Research Article

Quantile-Based Estimation of Liu Parameter in the Linear Regression Model: Applications to Portland Cement and US Crime Data

Muhammad Suhail (1,2), Iqra Babar, 1 Yousaf Ali Khan (3), 4 Muhammad Imran, and Zeeshan Nawaz

1College of Statistical and Actuarial Sciences University of the Punjab, Lahore, Pakistan
2Department of Statistics, The University of Agriculture Peshawar, Amir Muhammad Khan Campus Mardan, Peshawar, Pakistan
3Department of Mathematics and Statistics, Hazara University Mansehra, Mansehra, Pakistan
4SABIC Technology and Innovation, 2nd Industrial Area, Khraj Highway, Riyadh 11551, Saudi Arabia

Correspondence should be addressed to Yousaf Ali Khan; yousaf_hu@yahoo.com

Received 7 April 2021; Accepted 28 June 2021; Published 10 July 2021

Academic Editor: Ishfaq Ahmad

Copyright © 2021 Muhammad Suhail et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In multiple linear regression models, the multicollinearity problem mostly occurs when the explanatory variables are correlated among each other. It is well known that when the multicollinearity exists, the variance of the ordinary least square estimator is unstable. As a remedy, Liu in [1] developed a new method of estimation with biasing parameter \( d \). In this paper, we have introduced a new method to estimate the biasing parameter in order to mitigate the problem of multicollinearity. The proposed method provides the class of estimators that are based on quantile of the regression coefficients. The performance of the new estimators is compared with the existing estimators through Monte Carlo simulation, where mean squared error and mean absolute error are considered as evaluation criteria of the estimators. Portland cement and US Crime data is used as an application to illustrate the benefit of the new estimators. Based on simulation and numerical study, it is concluded that the new estimators outperform the existing estimators in certain situations including high and severe cases of multicollinearity. 95% mean prediction interval of all the estimators is also computed for the Portland cement data. We recommend the use of new method to practitioners when the problem of high multicollinearity exists among the explanatory variables.

1. Introduction

The commonly used method of estimation in multiple linear regression models (MLRM) is the method of ordinary least squares (OLS) [1]. The results obtained through this method might be misleading when the problem of multicollinearity is present among the explanatory variables [2]. To overcome such problem, Ridge regression (RR) and Liu regression suggested by [1, 3], respectively, are the two widely used alternative methods. Liu estimator (LE) is usually preferred over RR, because it is the linear function of its biasing parameter \( d \) [4].

The optimal value of biasing parameter \( d \) in Liu regression plays an important role in minimizing the variance. Many researchers have suggested several estimators for estimating \( d \). Few of them are [4–6]. In this paper, the performance of some existing LEs is investigated, and a new method called as quantile based estimation of Liu or biasing parameter \( d \) is proposed. Also, the new estimators are compared with the existing ones through a Monte Carlo simulation based on mean squared error (MSE) and mean absolute error (MAE) performance criterions. The rest of the article is written as follows. The model estimation, newly proposed, and existing LEs are discussed in Section 2.
2. Statistical Methodology

Consider the following MLRM:

\[ y = X\beta + \epsilon, \]  

where \( y \) is the vector of random response variable of order \((n \times 1)\), \( X \) shows the fixed design matrix of explanatory variables with order \((n \times p)\) and \( \beta \) is the \((p + 1) \times 1\) vector of population regression coefficients. \( \epsilon \) is the vector of stochastic or random errors with order \((n \times 1)\) and is distributed as normal with mean \( E(\epsilon) = 0 \) and variance covariance matrix \( E(\epsilon\epsilon') = \sigma^2 I_n \), \( I_n \) is an \((n \times n)\) identity matrix. The vector of OLS estimators for \( \beta \) is given below:

\[ \hat{\beta}_{\text{OLS}} = (X'X)^{-1}X'y. \]  

The OLS estimator is unbiased and more efficient than all other unbiased estimators [2]. However, in the presence of multicollinearity, OLS estimator becomes inefficient and provide large variance [4]. To circumvent such situation, numerous biased estimation methods are available, which provide smaller MSE than OLS, and LE is one of them. The Liu estimator defined by [1] is given as

\[ \hat{\beta}_{\text{LIU}} = (X'X + I)^{-1}(X'X + dI)\hat{\beta}_{\text{OLS}}, \quad 0 \leq d \leq 1. \]  

In the presence of multicollinearity, \( \hat{\beta}_{\text{LIU}} \) provides the smaller MSE than OLS [7]. The optimal choice of Liu parameter \( d \) plays a vital role in minimizing the MSE of \( \hat{\beta}_{\text{LIU}} \) [8]. Some existing LEs for the biasing parameter \( d \) are given in the following subsection.

