Modified Liu estimators in the linear regression model: An application to Tobacco data

Iqra Babar\(^1\), Hamdi Ayed\(^2\), Sohail Chand\(^1\), Muhammad Suhail\(^3\), Yousaf Ali Khan\(^4\)*, Riadh Marzouki\(^5\)

\(^1\) College of Statistical and Actuarial Sciences, University of the Punjab, Lahore, Pakistan, \(^2\) Department of Civil Engineering, College of Engineering, King Khalid University, Abha, KSA, \(^3\) Department of Statistics, The University of Agriculture Peshawar, Amir Muhammad Khan Campus, Mardan, Pakistan, \(^4\) Department of Mathematics and Statistics, Hazara University Mansehra, Mansehra, Pakistan, \(^5\) Department of Chemistry, College of Science, King Khalid University, Abha, KSA

* yousaf_hu@yahoo.com

Abstract

Background

The problem of multicollinearity in multiple linear regression models arises when the predictor variables are correlated among each other. The variance of the ordinary least squared estimator become unstable in such situation. In order to mitigate the problem of multicollinearity, Liu regression is widely used as a biased method of estimation with shrinkage parameter ‘\(d\)’. The optimal value of shrinkage parameter plays a vital role in bias-variance trade-off.

Limitation

Several estimators are available in literature for the estimation of shrinkage parameter. But the existing estimators do not perform well in terms of smaller mean squared error when the problem of multicollinearity is high or severe.

Methodology

In this paper, some new estimators for the shrinkage parameter are proposed. The proposed estimators are 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. Mean squared error and mean absolute error is considered as evaluation criteria of the estimators. Tobacco dataset is used as an application to illustrate the benefits of the new estimators and support the simulation results.

Findings

The new estimators outperform the existing estimators in most of the considered scenarios including high and severe cases of multicollinearity. 95% mean prediction interval of all the estimators is also computed for the Tobacco data. The new estimators give the best mean prediction interval among all other estimators.
The implications of the findings
We recommend the use of new estimators to practitioners when the problem of high to severe multicollinearity exists among the predictor variables.

1. Introduction
Ordinary least squared (OLS) method of estimation is commonly used in linear regression models. When the problem of multicollinearity exists among the predictor variables then the results obtained by the method of OLS can be misleading [1]. Ridge regression (RR) and Liu regression (LR) suggested by [2, 3] respectively are the two commonly used methods in order to mitigate this problem. LR is usually preferred over RR, because it is the linear function of its shrinkage parameter $d$ [4].

The optimal value of shrinkage parameter $d$ in LR plays an important role in minimizing the variance. Many researchers have suggested several LR estimators for estimating $d$. Few of them are [4–6] and very recently [7–9]. The existing estimators perform better only when the problem of multicollinearity is not very high. In case of very high or severe multicollinearity, the existing estimators do not perform well in terms of smaller mean squared error (MSE) and mean absolute error (MAE) respectively. To overcome this problem, it was necessary to develop some new estimators.

Therefore, the objective of this paper is to propose some new estimators that are robust to the presence of very high to severe level of multicollinearity. In this paper, the performance of some existing LR estimators is investigated and some new estimators for shrinkage parameter $d$ are proposed. The new proposed estimators give the optimal choice of shrinkage parameter and are robust to the presence of very high and severe multicollinearity. Also, the new estimators are compared with the existing ones through a Monte Carlo simulation based on MSE and MAE performance criterions. The MSE and MAE of the new estimators is smaller than OLS and other existing LR estimators and outperform in most of the considered scenarios.

Rest of the article is organized distributed as follows. The statistical methodology that includes the model estimation, new proposed and existing LR estimators are discussed in Section 2. The simulation design and results are discussed in Section 3. Section 4 includes the empirical application to demonstrate the benefits of the new estimators. The conclusion of the paper is given in Section 5.

