda_q^{-1} G'_q X'y \qquad (15)$$

The Simulation Results

To exhibit multi-collinearity in the simulated data, we use different degrees of correlation between the variables included in the model. Specifically, we assume correlation values to be $\rho = 0.75, 0.80$ and 0.95,four predictor variables have been generated. Since the performance of different estimators is influenced by the sample size, we have used three types of samples, small of size 20, median of size 50,80 and large of size 200. The standard deviations of the error terms are taken as $\sigma = 10, 25$ and 30 . Ordinary ridge estimates are computed using different ridge parameters given in equation (8) to (11) and the principal components regression given equations (12) to (15).

The mean square error (MSE) is used as a criterion in order to assess the performance of the stated methods. This experiment is repeated 1000 times. And the results are presented in tables (1), (2) and (3).

Table (1):The values of MSE at $\rho = 0.75$

n	Method	Standard deviation σ		
		10	25	30
20	PC	3.9252e-018	9.4206e-018	1.1776e-017
	$\hat{k}_{\text{HKB}}$	0.0215	0.0215	0.0214
	$\hat{k}_{\text{LW}}$	0.0027	0.0027	0.0027
	$\hat{k}_{\text{Bayes}}$	0.0275	0.0276	0.0276
	$\hat{k}_{\text{CN}}$	1.2561e-017	8.6355e-018	1.9626e-018
50	PC	2.9790e-018	4.9651e-018	1.4895e-017
	$\hat{k}_{\text{HKB}}$	0.0217	0.0220	0.0220
	$\hat{k}_{\text{LW}}$	8.2379e-004	8.2387e-004	8.2388e-004
	$\hat{k}_{\text{Bayes}}$	0.0151	0.0151	0.0151
	$\hat{k}_{\text{CN}}$	2.8797e-017	1.2909e-017	1.9860e-018
80	PC	3.9252e-018	4.7103e-018	1.2561e-017
	$\hat{k}_{\text{HKB}}$	0.0287	0.0287	0.0287
	$\hat{k}_{\text{LW}}$	6.2243e-004	6.2243e-004	6.2243e-004
	$\hat{k}_{\text{Bayes}}$	0.0575	0.0576	0.0576
	$\hat{k}_{\text{CN}}$	2.0411e-017	4.2392e-017	8.6355e-018
200	PC	2.4205e-018	2.7680e-017	4.2948e-017
	$\hat{k}_{\text{HKB}}$	0.0271	0.0273	0.0273
	$\hat{k}_{\text{LW}}$	1.1161e-004	1.1161e-004	1.1161e-004
	$\hat{k}_{\text{Bayes}}$	0.0370	0.0372	0.0372
	$\hat{k}_{\text{CN}}$	3.4524e-004	7.7579e-017	2.8425e-017

Table (2):The values of MSE at $\rho = 0.80$

n	Method	Standard deviation σ		
		10	25	30
20	PC	5.8878e-018	9.4206e-018	2.3551e-018
	$\hat{k}_{\text{HKB}}$	0.0215	0.0215	0.0214
	$\hat{k}_{\text{LW}}$	0.0027	0.0027	0.0027
	$\hat{k}_{\text{Bayes}}$	0.0275	0.0276	0.0276
	$\hat{k}_{\text{CN}}$	1.8056e-017	3.9252e-018	1.4916e-017
50	PC	7.9441e-018	1.6881e-017	6.9511e-018
	$\hat{k}_{\text{HKB}}$	0.0217	0.0220	0.0220
	$\hat{k}_{\text{LW}}$	8.2379e-004	8.2387e-004	8.2388e-004
	$\hat{k}_{\text{Bayes}}$	0.0151	0.0151	0.0151
	$\hat{k}_{\text{CN}}$	3.4755e-017	6.7525e-017	2.1846e-017
80	PC	2.7477e-017	1.2561e-017	4.7103e-018
	$\hat{k}_{\text{HKB}}$	0.0287	0.0287	0.0287
	$\hat{k}_{\text{LW}}$	6.2243e-004	6.2243e-004	6.2243e-004
	$\hat{k}_{\text{Bayes}}$	0.0575	0.0576	0.0576
	$\hat{k}_{\text{CN}}$	7.8505e-018	1.8056e-017	5.4953e-018
200	PC	2.6067e-018	4.2762e-017	1.8867e-017
	$\hat{k}_{\text{HKB}}$	0.0271	0.0273	0.0273
	$\hat{k}_{\text{LW}}$	1.1161e-004	1.1161e-004	1.1161e-004
	$\hat{k}_{\text{Bayes}}$	0.0370	0.0372	0.0372
	$\hat{k}_{\text{CN}}$	3.4879e-004	1.1935e-016	2.4701e-017