2.1. Some Existing Liu Estimators. Consider the canonical form of model (1):

\[ y = Z\alpha + \epsilon, \]  

where \( Z = X \) and \( \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_p)' = D'\beta \). \( D \) is an orthogonal matrix such that \( D'D = I \) and \( Z'Z = D'X'X \). \( D = \Lambda \), \( \Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p) \) consists of the eigenvalues of the \( X'X \) matrix. Note here that MSE \( (\hat{\alpha}) = \text{MSE}(\hat{\beta}) \) so it suffices to consider the canonical form only. The OLS estimator can be defined in canonical form as follows:

\[ \hat{\alpha}_{\text{OLS}} = \Lambda^{-1}Z'y. \]  

The LE is defined as

\[ \hat{\alpha}_{\text{LIU}} = (\Lambda + I)^{-1}(\Lambda + dI)\hat{\alpha}_{\text{OLS}}, \]  

The first estimator for \( d \) was suggested by [1] and is given below:

\[ \tilde{d}_j = \frac{\hat{\alpha}_j^2 - \sigma^2 j}{\sigma^2 j^2 + \hat{\alpha}_j^2}, \]  

where \( \hat{\alpha}_j \) is the \( j \)th element of \( \hat{\alpha} \), an OLS estimator of \( \alpha \). \( \hat{\alpha}^2 \) is the unbiased estimator of population error variance \( \sigma^2 \) and \( \lambda_j \) is the \( j \)th eigenvalue of the matrix \( X'X \). Liu in [1] also suggested the following estimator:

\[ D1 = \max \left( 0, \min \left( \frac{\hat{\alpha}_{\max}^2 - \sigma^2 \lambda_{\max}}{1/(1/\lambda_{\max}) + \hat{\alpha}_{\max}^2} \right) \right). \]  

Shukur et al. [6] considered the idea of [5, 9] and suggested the following four estimators:

\[ D2 = \max \left( 0, \min \left( \frac{\hat{\alpha}_{\max}^2 - \sigma^2 \lambda_{\max}}{1/(1/\lambda_{\max}) + \hat{\alpha}_{\max}^2} \right) \right). \]

\[ D3 = \max \left( 0, \min \left( \frac{\hat{\alpha}_{\max}^2 - \sigma^2 \lambda_{\max}}{1/(1/\lambda_{\max}) + \hat{\alpha}_{\max}^2} \right) \right). \]

\[ D4 = \max \left( 0, \min \left( \frac{\hat{\alpha}_{\max}^2 - \sigma^2 \lambda_{\max}}{1/(1/\lambda_{\max}) + \hat{\alpha}_{\max}^2} \right) \right). \]

\[ D5 = \max \left( 0, \min \left( \frac{\hat{\alpha}_{\max}^2 - \sigma^2 \lambda_{\max}}{1/(1/\lambda_{\max}) + \hat{\alpha}_{\max}^2} \right) \right). \]

Based on the work of [10, 11], we proposed five new quantile based LEs, which are defined in the following section.

2.2. Proposed Method. Let \( \tilde{d}_1, \tilde{d}_2, \ldots, \tilde{d}_p \) be the realized values of equation (7) and it can be written in the ascending order of magnitude as

\[ \tilde{d}_{(1)} \leq \tilde{d}_{(2)} \leq \cdots \leq \tilde{d}_{(p)}, \]  

where \( \tilde{d}_{(1)} = \min(\tilde{d}_1, \tilde{d}_2, \ldots, \tilde{d}_p) \) and \( \tilde{d}_{(p)} = \max(\tilde{d}_1, \tilde{d}_2, \ldots, \tilde{d}_p) \). The set \( \{\tilde{d}_{(1)}, \tilde{d}_{(2)}, \ldots, \tilde{d}_{(p)}\} \) is the order statistics for \( (\tilde{d}_1, \tilde{d}_2, \ldots, \tilde{d}_p) \) and \( \tilde{d}_{(j)}, j = 1, 2, \ldots, p \), is the \( j \)th ordered observation. Now let \( \tilde{d}_y, 0 < y < 1 \), be the 100\( y \)th quantile of \( \{\tilde{d}_{(1)}, \tilde{d}_{(2)}, \ldots, \tilde{d}_{(p)}\} \), and then the new proposed quantile estimator is

\[ \tilde{\alpha}_y = \tilde{\alpha}_{(j)} = \frac{\left( \tilde{\alpha}_{(j)}^2 \right)_y - \sigma^2 j}{\sigma^2 j^2 + \left( \tilde{\alpha}_{(j)}^2 \right)_y}, \]  

such that

\[ P(\tilde{\alpha}_y < \tilde{d}_y) = y, \]  

where “\( y \)” is the quantile probability, \( \tilde{\alpha}_{(j)} \) is the \( j \)th ordered element of \( \tilde{\alpha} \), an OLS estimator of \( \alpha \), and \( \lambda_{(j)} \) is the \( j \)th ordered eigenvalue of the matrix \( X'X \). The quantile
probability generally depends on the degree of multicollinearity in order to obtain the minimum MSE or MAE. The new estimator $\hat{d}_y$ defined in equation (11) depends on the quantile probability whose value is selected according to the degrees of multicollinearity. So the new proposed estimator is more robust to high or severe degrees of multicollinearity. Since the range of LE must be between zero and one, therefore we rewrite the equation (11) as

$$D_y = \max(0, \hat{d}_y) = \max \left(0, \frac{\left(\hat{a}_{(j)} - \hat{a}_{(j)}^2 \right)}{\left(\hat{a}_{(j)}^2 \lambda_{(j)} + \hat{a}_{(j)}^2 \right)} \right).$$

