MATEMATIKA, MJIAM, Volume 41, Number 1, 57–76

A Theoretical and Simulation Study in Formulating the Optimal Biasing Parameter for an Almost Unbiased Regression Estimator

[1]Set Foong Ng*

[1]College of Computing, Informatics and Mathematics, Universiti Teknologi MARA,
Cawangan Johor Kampus Pasir Gudang, 81750 Masai, Johor, Malaysia

*Corresponding author: ngsetfoong061@uitm.edu.my

Article history
Received: 11 October 2024
Received in revised form: 17 January 2025
Accepted: 6 March 2025
Published on line: 1 April 2025

Abstract To combat multicollinearity problem in linear regression model, a biased estimator named as the k-almost unbiased regression estimator (KAURE) was investigated in this study. KAURE is associated with the biasing parameter, and the mean squared error (MSE) of KAURE is sensitive to changes in the biasing parameter. Theoretical and numerical comparisons in the previous research showed that KAURE outperformed ordinary least squares estimator (OLSE) in terms of MSE when the biasing parameter was within a specific range. However, KAURE is not unique within the specified range of the biasing parameter, which complicates its practical use in linear regression modeling. Hence, there is a need to find an optimal biasing parameter for KAURE to enable practitioners across various fields to effectively use KAURE. In this paper, some new methods to estimate the optimal biasing parameter for KAURE that minimizes its MSE were proposed. Extensive Monte Carlo simulations were conducted to evaluate the performance of the proposed methods based on the average mean squared error (AMSE) criterion by varying the values of different factors (sample size, error standard deviation and degree of multicollinearity). The simulation results were confirmed by the empirical application. Thus, the proposed optimal biasing parameter provides a novel approach to formulating KAURE, enhancing its effectiveness as a practical alternative to OLSE for addressing multicollinearity issues in linear regression models.

Keywords Multicollinearity; Biasing parameter; Mean squared error; Linear regression model.

Mathematics Subject Classification 62F10.

1 Introduction

Linear regression is a statistical method that has various applications across different fields. Some notable areas where linear regression is commonly used are environmental science, medicine and healthcare, engineering, sports analytics, economics and finance. The most widely utilized estimation technique in regression analysis is the ordinary least squares estimator (OLSE).

While OLSE is an unbiased estimator in the regression model, its efficacy diminishes in the presence of multicollinearity in the data [1–3]. Multicollinearity is a widespread challenge with notable implications across diverse fields, especially during the application of linear regression analysis.

According to Belsley [4], multicollinearity is an inherent imperfection in data, stemming from uncontrollable processes within the data-generating mechanism. This phenomenon manifests when two or more independent variables in a regression model exhibit high correlation. Hence, multicollinearity arises when there is a nearly exact relationship between two or more independent variables in a dataset. The accuracy of the OLSE is affected by its large variance when multicollinearity is present in the data. In practical terms, multicollinearity can result in elevated standard errors of the coefficients. The consequence of having large variance extends to the inflation of confidence interval widths for the parameters in the regression model [5]. Furthermore, multicollinearity can misguide significance tests, erroneously suggesting that certain crucial variables are unnecessary in the model. As a result, coefficients may be deemed statistically insignificant even when the variables hold theoretical importance [6]. Beyond this, multicollinearity causes a reduction of statistical power of statistical tests. In studies where the primary focus lies in parameter estimation and the identification of important variables in the process, the impact of multicollinearity emerges as a serious concern.

Suppose $\mathbf{Y}$ is an $n \times 1$ vector of standardized dependent variables, $\mathbf{X}$ is an $n \times p$ matrix of standardized independent variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of errors such that $\boldsymbol{\varepsilon} \sim N\left(0, \sigma^2 \mathbf{I}_n\right)$ and $\mathbf{I}_n$ is an identity matrix of dimension n. Then, we may use the matrix form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, to represent a linear regression model with p standardized independent variables and a standardized dependent variable, y.

Let the matrix $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_p)$ be a $p \times p$ diagonal matrix whose diagonal elements are the eigenvalues of $\mathbf{X}'\mathbf{X}$ where $\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p = \lambda_{\min} > 0$. Let the matrix $\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_p]$ be a $p \times p$ orthonormal matrix consisting of the p eigenvectors of $\mathbf{X}'\mathbf{X}$. Here, the matrix $\mathbf{Q}$ and $\boldsymbol{\Lambda}$ satisfy $\mathbf{Q}'\mathbf{X}'\mathbf{X}\mathbf{Q} = \boldsymbol{\Lambda}$ and $\mathbf{Q}'\mathbf{Q} = \mathbf{Q}\mathbf{Q}' = \mathbf{I}_p$, where $\mathbf{I}_p$ is a $p \times p$ identity matrix.

The linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ can be transformed into a canonical form $\mathbf{Y} = \mathbf{Z}\boldsymbol{a} + \boldsymbol{\varepsilon}$, where $\mathbf{Z} = \mathbf{X}\mathbf{Q}$ is an $n \times p$ matrix, $\boldsymbol{a} = \mathbf{Q}'\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters, and $\mathbf{Z}'\mathbf{Z} = \boldsymbol{\Lambda}$.

Let $\hat{\boldsymbol{\alpha}}$ be the OLSE of parameter $\boldsymbol{\alpha}$. The OLSE is given by

$$\hat{\boldsymbol{\alpha}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}. \tag{1}$$

OLSE is an unbiased estimator that has no bias. However, the variance of OLSE is unacceptably large when multicollinearity is present in the regression model. Therefore, the accuracy of the parameter estimates by using OLSE is affected. Hence, biased estimators are introduced as an alternative to the OLSE. Although there is an amount of bias in biased estimators, its smaller variance would result in a smaller mean squared error (MSE) compared to the MSE of OLSE. As a result, the accuracy of parameter estimates by using biased estimators is better. There are many biased estimators that have been introduced such as almost unbiased modified ridge-type estimator [7], modified two-parameter regression estimator [8], new Ridge-type estimator [9] modified Ridgetype estimator [2], almost unbiased Ridge regression estimator [10] Liu-type estimator [11], Liu estimator [12], restricted Ridge regression estimator [13], Principal component regression estimator [14–17], iteration estimator [18], Ridge regression estimator [19] and Shrunken estimator [20].

A biased estimator named as the k-almost unbiased regression estimator (KAURE) was developed in [1]. The KAURE of parameter $\boldsymbol{\alpha}$ is given by

$$\hat{\boldsymbol{a}}_{KAURE} = \left[\mathbf{I} - (\mathbf{Z}'\mathbf{Z} + k\mathbf{I})^{-2}(k-1)^2\right]\hat{\boldsymbol{a}} \tag{2}$$

Let $\mathbf{H}_{KAURE} = \mathbf{I} - (\boldsymbol{\Lambda} + k\mathbf{I})^{-2}(k-1)^2$. The equations of the estimator KAURE as well as its bias, variance-covariance, and MSE of KAURE are given in Equations (3) to (7).

$$\hat{\boldsymbol{a}}_{KAURE} = \mathbf{H}_{KAURE}\hat{\boldsymbol{a}} \tag{3}$$

$$\mathbf{bias}(\hat{\boldsymbol{a}}_{KAURE}) = (\mathbf{H}_{KAURE} - \mathbf{I})\boldsymbol{a} \tag{4}$$

$$\mathbf{cov}(\hat{\boldsymbol{a}}_{KAURE}) = \mathbf{H}_{KAURE}\boldsymbol{\Lambda}^{-1}\mathbf{H}'_{KAURE}\sigma^2 \tag{5}$$

$$\mathbf{MSE}(\hat{\boldsymbol{\alpha}}_{KAURE}) = \mathbf{H}_{KAURE}\boldsymbol{\Lambda}^{-1}\mathbf{H}'_{KAURE}\sigma^2 + \boldsymbol{\alpha}'(\mathbf{H}_{KAURE} - \mathbf{I})'(\mathbf{H}_{KAURE} - \mathbf{I})\boldsymbol{\alpha} \tag{6}$$

$$= \sigma^2\sum_{j=1}^{p}\frac{1}{\lambda_j}\left[1 - \left(\frac{k-1}{\lambda_j + k}\right)^2\right]^2 + \sum_{j=1}^{p}\alpha_j^2\left(\frac{k-1}{\lambda_j + k}\right)^4 \tag{7}$$