2. Statistical methodology
Consider the following multiple linear regression model in matrix form as:

$$ y = X\beta + \varepsilon, $$

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

$$ \hat{\beta}_{OLS} = (X'X)^{-1}X'Y. $$

The OLS estimator gives the unbiased and efficient results provided the basic assumptions of
the classical linear regression models are satisfied [1]. However, in the presence of multicollinearity, OLS estimator become inefficient and provide large variance [4]. To circumvent such situation, numerous biased estimation methods are available that provide smaller MSE than OLS and LR is one of them. The LR estimator defined in [2] is given below:

$$\hat{\beta}_{Liu} = (X'X + I)^{-1}(X'X + dI)\hat{\beta}_{OLS}, \quad 0 \leq d \leq 1$$

In the presence of multicollinearity, \(\hat{\beta}_{Liu}\) provide the smaller MSE than OLS [10]. The optimal choice of shrinkage parameter \(d\) plays a vital role in minimizing the MSE of \(\hat{\beta}_{Liu}\). Some existing LR estimators for the shrinkage parameter \(d\) are given in the following sub-section.

2.1 Some existing LR estimators

Consider the canonical form of model (1) as:

$$y = Z\alpha + \epsilon, \quad (4)$$

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

$$\hat{\alpha}_{OLS} = \Lambda^{-1}Z'y, \quad (5)$$

The LR estimator is defined as:

$$\hat{\alpha}_{Liu} = (\Lambda + I)^{-1}(\Lambda + dI)\hat{\alpha}_{OLS}, \quad (6)$$

The first estimator for \(d\) was suggested by [2] and is given below:

$$\hat{d}_j = \frac{\hat{\alpha}_j^2 - \hat{\sigma}^2}{\lambda_j + \hat{\alpha}_j^2} \quad (7)$$

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

$$D1 = \max(0, \frac{\hat{\alpha}_{\text{max}}^2 - \hat{\sigma}^2}{\hat{\sigma}_{\text{max}}^2 + \hat{\alpha}_{\text{max}}^2}) \quad (8)$$

Shukur et al., [6] considered the idea of [5, 11] and suggested the following three estimators:

$$D2 = \max(0, \text{median}(\frac{\hat{\alpha}_j^2 - \hat{\sigma}^2}{\lambda_j + \hat{\alpha}_j^2})) \quad (9)$$

$$D3 = \max(0, \frac{1}{p} \sum_{j=1}^{p} (\frac{\hat{\alpha}_j^2 - \hat{\sigma}^2}{\lambda_j + \hat{\alpha}_j^2})) \quad (10)$$

$$D4 = \max(0, \max(\frac{\hat{\alpha}_j^2 - \hat{\sigma}^2}{\lambda_j + \hat{\alpha}_j^2})) \quad (11)$$
Shukur et al. [6] also suggested the following estimators:

\[ \hat{q}_j = \frac{\hat{\sigma}_j^2 - 1}{\max(\frac{1}{L_j}) + \hat{\sigma}_j^2} \]  

\[ D_5 = \max(0, \text{median}(\hat{q}_j)) \]  

\[ D_6 = \max(0, \frac{1}{p} \sum_{j=1}^{p} \hat{q}_j) \]  

\[ D_7 = \max(0, \max(\hat{q}_j)) \]

Based on the work of [4, 7], we proposed three new LR estimators in the section to follow.

2.2 Proposed method

Following the idea of [4, 7], we propose the following new estimator:

\[ \hat{d}_j = \frac{(\hat{\sigma}_j)_\gamma - \hat{\sigma}^2}{\max(\frac{1}{L_j}) + \max(\hat{\sigma}^2)} \]  

where ‘\( \gamma \)’ is the quantile probability. In order to obtain the minimum MSE and MAE, the new estimator \( \hat{d}_j \) depends on the quantile probability whose value is selected according to the level of multicollinearity [7]. Since the range of shrinkage parameter must be between zero and one, therefore we rewrite the proposed estimator as:

\[ D_\gamma = \max(0, \hat{d}_j) = \max(0, \frac{(\hat{\sigma}_j)_\gamma - \hat{\sigma}^2}{\max(\frac{1}{L_j}) + \max(\hat{\sigma}^2)}) \]  