Equation (13) satisfies the interval condition for Liu parameter $d$ suggested by [1]. In order to present the role of quantile probability, we choose some specific values for “y” as: 0 (minimum), 0.25 (first quartile), 0.50 (median), 0.75 (third quartile) and 1 (maximum). The mathematical form of new LEs, denoted hereby D6, D7, D8, D9, and D10, is given below:

$$D6 = \max(0, \hat{d}_{0.0}) = \max \left(0, \frac{\left(\hat{a}_{(j)}^2 \right)_{0.0} - \hat{a}_{(j)}^2}{\left(\hat{a}_{(j)}^2 \lambda_{(j)} + \hat{a}_{(j)}^2 \right)_{0.0}} \right),$$

$$D7 = \max(0, \hat{d}_{0.25}) = \max \left(0, \frac{\left(\hat{a}_{(j)}^2 \right)_{0.25} - \hat{a}_{(j)}^2}{\left(\hat{a}_{(j)}^2 \lambda_{(j)} + \hat{a}_{(j)}^2 \right)_{0.25}} \right),$$

$$D8 = \max(0, \hat{d}_{0.50}) = \max \left(0, \frac{\left(\hat{a}_{(j)}^2 \right)_{0.50} - \hat{a}_{(j)}^2}{\left(\hat{a}_{(j)}^2 \lambda_{(j)} + \hat{a}_{(j)}^2 \right)_{0.50}} \right),$$

$$D9 = \max(0, \hat{d}_{0.75}) = \max \left(0, \frac{\left(\hat{a}_{(j)}^2 \right)_{0.75} - \hat{a}_{(j)}^2}{\left(\hat{a}_{(j)}^2 \lambda_{(j)} + \hat{a}_{(j)}^2 \right)_{0.75}} \right),$$

$$D10 = \max(0, \hat{d}_{1.0}) = \max \left(0, \frac{\left(\hat{a}_{(j)}^2 \right)_{1.0} - \hat{a}_{(j)}^2}{\left(\hat{a}_{(j)}^2 \lambda_{(j)} + \hat{a}_{(j)}^2 \right)_{1.0}} \right).$$

(14)

3. The Design of an Experiment

This section covers the Monte Carlo simulation experiment, performance criterion measures, and results and discussion.

3.1. The Monte Carlo Simulation. The performance of LEs is compared in this section through the simulation study. Following [12–14], the explanatory variables are generated as

$$x_{ij} = (1 - p^2)^{1/2} z_{ij} + \rho z_{x_{(j)}, i}, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, p,$$

(15)

where $\rho$ is the degree or level of multicollinearity between the explanatory variables, which are given as 0.80, 0.90, 0.99, and 0.999. $z_{ij}$ are the random numbers obtained from the standard normal distribution.

The $n$ observations on the response variable are computed as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

(16)

where $\epsilon_i \sim N(0, \sigma^2)$, $\sigma^2$ is the error variance. $\beta_0$ is considered to be identically zero. Following [9], the eigenvector corresponding to maximum eigenvalue of the $X'X$ matrix is taken as the vector of regression coefficients. Following [4–6], the different factors we choose to vary in our study are given below:

Error variance: $\sigma^2 = 0.1, 1, 5, 10$
Sample size: $n = 25, 50, 75, 100, 200, 400$
Explanatory variables: $p = 4, 8, 16, 32$
Multicollinearity: $\rho = 0.80, 0.90, 0.99, 0.999$

3.2. Performance Evaluation Criteria. Following [4], MSE and MAE criteria are used to judge the performance of the different LEs. Estimated MSE (EMSE) and MAE (EMAE) are defined as

$$\text{MSE} = \frac{\sum_{i=1}^{M} (\hat{\beta}_i - \beta)^2}{M},$$
$$\text{MAE} = \frac{\sum_{i=1}^{M} |\hat{\beta}_i - \beta|}{M},$$

(17)

where $\hat{\beta}_i$ is the estimated value of $\beta$. $M$ shows the simulation runs. In this study, we choose $M = 5000$. The EMSE simulation results are presented in Tables 1–6 and EMAE in Tables 7, 8. The results are discussed in the following section.

3.3. Results and Discussion. The EMSE and EMAE values of the new and existing Liu estimators are presented in Tables 1–6 and 7, 8 respectively. The performance of the LEs is evaluated by varying the values of factors such as multicollinearity, error variance, sample size, and the explanatory variables. A general remark from the literature is that these factors have the significant effect on the simulation design. The effect of each factor on EMSE and EMAE of estimators is discussed below:

The first factor we considered is the effect of multicollinearity on the EMSE and EMAE of estimators. The EMSE and EMAE of all the estimators increase by increasing the degree of multicollinearity generally except D5–D8. Estimators D5–D8 first increase when multicollinearity increases from mild to high and decrease for severe multicollinearity. It is evident from these tables that the LE is always superior to the OLS estimator. For mild multicollinearity ($\rho = 0.80$), D5 outperforms others. Estimators D6 and D7 are the close competitors to D5. When the degree of multicollinearity is considered to be high ($\rho = 0.99$) or severe ($\rho = 0.999$), then the proposed estimators D6 and D7 outperform generally. However, the performance of D6 is good among others, because it yields lowest EMSE and EMAE when the degree of multicollinearity is high or severe.
Secondly, the effect of sample size on the EMSE and EMAE of estimators is considered. In general, the EMSE and EMAE of OLS and LE decrease by raising the sample size. The proposed estimators D6 and D7 exhibit lowest EMSE in most of the cases, while D5 remains the closest competitor to the proposed estimators.