Table (3):The values of MSE at $\rho = 0.95$

n	Method	Standard deviation σ		
		10	25	30
20	PC	3.9252e-18	9.4206e-018	9.8131e-018
	$\hat{k}_{\text{HKB}}$	0.0215	0.0215	0.0215
	$\hat{k}_{\text{LW}}$	0.0027	0.0027	0.0027
	$\hat{k}_{\text{Bayes}}$	0.0275	0.0276	0.0276
	$\hat{k}_{\text{CN}}$	2.5121e-017	8.6355e-018	1.0206e-017
50	PC	1.7874e-017	5.9581e-018	0
	$\hat{k}_{\text{HKB}}$	0.0216	0.0220	0.0220
	$\hat{k}_{\text{LW}}$	8.2376e-004	8.2387e-004	8.2387e-004
	$\hat{k}_{\text{Bayes}}$	0.0151	0.0151	0.0151
	$\hat{k}_{\text{CN}}$	7.9441e-018	1.7874e-017	2.3832e-017
80	PC	7.8505e-018	1.5701e-017	4.7103e-018
	$\hat{k}_{\text{HKB}}$	0.0287	0.0287	0.0287
	$\hat{k}_{\text{LW}}$	6.2243e-004	6.2243e-004	6.2243e-004
	$\hat{k}_{\text{Bayes}}$	0.0575	0.0576	0.0576
	$\hat{k}_{\text{CN}}$	1.3346e-017	2.0411e-017	5.4953e-018
200	PC	2.5446e-018	4.4686e-018	2.4081e-017
	$\hat{k}_{\text{HKB}}$	0.0271	0.0273	0.0273
	$\hat{k}_{\text{LW}}$	1.1161e-004	1.1161e-004	1.1161e-004
	$\hat{k}_{\text{Bayes}}$	0.0370	0.0372	0.0372
	$\hat{k}_{\text{CN}}$	3.8924e-004	1.1116e-016	1.6720e-016

Conclusions

In the sense of mean square error (MSE) as a criterion of performance. In our paper "An Alternative Approach for selecting Ridge Parameter for Ordinary Ridge Regression Estimator Regression Estimator" International Journal of Science and Research (IJSR).We made the comparison between the performance of different type ordinary ridge regression as well as the generalize ridge regression and we found the proposed method was the best in the since of MSE, while in this paper we introduced another which will be known method of estimation which is the principal component method and compared it with many different types of ridge regression estimator, the simulation results displayed that the principal components estimator performs better than all types of ordinary ridge regression estimators that are included in this articles while the ordinary regression estimator based on the ridge parameter $\hat{k}_{CN}$ seems to be better than other studied types of ordinary ridge regression estimators in all conditions of multi-collinearity levels, sample sizes and the standard deviations of the error terms. However, for the purpose of future works, many other methods can be used to overcome the multi-collinearity problem such as the generalized inverse method and the jackknife ridge regression method**.**

References

[1] H.M. Gorgees, “Using Singular Value Decomposition Method for Estimating the Ridge Parameter", Journal of Economic and Administrative Science, 1-10. 2009

[2] A. E. Hoerl and R.W. Kennard. "Ridge Regression: Biased Estimation of Nonorthogonal Problems", Techno metrics. 55-67. 1970

[3] A.E. Hoerl, R. W. Kennard and K.F. Baldwin. "Ridge Regression: some simulation" Communications in Statistics. 105-123. 1975

[4] J.F Lawless and p. Wang. "A simulation study of Ridge and other Regression Estimators", Communications in Statistics - theory and Methods .1177-1182. 2005

[5] M.S srivastava. "Methods of Multivariate Statistics", Wiley, New York. 2002

[6] H.M. Gorgees, and F. A. Mahdi, "An Alternative Approach for selecting Ridge Parameter for Ordinary Ridge Regression Estimator Regression Estimator", International Journal of Science and Research (IJSR). 2426-2429. 2017