International Journal of Mathematics and
Computer Science, **16**(2021), no. 4, 1721–1727

# The Efficiency of Ridge Estimations for Multicollinearity Multiple Linear Regression: A Monte-Carlo Simulation-Based Study

**Jeeraporn Thaithanan**[1], **Wandee Wanishsakpong**[1], **Thammarat Panityakul**[2], **Dulyawit Prangchumpol**[3]

[1]Department of Statistics
Faculty of Science
Kasetsart University
Bangkok 10900, Thailand

[2]Division of Computational Science
Faculty of Science
Prince of Songkla University
Hat Yai, Songkhla 90110, Thailand

[3]Department of Information Technology
Faculty of Science and Technology
Suan Sunandha Rajabhat University
Bangkok 10300, Thailand

email: jeeraporn.t@ku.th, wandee.w@ku.th,
thammarat.p@psu.ac.th, dulyawit.pr@ssru.ac.th

(Received April 30, 2021, Accepted June 9, 2021)

**Abstract**

The aim of this study is to compare the coefficients estimations of the multiple linear regression with multicollinearity. The ordinary least squares method (OLS), modified ridge regression method (MRR) and generalized Liu-Kejian method (LKM) are being compared in order to measure the efficiency of estimations by the mean of average mean square error (AMSE). The simulation scenarios for this study are

**Key words and phrases:** Multiple Linear Regression, Least Squares Method, Multicollinearity, Ridge Regression, Generalized Liu Kejian Method, Monte-Carlo Simulation.
**AMS (MOS) Subject Classifications**: 62J07.
**ISSN** 1814-0432, 2021, http://ijmcs.future-in-tech.net

3 and 5 independent variables with the zero mean normally distributed random error of variance 1, 5, and 10, three correlation coefficients levels; i.e., low (0.3), medium (0.6) and high (0.9) are determined for independent variables, all combinations be performed with the sample sizes 20, 50 and 100 by Monte Carlo simulation technique for 1,000 times in each situation. The AMSE decrease as the sample size grew. The MRR and LKM performed better than LSM. The MRR is the most suitable for all scenarios at random error of variance 10.

# 1 Introduction and Preliminaries

Nowadays, a most useful and famous statistical application like multiple linear regression (MLR) has been applied in various fields; for instance, natural science, social science, engineering, economics and demographics. This approach is a statistical technique that uses several predictors (independent variables) in order to predict the values of a response (dependent variable). The goal of MLR is to find out the best model which can describe the linear relationship between the predictor and response variables. A major task of MLR after the best subsets of predictors obtained is the coefficient estimation, the most fitted estimates, and the least of errors. The common and reasonable approach is called the least squared approach has been the household tool for estimation. However, this approach has limitation of multicollinearity, a huge obstacle of MLR.

## 1.1 MLR coefficients estimations

To deal with this situation, ridge regression is modified, especially when a high correlation of predictors exist. Ridge regression estimator was first proposed by Hoerl and Kennard in 1970 [5], by adding a scalar multiplication, the product of a positive real number and identity matrix, within the inverse component of the least square estimator. This provided a more precise ridge parameters estimates than least square estimates, and its variance and mean square errors are most often smaller than the least square estimates as well.

Besides the ridge estimation, many modifications of ridge estimation have been studied; for instance, a modified ridge-type and principal component regression estimators [9], prior information-based ridge estimators [11], unbiased ridge estimator [4], and the Liu Kejian Method (LKM) [6]. This study compares the three methods of estimating multiple linear regression coefficients; i.e., Least Squares Method (OLS), Modified Ridge Regression Method

(MRR) and the Liu Kejian Method (LKM) as follows:

1. Ordinary Least Square Method (OLS), the multiple linear regression coefficients estimated by this method will be unbiased estimates and has the least variance of estimation, called the best linear unbiased estimator (BLUE). The estimated value is in the form of

$$\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}, \tag{1.1}$$

where $\boldsymbol{X}$ is the $n \times p$ predictors matrix, $\boldsymbol{Y}$ is the $n \times 1$ observation vector, $\hat{\boldsymbol{\beta}}_{OLS}$ is the vector of coefficients estimates. The mean square error of $\hat{\boldsymbol{\beta}}_{OLS}$ is $\sigma^2 tr(\boldsymbol{X}'\boldsymbol{X})^{-1}$.