Eq (17) satisfies the interval condition for shrinkage parameter \( d \) suggested by [2]. In order to present the role of quantile probability, we choose some specific values for ‘\( \gamma \)’ as: 0 (minimum), 0.25 (first quartile) and 0.50 (median). The mathematical form of three new LR estimators obtained is given below:

\[ D_8 = \max(0, \hat{d}_{0.0}) = \max(0, \frac{(\hat{\sigma}_j)_{0.0} - \hat{\sigma}^2}{\max(\frac{1}{L_j}) + \max(\hat{\sigma}^2)}) \]  

\[ D_9 = \max(0, \hat{d}_{0.25}) = \max(0, \frac{(\hat{\sigma}_j)_{0.25} - \hat{\sigma}^2}{\max(\frac{1}{L_j}) + \max(\hat{\sigma}^2)}) \]  

\[ D_{10} = \max(0, \hat{d}_{0.50}) = \max(0, \frac{(\hat{\sigma}_j)_{0.50} - \hat{\sigma}^2}{\max(\frac{1}{L_j}) + \max(\hat{\sigma}^2)}) \]

The procedure for generating and analyzing data is given in the next section.
3. The design of an experiment

In this section Monte Carlo simulation experiment, a commonly used procedure in literature for the data generation and analysis. Performance evaluation criterion and results are also discussed in this section.

3.1 The Monte Carlo simulation

In this section, the performance of LR estimators is compared through extensive simulations. Following [12], the predictor variables are generated as:

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

where \( \rho \) is the degree or level of multicollinearity between the predictor variables and are given as 0.90, 0.99, 0.999 and 0.9999. \( z_{ij} \) are the random numbers obtained from the standard normal distribution. The \( n \) observations on the response variable are computed as:

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

where \( \epsilon_i \sim N(0, \sigma^2) \), \( \sigma^2 \) is the error variance. \( \beta_0 \) is considered to be identically zero. Following [11], the eigen vector corresponding to maximum eigen value 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: \( \sigma^2 = 0.5, 1, 2 \)
- Sample size: \( n = 25, 50, 100, 200 \)
- Predictor variables: \( p = 4, 8, 16, 32 \)
- Multicollinearity: \( \rho = 0.90, 0.99, 0.999, 0.9999 \)

3.2 Performance evaluation criteria

Following [4], MSE and MAE criterions are used to judge the performance of the different LR estimators. Estimated MSE (EMSE) and MAE (EMAE) are defined as:

\[ \text{MSE} = \frac{\sum_{i=1}^{M} (\beta_1^i - \beta_1) (\bar{\beta}_1 - \beta_1)}{M}, \]

\[ \text{MAE} = \frac{\sum_{i=1}^{M} |\beta_1^i - \beta_1|}{M}, \]

where \( \hat{\beta}_1^i \) is the estimated value of \( \beta_1 \). \( M \) shows the simulation runs. In this study we choose \( M = 5000 \). The EMSE simulation results are presented in Tables 1–3 and Fig 1 and EMAE in Tables 4–6. The results are discussed in the section to follow.

3.3 Results and discussion

The EMSE and EMAE values of the new and existing LR estimators are presented in Tables 1–6 and Fig 1 respectively. The performance of the LR estimators is evaluated with respect to different factors such as multicollinearity, error variance, sample size and the predictor variables. These factors affect the simulation design [8]. The effect of each factor on EMSE and EMAE of estimators are discussed below:
Multicollinearity: Increase in the level of multicollinearity increases the EMSE and EMAE of all the estimators. The performance of OLS estimator deteriorates when the multicollinearity becomes very high. LR estimators outperform OLS for all the levels of multicollinearity. However, among all LR estimators, the EMSE and EMAE of the proposed estimators D8-D10 is generally smaller than existing estimators. While the estimator D5 remain close competitor to the proposed estimators only in the case mild to high multicollinearity. But in case of high to severe multicollinearity only the proposed estimators outperform. Fig 1 also support the proposed estimators.

Sample size: Increase in the sample size generally decreases the EMSE and EMAE of all the estimators. But the variation in the sample size does not alter the best performance of proposed estimators as observed in the case of multicollinearity.

Predictors: When the number of predictors increases the EMSE and EMAE of all the estimator’s increases. But the performance pattern of the estimators remains same as in the case of multicollinearity and sample size. It is also seen from the tables that increase in the EMSE of OLS estimator is relatively higher than all LR estimators. LR with Liu parameter D8 exhibits the lowest EMSE and EMAE.

Error variance or Standard deviation: The EMSE and EMAE of the estimators increases with the increase in the value of error variance. However, the performance of proposed estimators is better than other existing estimators.