KAURE is associated with the biasing parameter k, and it has been observed that the MSE of KAURE is sensitive to changes in k. The theoretical and numerical comparisons in [1] showed that KAURE outperforms OLSE in terms of MSE when k is within a specific range. However, KAURE is not unique within the specified range of k, making its practical application in linear regression modeling challenging. Therefore, it is essential to identify an optimal biasing parameter for KAURE, allowing practitioners in various fields to effectively utilize it in linear regression models with multicollinear data. This paper expands on the research presented in [1]. The objective of this paper is to propose and evaluate methods for estimating the optimal biasing parameter for KAURE, aimed at minimizing its MSE.

The rest of the paper is organized as follows. The methodology for estimating the optimal biasing parameter for KAURE is detailed in Section 2. Section 3 details the simulation design and evaluates the performance of the proposed methods. Section 4 presents a numerical example to illustrate the application of the proposed optimal biasing parameter for KAURE. Some industrial applications of linear regression are discussed Section 5. Section 6 concludes the work.

2 The Proposed Optimal Biasing Parameter for KAURE

Biasing parameters are crucial for formulating a biased estimator and determining its MSE. The estimation of biasing parameters for biased estimators is fundamentally linked to the minimization of the MSE of estimator to enhance estimator accuracy. In the case of Generalized Ridge regression estimator (GRRE), $\hat{\boldsymbol{\alpha}}_k = (\mathbf{Z}'\mathbf{Z} + \mathbf{K})^{-1}\mathbf{Z}'\mathbf{Y}$ that was proposed by Hoerl and Kennard [19]. Here, $\mathbf{K} = \operatorname{diag}(k_1, k_2, \ldots, k_p)$ consists of biasing parameters named ridge parameters. Hoerl and Kennard [19] proposed $k_j = \dfrac{\hat{\sigma}^2}{\hat{\alpha}_j^2}$ to estimate the biasing parameter in GRRE. Throughout literature, many researchers have suggested various methods in estimating ridge parameters [5, 21–29]. For instance, Kibria [25] proposed the arithmetic mean of $k_j = \dfrac{\hat{\sigma}^2}{\hat{\alpha}_j^2}$ and median of $k_j = \dfrac{\hat{\sigma}^2}{\hat{\alpha}_j^2}$ to estimate ridge parameters. Dorugade [27] utilized the concepts of arithmetic mean, geometric mean, harmonic mean, and median in determining the value of the ridge parameter. Some studies suggested including the largest eigenvalue $\lambda_{\max}$ of $\mathbf{X'X}$ in estimating the biasing parameter [5, 26, 27]. Jacob and Varadharajan [29] proposed the robust variance inflation factor and applied it in ridge parameter estimation.

Hoerl et al. [19] stated that the biasing parameter should be selected such that the MSE of the biased estimator is smaller than that of the OLSE. Therefore, the key to estimating biasing parameter for biased estimators lies in employing search methods that minimize the MSE of the estimator. This approach is supported by a substantial body of research demonstrating that effective selection of biasing parameters can lead to improved estimator performance [28].

The methods used in previous studies on estimating biasing parameters for other estimators were adopted as a fundamental guideline to estimate biasing parameter for KAURE in this study. In this paper, the optimal biasing parameter for KAURE was developed based on the approach of minimizing MSE of KAURE.

Let $u_j = \dfrac{k-1}{\lambda_j + k}$. The $\mathbf{MSE}\left(\hat{\boldsymbol{\alpha}}\right)$ in Equation (7) can be expressed

$$\mathbf{MSE}(\hat{\boldsymbol{\alpha}}_{KAURE}) = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}\left(1 - u_j^2\right)^2 + \sum_{j=1}^{p} \alpha_j^2 u_j^4 \tag{8}$$

Hence, we obtain the following derivatives.

$$\frac{du_j}{dk} = \frac{\lambda_j + 1}{(\lambda_j + k)^2} \tag{9}$$

$$\frac{d\left[\mathbf{MSE}(\hat{\boldsymbol{\alpha}}_{KAURE})\right]}{du_j} = u_j \left[\left(\frac{4\sigma^2 + 4\lambda_j \alpha_j^2}{\lambda_j}\right) u_j^2 - \frac{4\sigma^2}{\lambda_j}\right] \tag{10}$$

$$\begin{aligned}\frac{d\left[\mathbf{MSE}(\hat{\boldsymbol{\alpha}}_{KAURE})\right]}{dk} &= \frac{d\left[\mathbf{MSE}(\hat{\boldsymbol{\alpha}}_{KAURE})\right]}{du_j} \cdot \frac{du_j}{dk} \\ &= u_j \left[\left(\frac{4\sigma^2 + 4\lambda_j \alpha_j^2}{\lambda_j}\right) u_j^2 - \frac{4\sigma^2}{\lambda_j}\right] \cdot \frac{\lambda_j + 1}{(\lambda_j + k)^2}\end{aligned} \tag{11}$$

The value of k that minimizes MSE of KAURE is obtained by solving $\dfrac{d\left[\mathbf{MSE}(\hat{\boldsymbol{\alpha}}_{KAURE})\right]}{dk} =$

0 as below.

$$\left(\frac{4\sigma^2 + 4\lambda_j\alpha_j^2}{\lambda_j}\right) u_j^2 - \frac{4\sigma^2}{\lambda_j} = 0$$

$$u_j = \sqrt{\frac{\sigma^2}{\sigma^2 + \lambda_j\alpha_j^2}}$$

$$k = \frac{1 + \lambda_j\sqrt{\frac{\sigma^2}{\sigma^2 + \lambda_j\alpha_j^2}}}{1 - \sqrt{\frac{\sigma^2}{\sigma^2 + \lambda_j\alpha_j^2}}} \tag{12}$$

Adopting algorithms outlined in [25,27], we propose the following methods (k_A, k_M and k_G) to estimate the optimal biasing parameter for KAURE.

$$k_A = \text{Arithmetic Mean}\left[f_j\right], \tag{13}$$

$$k_M = \text{Median}\left[f_j\right], \tag{14}$$

$$k_G = \text{Geometric Mean}\left[f_j\right], \tag{15}$$

where $f_j = \dfrac{1 + \lambda_{\max}\sqrt{\dfrac{\sigma^2}{\sigma^2 + \lambda_{\max}\alpha_j^2}}}{1 - \sqrt{\dfrac{\sigma^2}{\sigma^2 + \lambda_{\max}\alpha_j^2}}}$ and $j = 1, 2, \cdots, p$.

3 Monte Carlo Simulation Study

In this section, Monte Carlo simulation was conducted to evaluate the performance of the proposed methods to estimate the optimal biasing parameter for KAURE. The simulation design was first explained in Section 3.1. The simulation results were presented and discussed in Section 3.2. Python programming was used to conduct the simulation.

3.1 Simulation Design

The explanatory variables are generated following the methods of Xu & Yang [31] and Liu [11]. The formulation for the explanatory variables is defined as

$$x_{ij}^* = \left(1 - \rho^2\right)^{\frac{1}{2}} u_{ij} + \rho u_{i,p+1}, \tag{16}$$

where u_{ij} are independent pseudo-random numbers that follow standard normal distribution, $i = 1, 2, ..., n$ and $j = 1, 2, ..., p$.