2. Modified Ridge Regression (MRR), the ridge estimation for multiple linear regression coefficients is

$$\hat{\boldsymbol{\beta}}_{Ridge} = (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}_n)^{-1}\boldsymbol{X}'\boldsymbol{Y}, k > 0, \tag{1.2}$$

where $\hat{\boldsymbol{\beta}}_{Ridge}$ is $p \times 1$ ridge estimator, $k$ is a positive real number also known as a constant bias ridge, $\boldsymbol{I}_n$ is an identity matrix of size $n$. If we apply the historical data to the ridge regression method, the approximation of the linear regression coefficients are more accurate and close to the real values. This method is called modified ridge regression (MRR) [4] as follows:

$$\hat{\boldsymbol{\beta}}_{MRR} = (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}_n)^{-1}(\boldsymbol{X}'\boldsymbol{Y} + k\boldsymbol{J}), \tag{1.3}$$

where $\boldsymbol{J}$ is $p \times 1$ historical observation vector, $\boldsymbol{J} = (\sum_{i=1}^{p} \frac{\hat{\beta}_{OLS_i}}{p})\mathbf{1}$, $\mathbf{1}$ is $p \times 1$ vector of ones where every element is equal to one. From equation (1.3), $\hat{\boldsymbol{\beta}}_{MRR} = \hat{\boldsymbol{\beta}}_{OLS}$ when $k = 0$

The estimation of $k$ is considered using the following two cases:

1. $\sigma^2$ is known,

$$\hat{k} = \begin{cases} \frac{p\boldsymbol{\sigma}^2}{(\hat{\boldsymbol{\beta}}_{OLS}-\boldsymbol{J})'(\hat{\boldsymbol{\beta}}_{OLS}-\boldsymbol{J})-\boldsymbol{\sigma}^2 tr(\boldsymbol{X}'\boldsymbol{X})^{-1}}, \\ \text{if}(\hat{\boldsymbol{\beta}}_{OLS}-\boldsymbol{J})'(\hat{\boldsymbol{\beta}}_{OLS}-\boldsymbol{J}) - \boldsymbol{\sigma}^2 tr(\boldsymbol{X}'\boldsymbol{X})^{-1} > 0 \\ \frac{p\boldsymbol{\sigma}^2}{(\hat{\boldsymbol{\beta}}_{OLS}-\boldsymbol{J})'(\hat{\boldsymbol{\beta}}_{OLS}-\boldsymbol{J})}, \text{otherwise} \end{cases}$$

2. $\sigma^2$ is unknown,

$$\hat{k} = \begin{cases} \frac{p\hat{\boldsymbol{\sigma}}^2}{(\hat{\boldsymbol{\beta}}_{OLS}-\boldsymbol{J})'(\hat{\boldsymbol{\beta}}_{OLS}-\boldsymbol{J})-\hat{\boldsymbol{\sigma}}^2 tr(\boldsymbol{X}'\boldsymbol{X})^{-1}}, \\ \text{if}(\hat{\boldsymbol{\beta}}_{OLS}-\boldsymbol{J})'(\hat{\boldsymbol{\beta}}_{OLS}-\boldsymbol{J})-\hat{\boldsymbol{\sigma}}^2 tr(\boldsymbol{X}'\boldsymbol{X})^{-1} > 0 \\ \frac{p\hat{\boldsymbol{\sigma}}^2}{(\hat{\boldsymbol{\beta}}_{OLS}-\boldsymbol{J})'(\hat{\boldsymbol{\beta}}_{OLS}-\boldsymbol{J})}, \text{otherwise,} \end{cases}$$

where $\hat{\boldsymbol{\sigma}}^2 = \frac{(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS})'(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS})}{n-p}$ is an unbiased estimator of $\boldsymbol{\sigma}^2$. The mean square of $\hat{\boldsymbol{\beta}}_{MRR}$ is $\hat{\boldsymbol{\sigma}}^2 tr((\boldsymbol{X}'\boldsymbol{X}+\hat{k}\boldsymbol{I}_n)^{-1}(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X}+\hat{k}\boldsymbol{I}_n)^{-1}) + \hat{k}^2(\hat{\boldsymbol{\beta}}_{OLS}-\boldsymbol{J})(\boldsymbol{X}'\boldsymbol{X}+\hat{k}\boldsymbol{I}_n)^{-2}(\hat{\boldsymbol{\beta}}_{OLS}-\boldsymbol{J})$.