### Table 1: EMSE with $\varepsilon \sim N(0, 1)$ and $p = 4$.

| $n$  | $\rho$ | 25                | 50                |
|------|--------|-------------------|-------------------|
| 50   | 0.80   | 0.367             | 0.886             |
| OLS  | 0.8647 | 17.903            | 169.983           |
| D1   | 0.5808 | 15.424            | 166.540           |
| D2   | 0.6464 | 0.840             | 1.646             |
| D3   | 0.5342 | 0.104             | 0.837             |
| D4   | 0.6343 | 2.957             | 21.734             |
| D5   | 0.4714 | 0.569             | 0.267             |
| D6   | 0.5933 | 0.5274            | 0.1105            |
| D7   | 0.5056 | 0.558             | 0.187             |
| D8   | 0.5104 | 0.640             | 0.939             |
| D9   | 0.5353 | 11.546            | 158.740           |
| D10  | 0.5808 | 11.134            | 156.540           |

Bold values indicate the minimum MSE.

### Table 2: EMSE with $\varepsilon \sim N(0, 1)$ and $p = 8$.

| $n$  | $\rho$ | 25                | 50                |
|------|--------|-------------------|-------------------|
| 50   | 0.80   | 0.367             | 0.886             |
| OLS  | 0.8647 | 17.903            | 169.983           |
| D1   | 0.5808 | 15.424            | 166.540           |
| D2   | 1.4005 | 5.967             | 1.487             |
| D3   | 1.5054 | 2.344             | 8.705             |
| D4   | 2.6862 | 3.275             | 17.966             |
| D5   | 1.0341 | 1.262             | 0.026             |
| D6   | 1.0015 | 1.224             | 0.106             |
| D7   | 1.1855 | 1.369             | 1.086             |
| D8   | 1.4120 | 1.495             | 1.272             |
| D9   | 1.3468 | 1.546             | 5.134             |
| D10  | 3.2642 | 8.266             | 107.831           |

Bold values indicate the minimum MSE.
In the third case, we varied the number of explanatory variables from 4 to 8. It is found that EMSE and EMAE of all the estimators increase. But the performance pattern remains the same as in the case of multicollinearity and sample size. It is seen that the increase in the EMSE of OLS estimator is relatively higher than all LEs. LE with Liu

| ρ    | OLS | D1  | D2  | D3  | D4  | D5  | D6  | D7  | D8  | D9  | D10 |
|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 0.8  | 41.4642 | 26.1760 | 137.0247 | 378.0522 | 2975.1719 | 15.7533 | 32.8340 | 82.1737 | 104.7290 | 26.1760 |
| 0.9  | 42.9732 | 39.7862 | 174.2478 | 1374.5314 | 15340.5083 | 23.2637 | 37.9098 | 109.0638 | 129.5769 | 39.7862 |
| 0.99 | 454.5090 | 403.0533 | 188.2420 | 39523.4204 | 594331.0632 | 57.2993 | 348.4940 | 54.0141 | 53.2101 | 39.7862 |
| 0.999 | 4496.5388 | 4226.9805 | 1675.7301 | 490826.3305 | 7470388.8610 | 306.0084 | 4426.9805 | 156.5435 | 490826.3305 | 403.0533 |

**Table 3: EMSE with ε ~ N(0, 5) and p = 4.**

| n   | 25   | 50   |
|-----|------|------|
| ρ   | 0.8  | 0.9  | 0.99 | 0.999 | 0.8 | 0.9 | 0.99 |
| OLS | 41.4642 | 42.9732 | 454.5090 | 4496.5388 | 18.2290 | 18.9241 | 211.7116 |
| D1  | 26.1760 | 39.7862 | 403.0533 | 4226.9805 | 13.3709 | 23.2564 | 170.5487 |
| D2  | 137.0247 | 174.2478 | 188.2420 | 1675.7301 | 20.7851 | 37.1738 | 66.7591 |
| D3  | 378.0522 | 1374.5314 | 39523.4204 | 490826.3305 | 30.1051 | 53.7916 | 8463.1000 |
| D4  | 2975.1719 | 15340.5083 | 7470388.8610 | 12751605 | 1234016391 | 19269144380 |

| ρ | OLS | D1  | D2  | D3  | D4  | D5  | D6  | D7  | D8  | D9  | D10 |
|---|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 0.8  | 6.4770 | 7.7849 | 10.7290 | 12.9579 | 35.4892 | 4.8090 | 5.8656 | 8.7665 | 9.5607 | 7.7849 |
| 0.9  | 9.1141 | 14.6692 | 20.9964 | 41.9918 | 219.4767 | 7.9526 | 10.1178 | 16.8892 | 19.1424 | 14.6692 |
| 0.99 | 99.9004 | 77.4999 | 42.3556 | 2197.7497 | 29783.5059 | 23.2637 | 22.3379 | 33.6034 | 32.5339 | 77.4999 |
| 0.999 | 1013.3213 | 950.8600 | 174.1792 | 41734.7737 | 62723.7990 | 57.2993 | 22.3379 | 19.3000 | 16.8892 | 77.4999 |