The concluded remarks from Tables 1–6 are that the new LR estimators D8-D10 perform efficiently than the other existing LR estimators particularly in the case high to severe

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

| $\rho$ | 0.90 | 0.99 | 0.999 | 0.9999 | 0.90 | 0.99 | 0.999 | 0.9999 |
|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| $n$   |       |       |       |       |       |       |       |       |
| OLS   | 1.1596 | 11.8198 | 119.5606 | 1185.9839 | 0.8299 | 8.8919 | 87.0138 | 860.7005 |
| D1    | 0.7348  | 9.4928  | 116.0962 | 1182.3169 | 0.5410 | 6.9544 | 83.7579 | 857.1885 |
| D2    | 0.5855  | 0.7127  | 0.8423  | 5.3231  | 0.4518 | 0.6859 | 0.4039 | 1.2142  |
| D3    | 0.5895  | 0.7627  | 1.0614  | 7.7797  | 0.4531 | 0.7210 | 0.5467 | 2.2779  |
| D4    | 0.6466  | 1.4899  | 5.9986  | 55.1485 | 0.4803 | 1.0350 | 2.4882 | 18.0555 |
| D5    | 0.5816  | 0.5747  | 0.2235  | 0.5026  | 0.4515 | 0.6364 | 0.1841 | 0.0782  |
| D6    | 0.5825  | 0.6304  | 0.3973  | 1.5679  | 0.4518 | 0.6599 | 0.2436 | 0.3090  |
| D7    | 0.6030  | 0.9241  | 2.0217  | 14.7750 | 0.4592 | 0.7610 | 0.6822 | 3.0010  |
| D8    | 0.5812  | 0.5340  | 0.1000  | 0.0200  | 0.4514 | 0.6289 | 0.1564 | 0.0222  |
| D9    | 0.5814  | 0.5434  | 0.1310  | 0.1171  | 0.4514 | 0.6302 | 0.1637 | 0.0351  |
| D10   | 0.5827  | 0.5783  | 0.2271  | 0.5706  | 0.4515 | 0.6368 | 0.1852 | 0.0800  |
| $n$   | 100    |       |       |       |       |       |       |       |
| OLS   | 0.3009  | 3.1852  | 32.7389 | 316.5423 | 0.1701 | 1.7969 | 17.8274 | 178.7094 |
| D1    | 0.2380  | 2.0586  | 29.8277 | 313.0200 | 0.1454 | 1.0979 | 15.3159 | 175.2989 |
| D2    | 0.2332  | 0.6859  | 0.3835  | 0.2193  | 0.1443 | 0.5792 | 0.4920 | 0.1406  |
| D3    | 0.2332  | 0.6988  | 0.4384  | 0.4404  | 0.1443 | 0.5849 | 0.5199 | 0.2306  |
| D4    | 0.2367  | 0.7789  | 0.9362  | 3.2281  | 0.1453 | 0.6167 | 0.7018 | 1.1503  |
| D5    | 0.2332  | 0.6789  | 0.3261  | 0.0616  | 0.1443 | 0.5781 | 0.4719 | 0.0886  |
| D6    | 0.2332  | 0.6837  | 0.3453  | 0.0989  | 0.1443 | 0.5792 | 0.4795 | 0.1022  |
| D7    | 0.2335  | 0.7062  | 0.4353  | 0.4108  | 0.1443 | 0.5855 | 0.5092 | 0.1854  |
| D8    | 0.2332  | 0.6779  | 0.3145  | 0.0425  | 0.1443 | 0.5780 | 0.4685 | 0.0816  |
| D9    | 0.2332  | 0.6780  | 0.3176  | 0.0474  | 0.1443 | 0.5780 | 0.4693 | 0.0835  |
| D10   | 0.2332  | 0.6790  | 0.3262  | 0.0616  | 0.1443 | 0.5781 | 0.4720 | 0.0886  

https://doi.org/10.1371/journal.pone.0259991.t001
multicollinearity. The new estimators also outperform the OLS estimator substantially. Therefore, it is concluded that the new estimators D8-D10 outperform in terms of smaller EMSE and EMAE. Also, among new estimators, the new estimator D8 is more efficient and is the best choice for the practitioners in the presence of high and severe multicollinearity.