In this simulation, we consider $p = 3$. Hence, the explanatory variables are generated using the formulation below

$$x_{ij}^* = \left(1 - \rho^2\right)^{\frac{1}{2}} u_{ij} + \rho u_{i4}, \tag{17}$$

where $i = 1, 2, ..., n$ and $j = 1, 2, 3$.

The dependent variable is determined by

$$y_i^* = \beta_1 x_{i1}^* + \beta_2 x_{i2}^* + \beta_3 x_{i3}^* + \varepsilon_i, \qquad i = 1, 2, ..., n, \tag{18}$$

where the error ε_i are independent pseudo-random numbers that follow normal distribution with mean zero and variance σ^2.

In this simulation, we choose $(\beta_1, \beta_2, \beta_3)' = (1, 2, 3)'$.

Standardization is done on dependent variable (y_i^*) and explanatory variables (x_{i1}^*, x_{i2}^* and x_{i3}^*) using the Equation (19) and Equation (20), respectively. The standardized dependent variable is denoted by y_i and the standardized independent variables are denoted by x_{ij}, where $i = 1, 2, ..., n$ and $j = 1, 2, 3$.

$$y_i = \frac{y_i^* - \frac{1}{n}\sum\limits_{i=1}^{n} y_i^*}{\sqrt{\sum\limits_{i=1}^{n}\left(y_i^* - \frac{1}{n}\sum\limits_{i=1}^{n} y_i^*\right)^2}}, \tag{19}$$

$$x_{ij} = \frac{x_{ij}^* - \frac{1}{n}\sum\limits_{i=1}^{n} x_{ij}^*}{\sqrt{\sum\limits_{i=1}^{n}\left(x_{ij}^* - \frac{1}{n}\sum\limits_{i=1}^{n} x_{ij}^*\right)^2}}. \tag{20}$$

The matrix $\mathbf{X}$ is an $n \times 3$ matrix of standardized independent variables. The vector $\mathbf{Y}$ is an $n \times 1$ vector of standardized dependent variable. The linear regression model is represented by $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. The matrix $\boldsymbol{\Lambda}$ is a 3×3 diagonal matrix in which the diagonal elements are the eigenvalues of $\mathbf{X}'\mathbf{X}$. The matrix $\mathbf{Q}$ is a 3×3 orthonormal matrix consisting of the eigenvectors of $\mathbf{X}'\mathbf{X}$. The linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ is transformed into a canonical form $\mathbf{Y} = \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$, where $\mathbf{Z} = \mathbf{XQ}$.

The OLSE of parameter $\boldsymbol{\alpha}$ is defined by $\hat{\boldsymbol{\alpha}}$ in Equation (1). Its MSE is given by

$$\mathbf{MSE}(\hat{\boldsymbol{\alpha}}_{OLSE}) = \hat{\sigma}^2 \sum_{j=1}^{p} \frac{1}{\lambda_j} \tag{21}$$

The settings of the simulation are as follows;

- $n = 50, 100$
- $\rho = 0.9, 0.95, 0.99$
- $\sigma = 5, 8, 10$
- $k = k_A, k_M, k_G$

For $n = 50$, nine sets of simulation are performed by varying $\rho = 0.9, 0.95, 0.99$ and $\sigma = 5, 8, 10$. For $n = 100$, another nine sets of simulation are performed by varying those values of ρ and σ. The simulation is done 1000 times by generating new pseudo-random numbers for each of these eighteen simulation sets.

For each replicate in the simulation, the MSE of OLSE and the MSE of KAURE corresponding to the proposed methods of the biasing parameter (k_A, k_M and k_G) for KAURE, are obtained. In this study, the average mean squared error of the regression estimators is used as the performance evaluation criteria. Average mean squared error is abbreviated as AMSE. The equations for AMSE of KAURE, AMSE of OLSE and their Ratio, are defined in Equation (22), Equation (23) and Equation (24), respectively.

$$\text{AMSE(KAURE)} = \frac{1}{1000}\sum_{l=1}^{1000}\left[\mathbf{MSE}(\hat{\boldsymbol{\alpha}}_{KAURE})_l\right], \tag{22}$$

$$\text{AMSE(OLSE)} = \frac{1}{1000}\sum_{l=1}^{1000}\left[\mathbf{MSE}(\hat{\boldsymbol{\alpha}}_{OLSE})_l\right], \tag{23}$$

$$\text{Ratio} = \frac{\text{AMSE(OLSE)}}{\text{AMSE(KAURE)}}, \tag{24}$$

where $\mathbf{MSE}(\hat{\boldsymbol{\alpha}}_{KAURE})_l$ denotes the MSE of KAURE at the l-th simulation and $\mathbf{MSE}(\hat{\boldsymbol{\alpha}}_{OLSE})_l$ denotes the MSE of OLSE at the l-th simulation.

3.2 Simulation Results and Discussion

For $n = 50$, nine sets of simulation were performed with the parameters as follows:

Set 1: $(n = 50, \rho = 0.90, \sigma = 5)$, Set 2: $(n = 50, \rho = 0.90, \sigma = 8)$,

Set 3: $(n = 50, \rho = 0.90, \sigma = 10)$, Set 4: $(n = 50, \rho = 0.95, \sigma = 5)$,

Set 5: $(n = 50, \rho = 0.95, \sigma = 8)$, Set 6: $(n = 50, \rho = 0.95, \sigma = 10)$,

Set 7: $(n = 50, \rho = 0.99, \sigma = 5)$, Set 8: $(n = 50, \rho = 0.99, \sigma = 8)$,

Set 9: $(n = 50, \rho = 0.99, \sigma = 10)$.

Table 1 summarizes the simulated results of AMSE(KAURE), AMSE(OLSE) and the ratio of AMSE(OLSE) over AMSE(KAURE) for simulation Set 1 to Set 9.

Table 1: Monte Carlo simulation results for $n = 50$

Set	ρ	σ	k	AMSE(KAURE)	AMSE(OLSE)	Ratio
1	0.90	5	k_A	0.089482	0.103548	1.157
			k_M	0.070790	0.103548	1.463
			k_G	0.070864	0.103548	1.461
2	0.90	8	k_A	0.105057	0.158491	1.509
			k_M	0.092525	0.158491	1.713
			k_G	0.091018	0.158491	1.741
3	0.90	10	k_A	0.110857	0.176610	1.593
			k_M	0.099909	0.176610	1.768
			k_G	0.097515	0.176610	1.811
4	0.95	5	k_A	0.138268	0.190414	1.377
			k_M	0.128205	0.190414	1.485
			k_G	0.120650	0.190414	1.578
5	0.95	8	k_A	0.177140	0.294120	1.660
			k_M	0.175498	0.294120	1.676
			k_G	0.164095	0.294120	1.792
6	0.95	10	k_A	0.189815	0.337265	1.777
			k_M	0.191437	0.337265	1.762
			k_G	0.180205	0.337265	1.872
7	0.99	5	k_A	0.633789	0.895387	1.413
			k_M	0.683643	0.895387	1.310
			k_G	0.633229	0.895387	1.414
8	0.99	8	k_A	0.937401	1.412004	1.506
			k_M	0.979219	1.412004	1.442
			k_G	0.928967	1.412004	1.520
9	0.99	10	k_A	0.980109	1.606567	1.639
			k_M	1.044154	1.606567	1.539
			k_G	0.984066	1.606567	1.633

For $n = 100$, another nine sets of simulation were performed with the parameters as follows:
Set 10: $(n = 100, \rho = 0.90, \sigma = 5)$, Set 11: $(n = 100, \rho = 0.90, \sigma = 8)$,
Set 12: $(n = 100, \rho = 0.90, \sigma = 10)$, Set 13: $(n = 100, \rho = 0.95, \sigma = 5)$,
Set 14: $(n = 100, \rho = 0.95, \sigma = 8)$, Set 15: $(n = 100, \rho = 0.95, \sigma = 10)$,
Set 16: $(n = 100, \rho = 0.99, \sigma = 5)$, Set 17: $(n = 100, \rho = 0.99, \sigma = 8)$,
Set 18: $(n = 100, \rho = 0.99, \sigma = 10)$.