3. Generalized Liu Kejian Method (LKM) [6], a method for estimating the multiple linear regression coefficient in the case of a multiple-relationship between the independent variables. By combining the advantages of the Ridge Regression method and the Stein method [10]. This method is called the Generalized Kejian Method and the form of the multiple linear regression coefficient estimator is

$$\hat{\boldsymbol{\beta}}_{LKM} = (\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{I}_n)^{-1}(\boldsymbol{X}'\boldsymbol{Y}+d\hat{\boldsymbol{\beta}}_{OLS}), 0 < d < 1 \tag{1.4}$$

when $d = 1$, $\hat{\boldsymbol{\beta}}_{LKM} = \hat{\boldsymbol{\beta}}_{OLS}$ and

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{LKM} &= (\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{I}_n)^{-1}(\boldsymbol{X}'\boldsymbol{Y}+\boldsymbol{D}\hat{\boldsymbol{\beta}}_{OLS}) \\ &= (\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{I}_n)^{-1}(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{D})\hat{\boldsymbol{\beta}}_{OLS} \\ &= (\boldsymbol{I}_n-(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{I}_n)^{-1}(\boldsymbol{I}_n-\boldsymbol{D}))\hat{\boldsymbol{\beta}}_{OLS} \\ &= (\boldsymbol{I}_n-(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{I}_n)^{-2}(\boldsymbol{I}_n-\boldsymbol{D})^2)\hat{\boldsymbol{\beta}}_{OLS} \end{aligned} \tag{1.5}$$

where $D = diag(d_1, d_2, \ldots, d_p), 0 < d_i < 1, i = 1, 2, \ldots, p$ and the estimates of $d_i$ is

$$\hat{d_i} = 1 - \frac{\hat{\sigma}(\lambda_i+1)}{\sqrt{\lambda_i\hat{\boldsymbol{\beta}}^2_{OLS_i}+\hat{\sigma}^2}}, i = 1, 2, \ldots, p$$

The mean square error of $\hat{\boldsymbol{\beta}}_{LKM}$ is $(\boldsymbol{I}_n-\Delta^2)(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{I}_n-\Delta^2)\sigma^2 + \Delta^2\boldsymbol{\beta}\boldsymbol{\beta}'\Delta^2$, where $\Delta = (\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{I}_n)^{-1}(\boldsymbol{I}_n-\boldsymbol{D})$

## 1.2 Monte Carlo simulation

A Monte Carlo simulation scenario for this study [8] are 3 and 5 independent variables with the zero mean normally distributed random error of variance 1, 5, and 10, three correlation coefficients level; i.e., low (0.3), medium (0.6)

and high (0.9) are determined for independent variables, all combinations are performed with the sample sizes 20, 50 and 100 by Monte Carlo simulation technique for 1,000 times in each situation. First, the random error ($\varepsilon$) is simulated as $\varepsilon \sim \boldsymbol{N}(0, \sigma_\varepsilon^2 \boldsymbol{I}_n)$, where $\sigma_\varepsilon^2 = 1, 5, 10$. Secondly, an observation matrix, $\boldsymbol{X}$, is simulated from $\boldsymbol{X} \sim \boldsymbol{N}_n(\boldsymbol{0}, \boldsymbol{I}_n)$ with different levels of polynomial relations such that $\rho = 0.3, 0.6, 0.9$. Thirdly, generate response values of $\boldsymbol{Y}$ from the model with multiple linear regression coefficient $\boldsymbol{\beta}$. Finally, multiple linear regression coefficients are estimated for all methods. Repeat all steps above 1,000 times in each scenario. Then calculate the mean of the mean square error of multiple linear regression, AMSE= $\frac{1}{1000}(\sum_{i=1}^{1000} MSE)$, the method with lowest AMSE is selected as the best method for the scenario involved.