4. Applications

In the previous section, the performance of estimators is evaluated through Monte Carlo simulation experiment where some ideal conditions are assumed. Contrary to the simulation study, in this section, a numerical example of Tobacco dataset taken from [13] is considered to evaluate the estimators in real world problems.

4.1 Tobacco data

The first numerical example used in this study is the Tobacco dataset taken from [13] to compare the performance of new estimators in applied scenario. This data has already been used in literature, see e.g., [14]. The dataset consists of 30 observations of tobacco blends. The percentage concentrations of four important components are considered as predictor variables and the amount of heat given off by the tobacco during the smoking process as a response variable. The model for this dataset is defined as:

$$y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \epsilon$$  (25)

Condition number (CN) is used to measure the severity of multicollinearity among predictor variables.
Table 3. EMSE with \( n = 100 \) and \( \epsilon \sim N(0, \sigma) \).

| \( \sigma \) | \( 0.5 \) | \( 0.99 \) | \( 0.999 \) | \( 0.9999 \) | \( 0.5 \) | \( 0.99 \) | \( 0.999 \) | \( 0.9999 \) |
|---|---|---|---|---|---|---|---|---|
| \( \rho \) | 0.90 | 0.99 | 0.999 | 0.9999 | 0.90 | 0.99 | 0.999 | 0.9999 |
| \( \rho \) | 0.3763 | 4.0681 | 45.2599 | 460.6382 | 5.0758 | 64.6146 | 723.5484 | 7352.3531 |
| \( \rho \) | 0.2552 | 0.8122 | 0.4619 | 0.0769 | 4.0668 | 13.0077 | 8.4506 | 6.6499 |
| \( \rho \) | 0.2562 | 0.8783 | 0.7998 | 1.3977 | 4.0668 | 13.0061 | 7.1471 | 1.0051 |
| \( \rho \) | 0.3736 | 3.0591 | 24.6822 | 232.2378 | 4.1879 | 16.5939 | 39.7142 | 241.8414 |
| \( \rho \) | 0.2552 | 0.8122 | 0.4548 | 0.0637 | 4.0668 | 13.0163 | 7.2507 | 1.1671 |
| \( \rho \) | 0.2552 | 0.8122 | 0.4561 | 0.0657 | 4.0687 | 13.1268 | 7.6249 | 2.0437 |
| \( \rho \) | 0.2553 | 0.8173 | 0.4743 | 0.1162 | 4.2435 | 14.9294 | 77.2964 | 7.1471 |
| \( \rho \) | 0.2552 | 0.8122 | 0.4547 | 0.0632 | 4.0687 | 13.0061 | 7.1487 | 1.0129 |
| \( \rho \) | 0.2552 | 0.8122 | 0.4563 | 0.0656 | 4.0668 | 13.0062 | 7.1666 | 1.0424 |

Fig 1. EMSE values for \( n = 25, p = 4 \) and \( \sim N(0,1) \).

https://doi.org/10.1371/journal.pone.0259991.g001

variables [15] given as:

\[
CN = \sqrt{\frac{\hat{\xi}_{\text{max}}}{\hat{\xi}_{\text{min}}}},
\]

https://doi.org/10.1371/journal.pone.0259991.t003
where $\lambda_{\text{max}}$ and $\lambda_{\text{min}}$ are the maximum and minimum eigen values of the matrix $XX$ respectively. Following [1], 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 CN for this dataset is 43.50096, which shows that high multicollinearity exists among the predictor variables. The Shapiro-Wilk ($W$) normality test is used to test the normality of response variable. We obtain the value test statistic $W = 0.91248$ and $P$-value = 0.06719 which shows that the response variable is normal at 5% level of significance. MSE of OLS and Liu estimators from [2] can be written as:

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

$$
\text{MSE}(\hat{a}(d)) = \sigma^2 \sum_{j=1}^{p} \frac{(\lambda_j + d)^2}{\lambda_j(\lambda_j + 1)} + (d - 1)^3 \sum_{j=1}^{p} \frac{\hat{a}_j^2}{\lambda_j(\lambda_j + 1)}.
$$

The estimated values for $d$, regression coefficients and MSE of estimators are presented in Table 7. This table shows that the LR estimators have smaller MSE than OLS. However, among the LR estimators, new estimator D8 outperform and therefore highly efficient among others.