The simulated results of AMSE(KAURE), AMSE(OLSE) and the ratio of AMSE(OLSE) over AMSE(KAURE) for simulation Set 10 to Set 18 are summarized in Table 2.

Table 2: Monte Carlo simulation results for $n = 100$

Set	ρ	σ	k	AMSE(KAURE)	AMSE(OLSE)	Ratio
10	0.90	5	$\hat{k}_A$	0.061378	0.049692	0.810
			$\hat{k}_M$	0.037985	0.049692	1.308
			$\hat{k}_G$	0.042391	0.049692	1.172
11	0.90	8	$\hat{k}_A$	0.062303	0.075989	1.220
			$\hat{k}_M$	0.050377	0.075989	1.508
			$\hat{k}_G$	0.050247	0.075989	1.512
12	0.90	10	$\hat{k}_A$	0.061603	0.085305	1.385
			$\hat{k}_M$	0.052384	0.085305	1.628
			$\hat{k}_G$	0.051793	0.085305	1.647
13	0.95	5	$\hat{k}_A$	0.030050	0.091829	1.147
			$\hat{k}_M$	0.066792	0.091829	1.375
			$\hat{k}_G$	0.065739	0.091829	1.397
14	0.95	8	$\hat{k}_A$	0.098781	0.140830	1.426
			$\hat{k}_M$	0.093011	0.140830	1.514
			$\hat{k}_G$	0.087280	0.140830	1.614
15	0.95	10	$\hat{k}_A$	0.098341	0.161385	1.641
			$\hat{k}_M$	0.100916	0.161385	1.599
			$\hat{k}_G$	0.093006	0.161385	1.735
16	0.99	5	$\hat{k}_A$	0.319133	0.433916	1.360
			$\hat{k}_M$	0.346713	0.433916	1.252
			$\hat{k}_G$	0.314766	0.433916	1.379
17	0.99	8	$\hat{k}_A$	0.464602	0.671173	1.446
			$\hat{k}_M$	0.500986	0.671173	1.341
			$\hat{k}_G$	0.461844	0.671173	1.454
18	0.99	10	$\hat{k}_A$	0.506189	0.770699	1.523
			$\hat{k}_M$	0.548304	0.770699	1.406
			$\hat{k}_G$	0.507490	0.770699	1.519

Figure 1 to Figure 6 present the graphs of AMSE(KAURE) and AMSE(OLSE) for different values of ρ based on the simulation results when $n = 50$ and $\sigma = 5$ (Figure 1), $n = 50$ and $\sigma = 8$ (Figure 2), $n = 50$ and $\sigma = 10$ (Figure 3), $n = 100$ and $\sigma = 5$ (Figure 4), $n = 100$ and $\sigma = 8$ (Figure 5), $n = 100$ and $\sigma = 10$ (Figure 6). It is observed that AMSE across all regression estimators tends to increase when ρ increases. The explanatory variables in the simulation sets are generated using the Equation (17). A higher value of ρ in the Equation (17) implies an increase in the degree of multicollinearity. For fixed values of n and σ, the AMSE of OLSE and the AMSE of KAURE for all biasing parameters (k_A, k_M and k_G) display a notable increasing trend when the values of ρ increase from 0.90 to 0.99. This finding is consistent with other studies [2–3], which indicate that multicollinearity often reduces estimation accuracy of regression estimators.

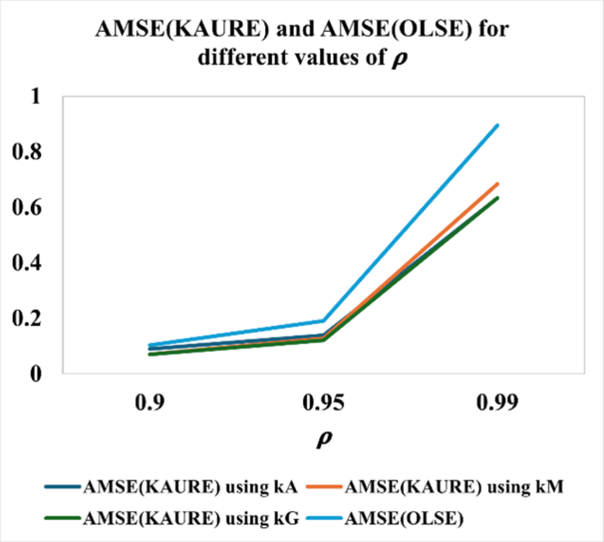


Figure 1: For $n = 50$ and $\sigma = 5$

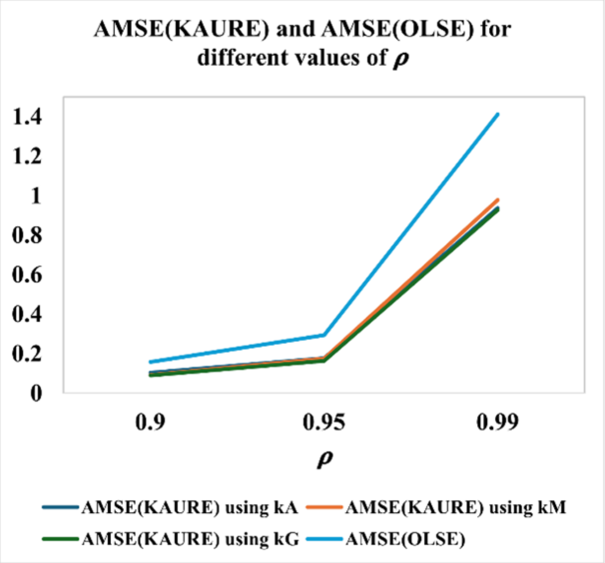


Figure 2: For $n = 50$ and $\sigma = 8$

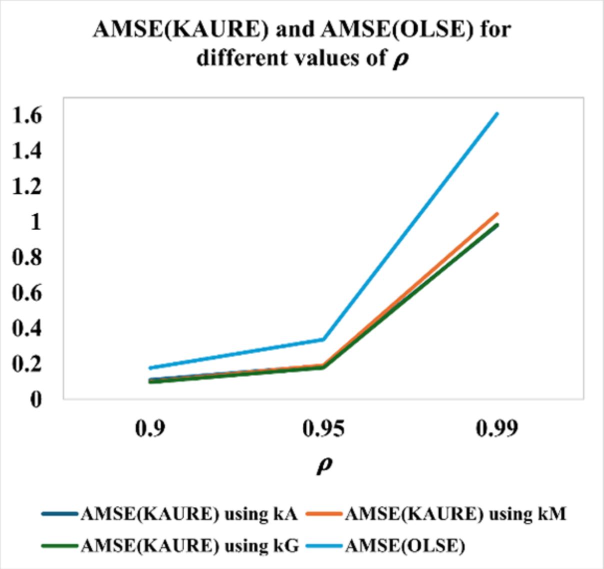


Figure 3: For $n = 50$ and $\sigma = 10$

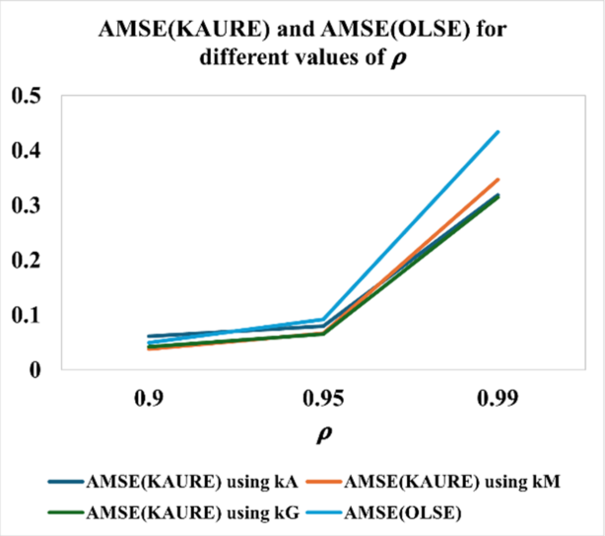


Figure 4: For $n = 100$ and $\sigma = 5$

In the comparison of regression estimator performance, the simulation results show that KAURE outperformed OLSE in terms of AMSE for all biasing parameters (k_A, k_M and k_G)