**Table 4: EMSE with ε ~ N(0, 5) and p = 8.**

| n   | 75   | 100  |
|-----|------|------|
| OLS | 23.6804 | 28.8332 | 293.2024 | 2994.0666 | 20.9489 | 22.8140 | 228.8529 |
| D1  | 19.3794 | 30.4080 | 240.8266 | 2903.6818 | 26.6420 | 32.8776 | 334.0245 |
| D2  | 47.5362 | 78.4761 | 88.1685 | 214.7922 | 27.6423 | 46.4392 | 45.1056 |
| D3  | 45.0039 | 43.3329 | 29.5639 | 5.7589 | 20.3380 | 50.1511 | 27.8849 |
| D4  | 111.9509 | 856.5399 | 261.2728 | 188.9864 | 72.1586 | 105.1502 | 44.5799 |
| D5  | 62.5967 | 164.4174 | 2542.1039 | 34753.5135 | 48.8018 | 15.3856 | 44.5799 |
| D6  | 42.9726 | 71.1838 | 68.5353 | 5.7589 | 20.3380 | 44.5799 | 27.8849 |
| D7  | 46.3333 | 80.2369 | 71.0713 | 130.2053 | 27.0422 | 46.7123 | 83.4833 |
| D8  | 2201.6875 | 12863.5533 | 16591.8930 | 45.9999 | 71.2587 | 1105.1502 | 44.5799 |
| D9  | 91.8201 | 208.6214 | 2627.6189 | 27216.2742 | 21.7582 | 32.8776 | 334.0245 |
| D10 | 7.7849 | 14.6692 | 77.4999 | 950.8600 | 4.8090 | 5.8656 | 22.8140 |

Bold values indicate the minimum MSE.
parameter D6 exhibits lowest EMSE and EMAE for both the cases of explanatory variables.

Finally, the performance of estimators is compared by varying error variance. When the error variance increases from one to five, the EMSE of all the estimators deteriorates, except the new estimator D6, which is least affected.

The concluded remarks from Tables 1–8 are that the new estimators D6 and D7 have shown better performance than
the other existing LEs in terms of smaller EMSE and EMAE, in situations of high multicollinearity disregarding the values of other factors. The proposed estimators also outperform the OLS estimator substantially. Therefore, based on EMSE and EMAE criterion, we can infer that the new estimator D6 is more efficient and is the best choice for the practitioners in the presence of high and severe multicollinearity.

| n   | 25    | 50    | 25    | 50    |
|-----|-------|-------|-------|-------|
| ρ   | 0.85  | 0.95  | 0.99  | 0.99  |
| OLS | 1.4382 | 2.0404 | 6.4201 | 19.8598 |
| D1  | 1.0641 | 1.3822 | 4.7235 | 16.1845 |
| D2  | 1.0385 | 1.1961 | 1.2746 | 1.4109 |
| D3  | 1.0467 | 1.2090 | 1.4512 | 2.0672 |
| D4  | 1.1205 | 1.3141 | 2.1683 | 4.7916 |
| D5  | **0.9892** | 1.1298 | 1.0873 | 0.6554 |
| D6  | 0.9964 | **1.1292** | **1.0482** | **0.4918** |
| D7  | 1.0260 | 1.2761 | 3.9183 | 15.5864 |
| D8  | 1.0370 | 1.1665 | 1.1648 | 1.1060 |
| D9  | 1.0611 | 1.2761 | 3.9183 | 15.5864 |
| D10 | 1.0641 | 1.3822 | 4.7235 | 16.1845 |

Table 7: EMAE with ε ~ N(0,1) and p = 4.

| n   | 75    | 100   |
|-----|-------|-------|
| ρ   | 0.85  | 0.95  |
| OLS | 0.6747 | 0.9665 |
| D1  | 0.5821 | 0.7682 |
| D2  | 0.5863 | 0.7686 |
| D3  | 0.5868 | 0.7685 |
| D4  | 0.5947 | 0.7801 |
| D5  | **0.5800** | **0.7568** |
| D6  | 0.5817 | 0.7591 |
| D7  | 0.5877 | 0.7682 |
| D8  | 0.5903 | 0.7701 |
| D9  | 0.5949 | 0.7885 |
| D10 | 0.5821 | 0.7682 |

Table 8: EMAE with ε ~ N(0,1) and p = 8.

| n   | 75    | 100   |
|-----|-------|-------|
| ρ   | 0.85  | 0.95  |
| OLS | 4.1906 | 6.1205 |
| D1  | 3.0025 | 4.4984 |
| D2  | 2.3971 | 2.6431 |
| D3  | 2.4615 | 2.7378 |
| D4  | 2.9882 | 3.3961 |
| D5  | **2.1188** | **2.3539** |
| D6  | 2.2661 | 2.4884 |
| D7  | 2.4127 | 2.5829 |
| D8  | 2.3726 | 2.6012 |
| D9  | 3.0025 | 4.4984 |
| D10 | 3.0025 | 4.4984 |

Bold values indicate the minimum MAE.
4. Applications

Since the simulation evidence is not enough to judge the performance of the proposed estimators, because the study is usually conducted assuming some ideal conditions, in practice, the ideal conditions may not be met. Therefore, contrary to the previous section, we used two real applications taken from the books of [2, 15] in order to compare the performance of our proposed estimators in practical situations.