### Table 4. EMAE with $\epsilon \sim N(0,1)$ and $p = 4.$

| $\rho$ | 0.90 | 0.99 | 0.999 | 0.9999 | 0.90 | 0.99 | 0.999 | 0.9999 |
|--------|------|------|-------|--------|------|------|-------|--------|
| $n$    | 25   | 50   |       |        | 25   | 50   |       |        |
| OLS    | 1.7028 | 5.3848 | 17.1241 | 53.8792 | 1.4083 | 4.5591 | 14.1983 | 44.9883 |
| D1     | 1.2204 | 3.9504 | 14.3493 | 45.9810 | 1.0159 | 3.1746 | 11.5845 | 37.6036 |
| D2     | 1.1298 | 1.1950 | 1.0779 | 2.1735 | 0.9656 | 1.1714 | 0.8180 | 1.0717 |
| D3     | 1.1335 | 1.2403 | 1.1869 | 2.5672 | 0.9667 | 1.2038 | 0.9260 | 1.4128 |
| D4     | 1.1783 | 1.6452 | 2.7749 | 7.7409 | 0.9889 | 1.4055 | 1.7821 | 4.2252 |
| D5     | 1.1277 | 1.1049 | 0.6551 | 0.7595 | 0.9655 | 1.1371 | 0.6125 | 0.3721 |
| D6     | 1.1281 | 1.1428 | 0.8046 | 1.2361 | 0.9656 | 1.1536 | 0.6809 | 0.5866 |
| D7     | 1.1396 | 1.3017 | 1.4257 | 3.2133 | 0.9694 | 1.2159 | 0.9296 | 1.3739 |
| D8     | 1.1275 | 1.0744 | 0.4854 | 0.2126 | 0.9654 | 1.1315 | 0.5704 | 0.2287 |
| D9     | 1.1276 | 1.0817 | 0.5370 | 0.3910 | 0.9654 | 1.1325 | 0.5823 | 0.2736 |
| D10    | 1.1284 | 1.1081 | 0.6587 | 0.7841 | 0.9655 | 1.1375 | 0.6137 | 0.3748 |

https://doi.org/10.1371/journal.pone.0259991.t004

where $\lambda_{\text{max}}$ and $\lambda_{\text{min}}$ are the maximum and minimum eigen values of the matrix $XX$ respectively. Following [1], 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 CN for this dataset is 43.50096, which shows that high multicollinearity exists among the predictor variables. The Shapiro-Wilk ($W$) normality test is used to test the normality of response variable. We obtain the value test statistic $W = 0.91248$ and $P$-value = 0.06719 which shows that the response variable is normal at 5% level of significance. MSE of OLS and Liu estimators from [2] can be written as:
4.2 Prediction interval

In this section, 95% mean prediction interval of all the estimators is computed from the Tobacco dataset. We consider the following values of predictor variables: \( X_0' = (X_{10}, X_{20}, X_{30}, X_{40}) = (20.6, 10.9, 33.62, 39.76) \). 100(1-\( \alpha \))% mean prediction interval for the response variable is given as:

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

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

where \( \hat{y}_0 = \hat{\beta}_1 X_{10} + \hat{\beta}_2 X_{20} + \hat{\beta}_3 X_{30} + \hat{\beta}_4 X_{40} \)
\( \tilde{y}_0 = \tilde{\beta}_1 X_{10} + \tilde{\beta}_2 X_{20} + \tilde{\beta}_3 X_{30} + \tilde{\beta}_4 X_{40} \) \( \hat{\beta} \) and \( \tilde{\beta} \) are the OLS and Liu estimators respectively. \( t_{1-\alpha/2, v} \) is the \( (1 - \frac{\alpha}{2}) \) quantile from the Student’s t-distribution with \( v = n - p \) degrees of freedom. \( \text{var}(\hat{y}_0) = \hat{\sigma}^2 X_0'(X'X)^{-1}X_0 \) and \( \text{var}(\tilde{y}_0) = \tilde{\sigma}^2 X_0'Q_d(X'X)^{-1}Q_d'X_0, Q_d = (X'X + I)^{-1}(X'X + dI) \). For detail see [1, 4]. The results for the 95% mean prediction interval is given in the Table 8. From this table, we see that the new estimator D8 gives the best mean prediction interval among all other estimators.
5. Concluding remarks