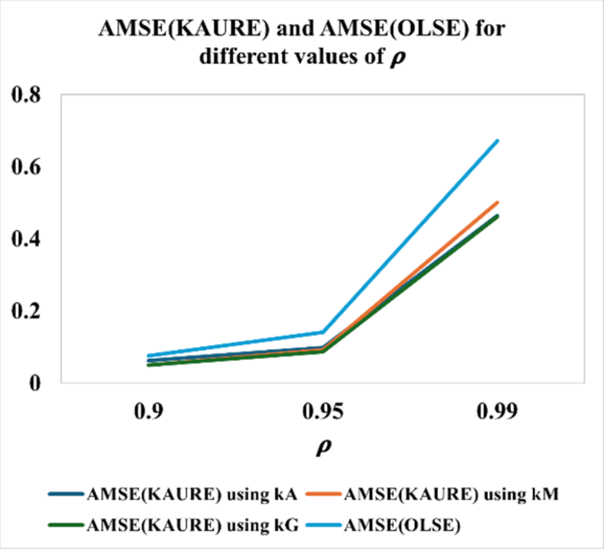


Figure 5: For $n = 100$ and $\sigma = 8$

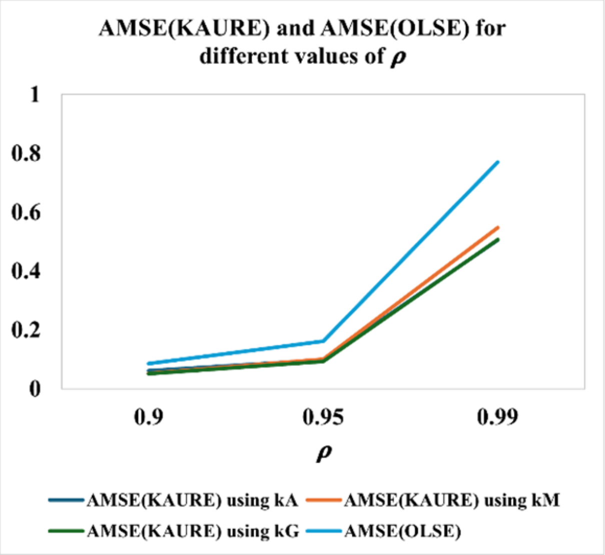


Figure 6: For $n = 100$ and $\sigma = 10$

in seventeen out of eighteen simulation sets, consistently achieving lower AMSE values than OLSE.

Based on the simulated results, Figure 7 to Figure 12 present the graphs of AMSE(KAURE) and AMSE(OLSE) for different values of σ when $n = 50$ and $\rho = 0.90$ (Figure 7), $n = 50$ and $\rho = 0.95$ (Figure 8), $n = 50$ and $\rho = 0.99$ (Figure 9), $n = 100$ and $\rho = 0.90$ (Figure 10), $n = 100$ and $\rho = 0.95$ (Figure 11), $n = 100$ and $\rho = 0.99$ (Figure 12). It is observed that there is an increase in the AMSE across all regression estimators when σ increases. For fixed values of n and ρ, the AMSE of OLSE and the AMSE of KAURE for all biasing parameters (k_4, k_M and k_G) show an increasing trend when the values of σ increase from 5 to 10. This indicates that lower data variability leads to better estimation accuracy across all estimators. This finding is consistent with the study in Jegede *et al.* [7].