4.1. Portland Cement Data. The first numerical example used in this study is the Portland cement dataset taken from [15] to compare the performance of new estimators in applied scenario. Some authors named this data as Hald’s or ideal conditions, in heat evolved after 180 days of curing and measured in calories/gram of cement with 40% water at 35°C (95°F). Besides, the response variable (y), the four explanatory variables considered are the clinker compounds (CALCD) defined as

- (i) $X_1$: Tricalcium aluminate (3CaO * Al₂O₃)
- (ii) $X_2$: Tricalcium silicate (3CaO * SiO₂)
- (iii) $X_3$: Tetracalcium aluminoferrite (4CaO * Al₂O₃ * Fe₂O₃)
- (iv) $X_4$: Dicalcium silicate (2CaO * SiO₂)

The model is defined as

$$ y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon \quad (18) $$

Condition number (CN) is used to measure the severity of multicollinearity among explanatory variables [23]. It can be defined as

$$ CN = \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}} \quad (19) $$

where $\lambda_{\text{max}}$ and $\lambda_{\text{min}}$ are the maximum and minimum eigenvalues of the matrix $X'X$, respectively. Following [2], a rule of thumb is that multicollinearity is moderate if the CN is between 10 and 30, high if it is between 30 and 100, and severe when it is greater than 100. The eigenvalues and CN for this dataset are computed and presented in Table 10. From Table 10, it is concluded that high multicollinearity exists among the explanatory variables. The Shapiro-Wilk (W) normality test is used to test the normality of response variable. We obtain the value test statistic $W = 0.96967$ and $p$-value = 0.8903, which shows that the response variable is normal at 5% level of significance. MSE of OLS and Liu estimators from [1] can be written as

$$ \text{MSE}(\hat{\alpha}) = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j} $$

$$ \text{MSE}(\hat{\alpha}(d)) = \sigma^2 \sum_{j=1}^{p} \left( \frac{\lambda_j + d}{\lambda_j + 1} \right)^2 + (d - 1)^2 \sum_{j=1}^{p} \frac{\alpha_j^2}{\lambda_j + 1} $$

The estimated values for $d$, coefficients, and MSE of estimators are presented in Table 11. This table shows that the LEs have smaller MSE than OLS. However, among the LEs, estimator D5 and new estimators D6 and D7 outperform others; therefore, they are highly efficient among others.

4.2. Prediction Interval. In this section, 95% mean prediction interval of all the estimators is computed from the Portland cement data. We consider the following values of explanatory variables: $X_0 = (X_{10}, X_{20}, X_{30}, X_{40}) = (23, 79, 27, 65)$.

100 $(1 - \alpha) \%$ mean prediction interval for the response variable is defined as

$$ \text{OLS: } \hat{y}_0 \pm t_{1-(\alpha/2),v} \sqrt{\text{var}(\hat{y}_0)}, $$

$$ \text{LIU: } \hat{y}_0 \pm t_{1-(\alpha/2),v} \sqrt{\text{var}(\hat{y}_0)}, $$

where $\hat{y}_0 = \hat{\beta}_1 X_{10} + \hat{\beta}_2 X_{20} + \hat{\beta}_3 X_{30} + \hat{\beta}_4 X_{40}$, $\hat{y}_0 = \bar{\beta}_1 X_{10} + \bar{\beta}_2 X_{20} + \bar{\beta}_3 X_{30} + \bar{\beta}_4 X_{40}$, $\bar{\beta}$ and $\bar{\beta}$ are the OLS and Liu

| Observation | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|-------------|------|------|------|------|----|
| 1           | 1    | 29   | 15   | 52   | 74.3 |
| 2           | 11   | 56   | 8    | 20   | 104.3 |
| 3           | 7    | 26   | 6    | 60   | 78.5 |
| 4           | 7    | 52   | 6    | 33   | 95.9 |
| 5           | 11   | 55   | 9    | 22   | 109.2 |
| 6           | 11   | 31   | 8    | 47   | 87.6 |
| 7           | 1    | 31   | 22   | 44   | 72.5 |
| 8           | 2    | 54   | 18   | 22   | 93.1 |
| 9           | 3    | 71   | 17   | 6    | 102.7 |
| 10          | 11   | 66   | 9    | 12   | 113.3 |
| 11          | 10   | 68   | 8    | 12   | 109.4 |
| 12          | 1    | 40   | 23   | 34   | 83.8 |
| 13          | 21   | 47   | 4    | 26   | 115.9 |

Table 9: Portland cement data [15].

| $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\lambda_4$ | CN |
|-------------|-------------|-------------|-------------|----|
| 2.2357      | 0.1376      | 0.0186      | 0.0016      | 37.38064 |

Parenthesis value gives the maximum percentage of first eigenvalue.
estimators, respectively. $t_{1-(α/2)}$ is the $(1 - (α/2))$ quantile from the Student’s $t$-distribution with $(ν = n − p)$ degrees of freedom. $\text{var} (\bar{y}_0) = \hat{σ}^2 X_0' (X' X)^{−1} X_0$ and $\text{var} (\hat{y}_0) = \hat{σ}^2 X_0' Q_d (X' X)^{−1} Q_d' X_0$. $Q_d = (X' X + I)^{−1} (X' X + dI)$. For details, see [2, 4]. The results for the 95% mean prediction interval are given in Table 12. It can be seen from the table that the new estimator D6 gives the best mean prediction interval among all other estimators.