In this paper, some new quantile based LR estimators for the shrinkage parameter \( \gamma \) are proposed in order to minimize the variance and mitigate the problem of multicollinearity. Monte Carlo simulation experiment was performed to compare the performance of estimators. MSE and MAE performance measures were used. Multicollinearity, Sample size, predictor variables

Table 6. EMAE with \( n = 100 \) and \( \epsilon \sim N(0,\sigma) \).

| \( \sigma \) | 0.5 | 0.99 | 0.999 | 0.9999 | 0.9 | 0.99 | 0.999 | 0.9999 |
|---|---|---|---|---|---|---|---|---|
| \( \rho \) | 0.9 | 0.99 | 0.999 | 0.9999 | 0.9 | 0.99 | 0.999 | 0.9999 |
| OLS | 1.8681 | 5.9671 | 18.9033 | 59.2703 | 7.4724 | 23.8839 | 75.1942 | 236.5916 |
| D1 | 1.6846 | 5.1351 | 17.3340 | 55.4982 | 6.2488 | 20.2901 | 69.1871 | 221.6485 |
| D2 | 1.5014 | 2.7710 | 2.0561 | 0.8446 | 6.0071 | 11.1240 | 8.8420 | 7.0330 |
| D3 | 1.5032 | 2.8909 | 2.7621 | 3.3132 | 6.0071 | 11.1235 | 8.0795 | 3.0193 |
| D4 | 1.6832 | 4.7629 | 13.1063 | 39.4865 | 6.0381 | 12.3402 | 17.3253 | 36.6493 |
| D5 | 1.5014 | 2.7710 | 2.0374 | 0.7611 | 6.0071 | 11.1281 | 8.1490 | 3.2867 |
| D6 | 1.5014 | 2.7710 | 2.0411 | 0.7754 | 6.0075 | 11.1754 | 8.3905 | 4.1168 |
| D7 | 1.5015 | 2.7792 | 2.0875 | 0.9369 | 6.0590 | 11.7235 | 11.0411 | 13.2700 |
| D8 | 1.5014 | 2.7710 | 2.0372 | 0.7575 | 6.0071 | 11.1235 | 8.0795 | 3.0190 |
| D9 | 1.5014 | 2.7710 | 2.0379 | 0.7613 | 6.0071 | 11.1235 | 8.0806 | 3.0336 |
| D10 | 1.5014 | 2.7710 | 2.0415 | 0.7741 | 6.0071 | 11.1235 | 8.0925 | 3.0853 |

Table 7. Estimated values for \( \gamma \), MSE and regression coefficients of Tobacco data.

| Estimators | \( \gamma \) | MSE | \( \hat{\beta}_1 \) | \( \hat{\beta}_2 \) | \( \hat{\beta}_3 \) | \( \hat{\beta}_4 \) |
|---|---|---|---|---|---|---|
| OLS | - | 32.4972 | 0.4857 | -0.6728 | 1.0744 | 1.4438 |
| D1 | 0.8719 | 24.7963 | 0.4852 | -0.6157 | 0.9582 | 1.2696 |
| D2 | 0.0070 | 3.6233 | 0.4815 | -0.2303 | 0.1742 | 0.0940 |
| D3 | 0.1011 | 3.2659 | 0.4819 | -0.2722 | 0.2594 | 0.2219 |
| D4 | 0.3859 | 6.2963 | 0.4831 | -0.3992 | 0.5177 | 0.6091 |
| D5 | 0.0000 | 3.6710 | 0.4815 | -0.2272 | 0.1678 | 0.0844 |
| D6 | 0.0144 | 3.5772 | 0.4816 | -0.2336 | 0.1808 | 0.1040 |
| D7 | 0.0077 | 3.3188 | 0.4815 | -0.2306 | 0.1748 | 0.0949 |
| D8 | 0.0310 | 3.4872 | 0.4816 | -0.2410 | 0.1959 | 0.1266 |
| D9 | 0.0310 | 3.4872 | 0.4816 | -0.2410 | 0.1959 | 0.1266 |
| D10 | 0.0310 | 3.4872 | 0.4816 | -0.2410 | 0.1959 | 0.1266 |
and error variance were the different factors we choose to vary in our study. It is concluded that all the LR estimators generally perform better than OLS estimator. Furthermore, among the LR estimators, the new estimators have shown best performance in the simulation and application. The LR is a robust choice than the OLS when the problem of multicollinearity is present. Moreover, among the new estimators, D8 performs better than other considered estimators in many evaluated instances particularly when the problem of multicollinearity is very high and severe. Estimators D5 and D9 were the close competitors to D8. Therefore, we recommend the use of LR method with shrinkage new estimator D8 over OLS when the problem of multicollinearity is present in the data.