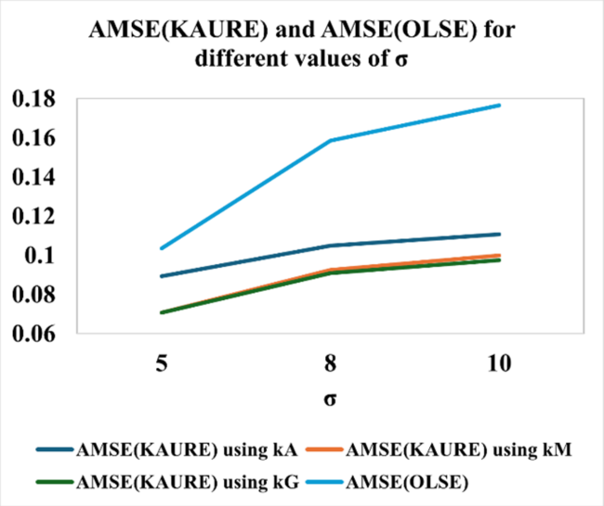


Figure 7: For $n = 50$ and $\rho = 0.90$

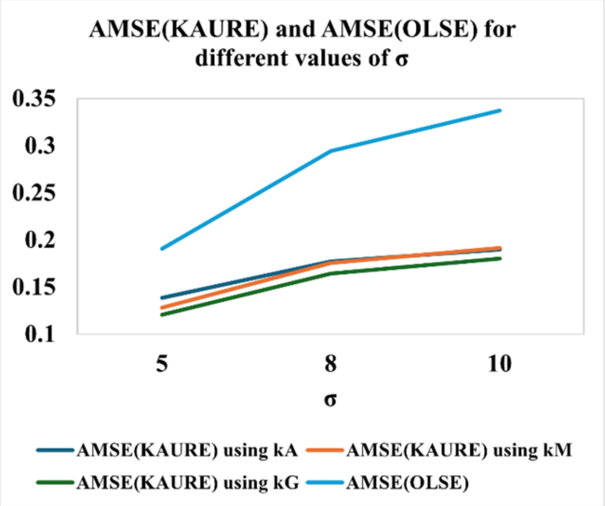


Figure 8: For $n = 50$ and $\rho = 0.95$

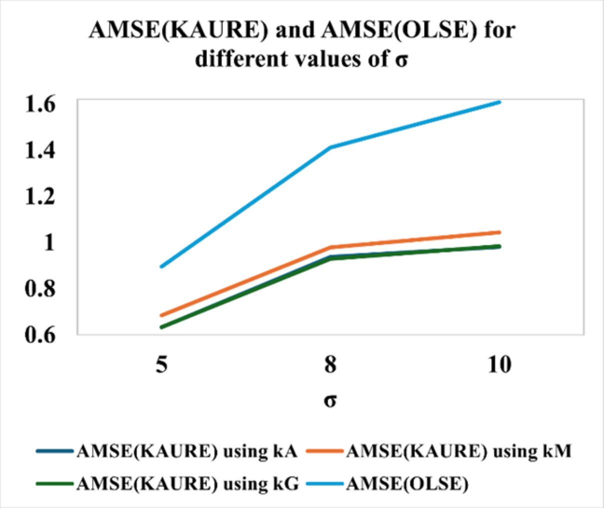


Figure 9: For $n = 50$ and $\rho = 0.99$

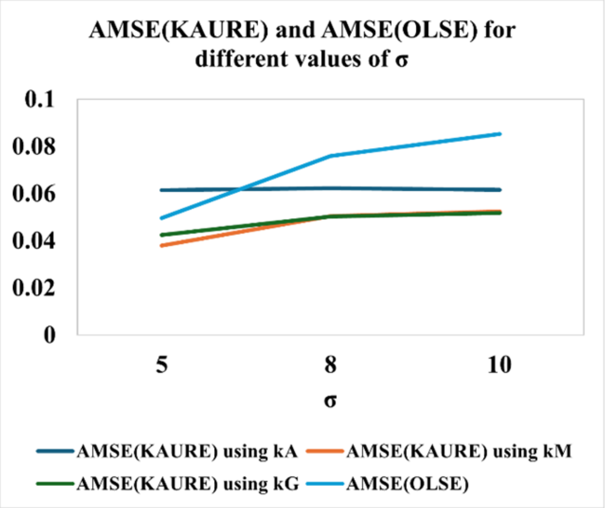


Figure 10: For $n = 100$ and $\rho = 0.90$

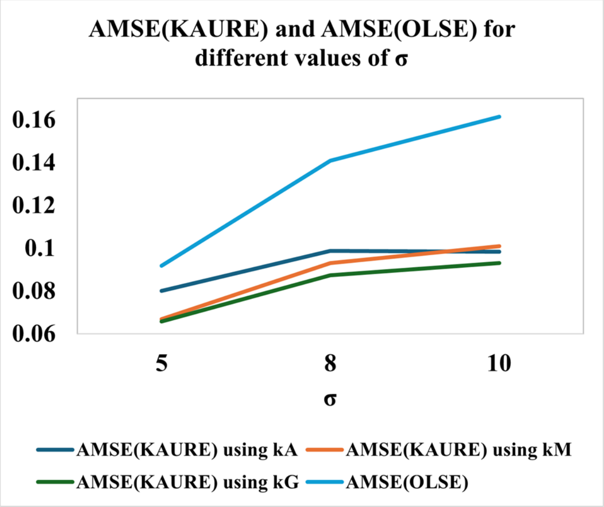


Figure 11: For $n = 100$ and $\rho = 0.95$

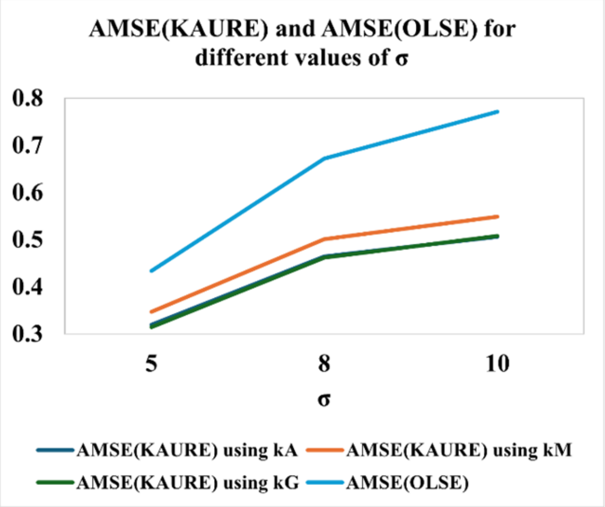


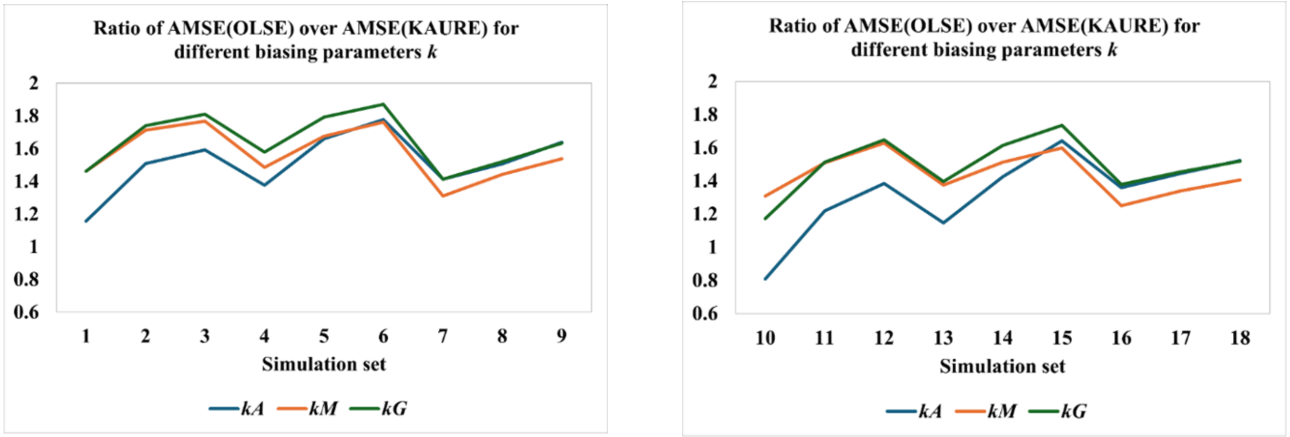
Figure 12: For $n = 100$ and $\rho = 0.99$

Comparing the performance of the regression estimators, the simulation results showed that KAURE corresponding to all biasing parameters (k_4, k_M and k_G) outperformed OLSE in terms of AMSE in 17 out of a total of 18 simulation sets. In the particular simulation set ($n = 100$, $\rho = 0.90$, $\sigma = 5$), OLSE outperformed KAURE corresponding to the biasing parameter k_4. However, the performance of KAURE associated with the biasing parameters k_G and k_M were better than OLSE in terms of AMSE (Figure 10).

The ratio of AMSE(OLSE) over AMSE(KAURE) was examined in order to compare the performance of the proposed methods (k_4, k_M and k_G) to estimate the biasing parameter for KAURE. The method that yields the highest ratio is identified as the optimal choice for estimating the biasing parameter for KAURE.

Figure 13 presents the simulated results of the Ratio $= \dfrac{\text{AMSE(OLSE)}}{\text{AMSE(KAURE)}}$ corresponding to various biasing parameters for KAURE across all eighteen simulation sets. The results indicated that the AMSE(KAURE) for the biasing parameter k_G achieved the highest ratio in 14 of the simulation sets. In contrast, k_4 produced the highest ratio in 2 simulation sets, while k_M also had the highest ratio in 2 simulation sets.

Therefore, it is recommended that $k_G = \text{Geometric Mean}\,[f_j]$ in Equation (15) be chosen as the optimal biasing parameter for KAURE.



(a) (b)

Figure 13: Figure 13: Ratio of AMSE(OLSE) over AMSE(KAURE) corresponding to the proposed methods (k_4, k_M, and k_G) to estimate the biasing parameter for KAURE when (a) $n = 50$ (simulation Set 1 to Set 9) and (b) $n = 100$ (simulation Set 10 to Set 18).

4 Numerical Example

To empirically apply the proposed methods to estimate the optimal biasing parameter for KAURE, we used the Portland cement data that was originally from Woods *et al.* [32]. This dataset has since been widely analyzed by researchers such as [1, 2, 9, 11, 33]. The linear regression model for the data includes one dependent variable (heat evolved after 180 days of curing) and $p = 4$ independent variables (Tricalcium aluminate, Tricalcium silicate, Tetracalcium aluminoferrite, and β-dicalcium silicate).

Standardization was done on the dependent variable and independent variables. The canonical form of the linear regression model was represented by $\mathbf{Y} = \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$. From this dataset, we obtained four eigenvalues (λ_j), four OLSE of the parameters ($\hat{\alpha}_j$), and estimated variance $\hat{\sigma}^2 = 0.002$. The data and computational formulations are detailed in Appendix A.