### 4.3. US Crime Data

The US crime dataset is taken from the book of [2] to explain the crime rate in relation to the eight socioeconomic variables in 47 states in the United States for the year 1960. Response variable and eight explanatory variables are defined as follows:

(i) $y$ = crime rate, number of offenses reported to police per million population

(ii) $X_1$ = mean number of years of schooling times 10 for persons age 25 or older

(iii) $X_2$ = 1960 per capita expenditure on police by state and local government

(iv) $X_3$ = 1959 per capita expenditure on police by state and local government

| Estimators | $d$ | MSE | $\hat{β}_1$ | $\hat{β}_2$ | $\hat{β}_3$ | $\hat{β}_4$ |
|------------|-----|-----|-------------|-------------|-------------|-------------|
| OLS        | —   | 16.4515 | -0.65696 | -0.00831 | 0.30277 | 0.38774 |
| D1         | 0.46099 | 3.6085 | -0.64423 | -0.00808 | 0.25239 | 0.18274 |
| D2         | 0.02676 | 0.2726 | -0.63398 | -0.00790 | 0.21180 | 0.01759 |
| D3         | 0.12868 | 0.4937 | -0.63639 | -0.00795 | 0.22133 | 0.05635 |
| D4         | 0.46099 | 3.6085 | -0.64423 | -0.00808 | 0.25239 | 0.18274 |
| D5         | 0.00020 | 0.2716 | -0.63336 | -0.00789 | 0.20932 | 0.00749 |
| D6         | 0.00004 | 0.2716 | -0.63335 | -0.00789 | 0.20930 | 0.00743 |
| D7         | 0.00588 | 0.2698 | -0.63349 | -0.00789 | 0.20985 | 0.00965 |
| D8         | 0.07531 | 0.3349 | -0.63513 | -0.00792 | 0.21634 | 0.03605 |
| D9         | 0.24428 | 1.1617 | -0.63912 | -0.00799 | 0.23213 | 0.10032 |
| D10        | 0.46099 | 3.6085 | -0.64423 | -0.00808 | 0.25239 | 0.18274 |

Bold values indicate the minimum MSE.

| Estimators | Lower | Upper | Difference |
|------------|-------|-------|------------|
| OLS        | 265.5974 | 300.8197 | 566.4171 |
| D1         | 130.2433 | 136.7162 | 266.9595 |
| D2         | -21.2485 | 4.5604 | 25.8089 |
| D3         | -46.8037 | 35.5516 | 82.3553 |
| D4         | 130.2433 | 136.7162 | 266.9595 |
| D5         | -14.6542 | -3.4506 | 11.2037 |
| D6         | -14.6155 | -3.4977 | 11.1178 |
| D7         | -16.0524 | -1.7492 | 14.3033 |
| D8         | -33.4124 | 19.3140 | 52.7264 |
| D9         | -75.8272 | 70.7413 | 146.5684 |
| D10        | 130.2433 | 136.7162 | 266.9595 |

Bold values indicate the best prediction interval.

| $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\lambda_4$ | $\lambda_5$ | $\lambda_6$ | $\lambda_7$ | $\lambda_8$ | CN |
|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|----|
| 4.0851      | 1.8390      | 0.9739      | 0.5119      | 0.3522      | 0.1536      | 0.0786      | 0.0057      | 26.77096 |

Parenthesis value gives the maximum percentage of first eigenvalue.

| Estimators | US crime data |
|------------|---------------|
| OLS        | 72.5334 |
| D1         | 35.5803 |
| D2         | 3.2247 |
| D3         | 6.9318 |
| D4         | 3.0783 |
| D5         | 3.0777 |
| D6         | 3.0777 |
| D7         | 3.1067 |
| D8         | 3.1654 |
| D9         | 3.2047 |
| D10        | 35.5803 |

Bold values indicate the minimum MSE.
(v) $X_4 =$ number of males per 1000 females
(vi) $X_5 =$ state population size in hundred thousands
(vii) $X_6 =$ unemployment rate of urban males per 1000 of age 35–39
(viii) $X_7 =$ median value of transferable goods and assets or family income in tens of dollars
(ix) $X_8 =$ the number of families per 1000 earnings 1/2 the median income

Table 13 gives the eigenvalues and condition number for these data, which shows that moderately high multicollinearity is present. The value of test statistic from Shapiro-Wilk normality test is $W = 0.99379$ with $p$-value = 0.9967, which shows that the response variable is normal at 5% level of significance. MSE results are tabulated in Table 14. From Table 14, we see that MSE of LEs outperforms OLS. Among all LEs, the proposed estimator D6 showed smallest MSE, which supports the simulation results given in Section 3.

Therefore, based on simulation results and illustrative examples, we recommend the use of LE with Liu parameter D6 and D7 to practitioners in presence of high or severe multicollinearity.

5. Concluding Remarks

In this paper, we introduced a new quantile based method to estimate the Liu parameter in order to minimize the variance and circumvent the problem of multicollinearity. Extensive Monte Carlo simulations were carried out to evaluate the performance of estimators with MSE and MAE criterions by varying the values of different factors such as sample sizes, number of explanatory variables, error variances, and multicollinearity. Results from the simulation study revealed that the LE performed generally better than the traditional OLS estimator.