## 2 The Simulation Results and Discussion

Table 1 demonstrates the best multicollinearity MLR coefficients estimation method of each simulation condition. Obviously, the OLS is not suitable for all conditions. The most appropriate method to estimate MLR coefficients when multicollinearity exists is MRR and LKM. MRR is suitable for all sample sizes and the data with the low correlation degree of the predictors with small at moderate variance of error. The Generalized Liu Kejian Method is suitable for small data and high degree of correlation of the predictors at high variance of the error. When the number of predictors increases, LKM is tend to do better than MRR, but more predictors, more risk of multicollinearity. Some additional modification should be considered; for instance, kernel ridge regression [2], weighted ridge regression [3], ridge-lasso regression [1] and Bayesian ridge regression [7].

| Predictors | $\sigma$ | $\rho$ | $n = 20$ | $n = 50$ | $n = 100$ |
|---|---|---|---|---|---|
| 3 | 1 | 0.3 | MRR | MRR | MRR |
| 3 | 1 | 0.6 | MRR | MRR | MRR |
| 3 | 1 | 0.9 | MRR | MRR | MRR |
| 3 | 5 | 0.3 | MRR | MRR | MRR |
| 3 | 5 | 0.6 | MRR | MRR | MRR |
| 3 | 5 | 0.9 | MRR | MRR | MRR |
| 3 | 10 | 0.3 | LKM | MRR | MRR |
| 3 | 10 | 0.6 | LKM | LKM | MRR |
| 3 | 10 | 0.9 | LKM | MRR | MRR |
| 5 | 1 | 0.3 | MRR | MRR | MRR |
| 5 | 1 | 0.6 | MRR | MRR | MRR |
| 5 | 1 | 0.9 | MRR | MRR | MRR |
| 5 | 5 | 0.3 | LKM | MRR | MRR |
| 5 | 5 | 0.6 | MRR | MRR | MRR |
| 5 | 5 | 0.9 | MRR | MRR | MRR |
| 5 | 10 | 0.3 | LKM | MRR | MRR |
| 5 | 10 | 0.6 | LKM | LKM | MRR |
| 5 | 10 | 0.9 | LKM | LKM | LKM |

Table 1: The best method of all scenarios from 1000-Monte Carlo simulations

**Acknowledgment.** The authors would like to thank Suan Sunandha Rajabhat University for scholarship support.

# References

[1] A. Bedoui, N. A. Lazar, Bayesian empirical likelihood for ridge and lasso regressions, Comput. Stat. Data Anal., **145**, (2020), 106917.

[2] H. Chen, J. Leclair, Optimizing etching process recipe based on Kernel Ridge Regression, J. Manufacturing Processes, **61**, (2021), 454–460.

[3] S. Chen, L. Xiong, Q. Ma, J. S. Kim, J. Chen, C. Y. Xu, Improving daily spatial precipitation estimates by merging gauge observation with multiple satellite-based precipitation products based on the geographically weighted ridge regression method, J. Hydrology, **589**, (2020), 125156.

[4] R. Crouse, C. Jin, R. Hanumara, Unbiased ridge estimation with prior information and ridge trace, Commun. Stat. Theory Methods, **24**, (1995), 2341–2354.

[5] A. E. Hoerl, R. W. Kennard, Ridge Regression: Biased Estimation for Nonorthogonal Problems, Technometrics, **12,** no. 1, (1970), 55–67.

[6] K. Liu, A new class of biased estimate in linear regression, Commun. Stat. Theory Methods, **22**, no. 2, (1993), 393–402.

[7] Y. Liu, L. Sun, C. Du, X. Wang, Near-infrared prediction of edible oil frying times based on Bayesian Ridge Regression, Optik, **218**, (2020).

[8] A. F. Lukman, K. Ayinde, S. A. Ajiboye, Monte-Carlo study of some classification-based ridge parameter estimators J. Modern Appl. Stat. Methods, **16**, no. 1, (2017), 428–451.

[9] A. F. Lukman, K. Ayinde, O. Oludoun, C. A. Onate, Combining modified ridge-type and principal component regression estimators, Scientific African, **9**, (2020), e00536.

[10] C. Stein, A bound for the error in the normal approximation to the distribution of a sum of dependent random variables, Proc. Sixth Berkeley Symp. on Math. Stat. Prob., Univ. of California, **2**, (1972), 583–602.

[11] B. F. Swindel, Good ridge estimators based on prior information, Commun. Stat. Theory Methods, **5**, (1976), 1065–1075.