Based on Equation (13), Equation (14), and Equation (15), the estimated value of f_j is given by

$$\hat{f}_j = \frac{1 + \lambda_{\max}\sqrt{\dfrac{\hat{\sigma}^2}{\hat{\sigma}^2 + \lambda_{\max}\hat{\alpha}_j^2}}}{1 - \sqrt{\dfrac{\hat{\sigma}^2}{\hat{\sigma}^2 + \lambda_{\max}\hat{\alpha}_j^2}}}. \tag{25}$$

The values of λ_j, $\hat{\alpha}_j$, and $\hat{f}_j$ are given in Table 3.

Table 3: The values of λ_j, $\hat{\alpha}_j$ and $\hat{f}_j$

j	λ_j	$\hat{\alpha}_j$	$\hat{f}_j$
1	2.235704	-1.198979	1.082756
2	1.576066	-0.018413	19.56334
3	0.186606	-1.549323	1.063682
4	0.001624	0.573396	1.177814

Table 4 presents the results of the proposed methods for determining the biasing parameter for KAURE, along with the corresponding MSE of KAURE. The equation of $\hat{f}_j$ in Table 4 is given by Equation 25.

Table 4: Biasing parameter for KAURE and the corresponding MSE($\hat{\boldsymbol{\alpha}}_{KAURE}$)

Proposed methods	Estimated biasing parameter k for KAURE	**MSE($\hat{\boldsymbol{\alpha}}_{KAURE}$)**
$k_A = \text{Arithmetic Mean}\left[\hat{f}_j\right] = \frac{1}{p}\sum_{j=1}^{p}\hat{f}_j$	$k_A = 5.722$	1.437540
$k_M = \text{Median}\left[\hat{f}_j\right]$	$k_M = 1.130$	1.212067
$k_G = \text{Geometric Mean}\left[\hat{f}_j\right] = \left(\prod_{j=1}^{p}\hat{f}_j\right)^{\frac{1}{p}}$	$k_G = 2.270$	0.802176

Figure 14 presents the graph of **MSE($\hat{\boldsymbol{\alpha}}_{KAURE}$)** versus the biasing parameter k. The minimum value of **MSE($\hat{\boldsymbol{\alpha}}_{KAURE}$)** occurs when k is between 2 and 3.

Comparing the three proposed methods (k_A, k_M and k_G), we find that $k_G = 2.270$ falls between 2 and 3. The **MSE** of KAURE with $k_G = 2.270$ is the lowest among these methods, yielding an **MSE** of 0.802176. Therefore, k_G is selected as the optimal biasing parameter for KAURE, a conclusion that is further supported by the simulation results presented in Section 3. Moreover, it is worth noting that **MSE**($\hat{\boldsymbol{\alpha}}_{KAURE}$) = 0.802176 corresponding to $k_G = 2.270$ is less than **MSE**($\hat{\boldsymbol{\alpha}}_{OLSE}$) $= \hat{\sigma}^2 \sum_{j=1}^{4} \frac{1}{\lambda_j} = 1.244601$. Therefore, the superiority of KAURE over OLSE remains valid when k_G is selected as the optimal biasing parameter.

5 Discussion on Industrial Applications

Some industrial applications of linear regression are discussed in this section. In various industrial applications, multicollinearity, a condition where two or more independent variables are

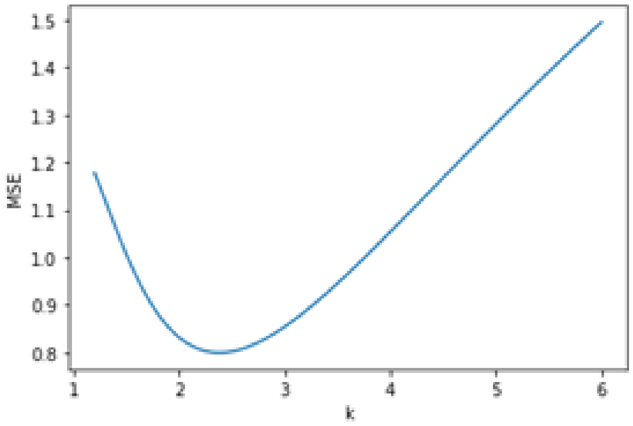


Figure 14: $\mathbf{MSE}(\hat{\boldsymbol{\alpha}}_{KAURE})$ versus the biasing parameter k

highly correlated, often arises. Multicollinearity is an inherent imperfection in data resulting from uncontrollable processes through the data-generating mechanism.

For example, linear regression is often used in environmental and climate modeling. In predicting air pollution levels, independent variables such as weather conditions, traffic volume, population density, land use and industrial emissions are often correlated. Some studies related to air pollution modeling using regression are [34–36]. In economic and financial modeling, multicollinearity frequently arises because independent variables are often interrelated. In predicting economic growth, independent variables such as investment, debt, interest rate, exchange rate, consumption, and exports are often correlated. Related study can be seen in [37].

In healthcare, multicollinearity arises when analyzing medical outcomes or genetic data. In predicting disease risk, lifestyle factors and biometric indicators such as blood pressure, cholesterol levels and diet are often intercorrelated. The results reported in [38] highlighted the adverse effects of multicollinearity in regression analysis conducted in epidemiologic studies. Multicollinearity is a common challenge in many industrial applications of linear regression, leading to unstable and unreliable coefficient estimates. While OLSE are unbiased, their high variance in the presence of multicollinearity necessitates the use of biased estimators. By using biased estimators in regression modeling, these techniques effectively reduce mean square eror of regression estimators, improve the accuracy in parameter estimation and identification of important variables in modeling, making them essential tools in modern statistical modeling.

6 Conclusion

In this paper, we introduced some new methods for estimating the optimal biasing parameter for k-almost unbiased regression estimator (KAURE) that minimizes its MSE. Extensive Monte Carlo simulations were conducted to assess the performance of estimators based on the AMSE criterion by varying factors such as sample size, error standard deviation and degree of multicollinearity. Among the proposed methods (k_A, k_M and k_G), the results indicated that KAURE showed the best performance when the biasing parameter k_G was utilized. Additionally, these simulation findings were confirmed by the empirical application. Based on both the simulation results and real-world application, we conclude that the biasing parameter k_G is proposed as the optimal choice for practitioners seeking to effectively apply KAURE as a regression estimator to address multicollinearity issues in linear regression models.

Appendix A

This Appendix provides the Portland Cement data. The dependent variable consists of heat evolved after 180 days of curing (T). The independent variables consist of Tricalcium aluminate (S_1), Tricalcium silicate (S_2), Tetracalcium aluminoferrite (S_3) and β-dicalcium silicate (S_4). Table A1 presents the Portland Cement data.

Table A1: Data

T	S_1	S_2	S_3	S_4
78.5	7	26	6	60
74.3	1	29	15	52
104.3	11	56	8	20
87.6	11	31	8	47
95.9	7	52	6	33
109.2	11	55	9	22
102.7	3	71	17	6
72.5	1	31	22	44
93.1	2	54	18	22
115.9	21	47	4	26
83.8	1	40	23	34
113.3	11	66	9	12
109.4	10	68	8	12

The standardized dependent variable and the standardized independent variables are obtained from the Equation (26) and Equation (27), respectively.

$$y_i = \frac{t_i - \bar{t}}{\sqrt{\sum\limits_{i=1}^{n}(t_i - \bar{t})^2}}, \tag{26}$$

$$x_{ij} = \frac{s_{ij} - \bar{s}_j}{\sqrt{\sum_{i=1}^{n} (s_{ij} - \bar{s}_j)^2}}, \tag{27}$$

where $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, p$, $n = 13$, $p = 4$, $\bar{t}$ is the mean of T and $\bar{s}_j$ is the mean of S_j.