The LE is a robust choice than the OLS when the problem of multicollinearity is present. Also, the proposed estimator D6 performs better than other considered estimators in many evaluated instances particularly when the problem of multicollinearity is extremely high. Estimator D5 is the closest competitor to D6. Furthermore, the benefits of the new estimators are evidently confirmed in the two empirical applications. Based on the simulation results along with the real applications, we conclude that the Liu method with proposed estimator D6 is the best choice for practitioners to overcome the problem of multicollinearity.

Data Availability

Data used in this research is taken from Portland cement limited available online at: H. Woods, H. H. Steinour, and H. R. Starke, “Effect of composition of Portland cement on heat evolved during hardening,” Ind. Eng. Chem., vol. 24, no. 11, pp. 1207–1214, 1932. Research codes will be provided on personal request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] Liu, “A new class of biased estimate in linear regression,” Communications in Statistics Methods, vol. 22, no. 2, pp. 393–402, 1993.
[2] D. N. Gujarati, Basic Econometrics, Tata McGraw-Hill Education, New Delhi, India, 2009.
[3] A. E. Hoerl and R. W. Kennard, “Ridge regression: biased estimation for nonorthogonal problems,” Technometrics, vol. 12, no. 1, pp. 55–67, 1970.
[4] M. Qasim, M. Amin, and T. Omer, “Performance of some new liu parameters for the linear regression model,” Communications in Statistics—Theory and Methods, vol. 49, no. 17, pp. 4178–4196, 2020.
[5] G. Khalaf and G. Shukur, “Choosing ridge parameter for regression problems,” Communications in Statistics—Theory and Methods, vol. 34, no. 5, pp. 1177–1182, 2005.
[6] G. Shukur, K. Månnsson, and P. Sjolander, “Developing interaction shrinkage parameters for the liu estimator—with an application to the electricity retail market,” Computational Economics, vol. 46, no. 4, pp. 539–550, 2015.
[7] K. Månnsson, B. M. G. Kibria, and G. Shukur, “On liu estimators for the logit regression model,” Economic Modelling, vol. 29, no. 4, pp. 1483–1488, 2012.
[8] I. Perveen and M. Suhail, “Bootstrap liu estimators for poisson regression model,” Communication in Statistics Computation, pp. 1–11, 2021.
[9] B. M. G. Kibria, “Performance of some new ridge regression estimators,” Communications in Statistics—Simulation and Computation, vol. 32, no. 2, pp. 419–435, 2003.
[10] S. Ali, H. Khan, I. Shah, M. M. Butt, and M. Suhail, “A comparison of some new and old robust ridge regression estimators,” Communications in Statistics—Simulation and Computation, pp. 1–19, 2019.
[11] M. Suhail, S. Chand, and B. M. G. Kibria, “Quantile based estimation of biasing parameters in ridge regression model,” Communications in Statistics—Simulation and Computation, vol. 49, no. 10, pp. 2732–2744, 2019.
[12] G. C. McDonald and D. I. Galarneau, “A monte carlo evaluation of some ridge-type estimators,” Journal of the American Statistical Association, vol. 70, no. 350, pp. 407–416, 1975.
[13] M. Suhail and S. Chand, “Performance of some new ridge regression estimators,” in Proceedings of the 2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS), pp. 1–4, Karachi, Pakistan, December 2019.
[14] M. S. Ayanullah and M. Ilyas, “Modified method for choosing ridge parameter,” Journal of Statistics, vol. 24, pp. 20–34, 2017.
[15] S. Chatterjee and A. S. Hadi, Regression Analysis by Example, John Wiley & Sons, Hoboken, NJ, USA, 2015.
[16] H. Woods, H. H. Steinour, and H. R. Starke, “Effect of composition of Portland cement on heat evolved during hardening,” Industrial & Engineering Chemistry, vol. 24, no. 11, pp. 1207–1214, 1932.
[17] G. Trenkler and H. Toutenburg, “Mean squared error matrix comparisons between biased estimators—an overview of
recent results,” *Statistical Papers*, vol. 31, no. 1, pp. 165–179, 1990.

[18] N. R. Draper and H. Smith, *Applied Regression Analysis*, John Wiley & Sons, Hoboken, NJ, USA, 1998.

[19] A. S. Hadi and R. F. Ling, “Some cautionary notes on the use of principal components regression,” *The American Statistician*, vol. 52, no. 1, pp. 15–19, 1998.

[20] A. V. Dorugade, “Adjusted ridge estimator and comparison with kibria’s method in linear regression,” *Journal of the Association of Arab Universities for Basic and Applied Sciences*, vol. 21, no. 1, pp. 96–102, 2016.

[21] K. Ayinde, A. F. Lukman, S. O. Olarenwaju, and M. O. Attah, “Some new adjusted ridge estimators of linear regression model,” *International Journal of Civil Engineering Technology*, vol. 9, no. 11, pp. 2838–2852, 2018.

[22] A. F. Lukman, K. Ayinde, S. Binuomote, and O. A. Clement, “Modified ridge-type estimator to combat multicollinearity: application to chemical data,” *Journal of Chemometrics*, vol. 33, no. 5, pp. 1–12, 2019.

[23] B. M. G. Kibria and S. Banik, “Some ridge regression estimators and their performances,” *Journal of Modern Applied Statistical Methods*, vol. 15, no. 1, pp. 206–238, 2017.