The linear regression model that is formed by the standardized variables is represented by $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. The regression model is then transformed into a canonical form $\mathbf{Y} = \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$, where $\mathbf{Z} = \mathbf{XQ}$, $\boldsymbol{\alpha} = \mathbf{Q}'\boldsymbol{\beta}$, $\mathbf{Z}'\mathbf{Z} = \mathbf{\Lambda}$ and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$.

Multicollinearity diagnostic analysis is done by evaluating variance inflation factor VIF_j and condition index CI_j. The values of VIF_j are obtained from the diagonal element of matrix $(\mathbf{X}'\mathbf{X})^{-1}$. In this dataset, VIF_j are 38.496211, 254.423162, 46.868386 and 282.512861. All values of VIF_j are greater than 10, indicating the existence of multicollinearity in the dataset. The values of CI_j are obtained from $CI_j = \sqrt{\frac{\lambda_{\max}}{\lambda_j}}$. In this dataset, CI_j are 1, 1.191022, 3.461339 and 37.106342. The largest condition index is higher than 30, indicating the existence of moderate to strong dependencies among the independent variables in the dataset.

Acknowledgments

The author would like to thank Universiti Teknologi MARA for the support in research and publication.

References

[1] Ng, S. F. An almost unbiased regression estimator: Theoretical comparison and numerical comparison in Portland cement data. *MATEMATIKA*. 2023. 39(3): 315-327.

[2] Lukman, A. F., Ayinde, K., Binuomote, S., and Clement, O. A. Modified ridgetype estimator to combat multicollinearity: Application to chemical data. *Journal of Chemometrics*. 2019. 33(5): 1-12.

[3] Mermi, S., Gkta, A., and Akku, . Are most proposed ridge parameter estimators skewed and do they have any effect on MSE values? *Journal of Statistical Computation and Simulation*. 2021. 91(10): 2074-2093.

[4] Belsley, D. A. *Conditioning Diagnostics: Collinearity and Weak Data in Regression.* New York: John Wiley & Sons. 1991.

[5] Adnan, N., Ahmad, M. H., and Adnan, R. A comparative study on some methods for handling multicollinearity problems. *MATEMATIKA*. 2006. 22(2): 109-119.

[6] Rawlings, J. O., Pantula, S. G., and Dickey, D. A. *Applied Regression Analysis - A Research Tool.* New York: Springer-Verlag. 1998.

[7] Jegede, S. L., Lukman, A. F., Alqasem, O. A., Abd Elwahab, M. E., Ayinde, K., Kibria, B. G., and Adewinbi, H. Handling linear dependency in linear regression models: Almost unbiased modified ridge-type estimator. *Scientific African*. 2024. e02324.

[8] Dawoud, I. Modified two parameter regression estimator for solving the multicollinearity. *Thailand Statistician*. 2022. 20(4): 842-859.

[9] Kibria, B. M., and Lukman, A. F. A new ridge-type estimator for the linear regression model: Simulations and applications. *Scientifica*. 2020. 1-16.

[10] Akdeniz, F., and Erol, H. Mean squared error matrix comparisons of some biased estimators in linear regression. *Communications in Statistics - Theory and Methods*. 2003. 32(12): 2389-2413.

[11] Liu, K. Using Liu-type estimator to combat collinearity. *Communications in Statistics - Theory and Methods*. 2003. 32(5): 1009-1020.

[12] Liu, K. A new class of biased estimate in linear regression. *Communications in Statistics - Theory and Methods*. 1993. 22(2): 393-402.

[13] Sarkar, N. A new estimator combining the ridge regression and the restricted least squares methods of estimation. *Communications in Statistics - Theory and Methods*. 1992. 21(7): 1987-2000.

[14] Greenberg, E. Minimum variance properties of principal component regression. *Journal of the American Statistical Association*. 1975. 70: 194-197.

[15] Marquardt, D. W. Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation. *Technometrics*. 1970. 12(3): 591-612.

[16] Massy, W. F. Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*. 1965. 60: 234-246.

[17] Hawkins, D. M. On the investigation of alternative regressions by principal component analysis. *Applied Statistics*. 1973. 22: 275-286.

[18] Trenkler, G. An iteration estimator for the linear model. *Compstat*. 1978. 125-131.

[19] Hoerl, A. E., and Kennard, R. W. Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics*. 1970. 12(1): 55-67.

[20] Stein, C. M. Multiple regression in Contributions to Probability and Statistics, ed. I. Olkin. Stanford CA: Stanford University Press. 1960.

[21] Lawless, J. F., and Wang, P. A simulation study of ridge and other regression estimators. *Communications in Statistics - Theory and Methods*. 1976. 5(4): 307-323.

[22] Nomura, M. On the almost unbiased ridge regression estimator. *Communications in Statistics - Simulation and Computation*. 1988. 17(3): 729-743.

[23] Troskie, C. G., and Chalton, D. A Bayesian estimate for the constants in ridge regression. *South African Statistical Journal*. 1996. 30(2): 119-137.

[24] Firinguetti, L. A generalized ridge regression estimator and its finite sample properties: A generalized ridge regression estimator. *Communications in Statistics - Theory and Methods*. 1999. 28(5): 1217-1229.

[25] Kibria, B. G. Performance of some new ridge regression estimators. *Communications in Statistics - Simulation and Computation*. 2003. 32(2): 419-435.

[26] Khalaf, G., and Shukur, G. Choosing ridge parameter for regression problems. *Communications in Statistics - Theory and Methods*. 2005. 34(5): 1177-1182.

[27] Dorugade, A. V. New ridge parameters for ridge regression. *Journal of the Association of Arab Universities for Basic and Applied Sciences*. 2014. 15: 94-99.

[28] Goktas, A., Akkus, O., and Kuvat, A. A new robust ridge parameter estimator based on search method for linear regression model. *Journal of Applied Statistics*. 2020. 48(13-15): 2457-2472.

[29] Jacob, J., and Varadharajan, R. Robust variance inflation factor: A promising approach for collinearity diagnostics in the presence of outliers. *Sankhya B*. 2024. 1-27.

[30] Hoerl, A. E., Kannard, R. W. and Baldwin, K. F. Ridge regression: Some simulations. *Communications in Statistics - Theory and Methods*. 1975. 4(2): 105-123.

[31] Xu, J. and Yang, H. More on the bias and variance comparisons of the restricted almost unbiased estimators. *Communications in Statistics - Theory and Methods*. 2011. 40(22): 4053-4064.

[32] Woods, H., Steinour, H. H. and Starke, H. R. Effect of composition of Portland cement on heat evolved during hardening. Industrial & Engineering Chemistry. 1932. 24(11): 1207-1214.

[33] Yang, H. and Chang, X. A new two-parameter estimator in linear regression. *Communications in Statistics - Theory and Methods*. 2010. 39(6): 923-934.

[34] Hoek, G., Beelen, R., De Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P. and Briggs, D. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment*. 2008. 42(33): 7561-7578.

[35] Mozumder, C., Reddy, K. V. and Pratap, D. Air pollution modeling from remotely sensed data using regression techniques. *Journal of the Indian Society of Remote Sensing*. 2013. 41: 269-277.

[36] Hsu, C. W., Chan, M. J., Weng, C. H., Tsai, T. Y., Yen, T. H. and Huang, W. H. Environmental PM2.5 exposure: An ignored factor associated with blood cadmium level in hemodialysis patients. *Therapeutics and Clinical Risk Management*. 2025. 1-13.

[37] Ebiwonjumi, A., Chifurira, R. and Chinhamu, K. A robust principal component analysis for estimating economic growth in Nigeria in the presence of multicollinearity and outlier. *Journal of Statistics Applications & Probability*. 2023. 12(2): 611-627.

[38] Vatcheva, K. P., Lee, M., McCormick, J. B. and Rahbar, M. H. Multicollinearity in regression analyses conducted in epidemiologic studies. *Epidemiology*. 2016, 6(2): 1-20.