5.1 Future research
Future research directions: In this research only the problem of multicollinearity is considered. When the outliers are also present in the data then the performance of the estimators will change. In future, we can develop some new robust LR estimators to overcome the joint problem of multicollinearity and outliers.

Supporting information
S1 File.

Author Contributions
Conceptualization: Iqra Babar, Sohail Chand, Yousaf Ali Khan.
Formal analysis: Iqra Babar, Sohail Chand, Yousaf Ali Khan.
Funding acquisition: Hamdi Ayed, Riadh Marzouki.
Investigation: Iqra Babar, Hamdi Ayed, Sohail Chand, Riadh Marzouki.
Methodology: Iqra Babar, Muhammad Suhail, Yousaf Ali Khan.
Resources: Hamdi Ayed, Yousaf Ali Khan, Riadh Marzouki.
Software: Iqra Babar, Yousaf Ali Khan.
Supervision: Sohail Chand, Muhammad Suhail, Yousaf Ali Khan.
Validation: Iqra Babar, Riadh Marzouki.

Visualization: Hamdi Ayed, Sohail Chand, Muhammad Suhail, Yousaf Ali Khan.

Writing – original draft: Sohail Chand, Muhammad Suhail.

Writing – review & editing: Yousaf Ali Khan.

References

1. Gujarati D. N., Basic econometrics. Tata McGraw-Hill Education, 2009.
2. Liu, “A new class of biased estimate in linear regression,” Commun. Stat. Methods, vol. 22, no. 2, pp. 393–402, 1993.
3. Hoerl A. E. and Kennard R. W., “Ridge regression: Biased estimation for nonorthogonal problems,” Technometrics, vol. 12, no. 1, pp. 55–67, 1970.
4. Qasim M., Amin M., and Omer T., “Performance of some new Liu parameters for the linear regression model,” Commun. Stat. Methods, vol. 49, no. 17, pp. 4178–4196, 2020.
5. Khalaf G. and Shukur G., “Choosing ridge parameter for regression problems,” Commun. Stat. Methods, vol. 34, no. 5, pp. 1177–1182, 2005, https://doi.org/10.1081/STA-200056836
6. Shukur G., Månsson K., and Sjölander P., “Developing Interaction Shrinkage Parameters for the Liu Estimator—with an Application to the Electricity Retail Market,” Comput. Econ., vol. 46, no. 4, pp. 539–550, 2015.
7. Suhail M., Babar I., Khan Y. A., Imran M., and Nawaz Z., “Quantile-Based Estimation of Liu Parameter in the Linear Regression Model: Applications to Portland Cement and US Crime Data,” Math. Probl. Eng., vol. 2021, 2021.
8. Perveen I. and Suhail M., “Bootstrap Liu estimators for Poisson regression model,” Commun. Stat. Comput., vol. 0, no. 0, pp. 1–11, 2021.
9. Rajput D. S. et al., “Providing diagnosis on diabetes using cloud computing environment to the people living in rural areas of India,” J. Ambient Intell. Humaniz. Comput., pp. 1–12, 2021.
10. Månsson K., Kibria B. M. G., and Shukur G., “On Liu estimators for the logit regression model,” Econ. Model., vol. 29, no. 4, pp. 1483–1488, 2012.
11. Kibria, “Performance of some New Ridge regression estimators,” Commun. Stat. Part B Simul. Comput., vol. 32, no. 2, pp. 419–435, 2003, https://doi.org/10.1081/SAC-120017499
12. McDonald G. C. and Galameau D. I., “A Monte Carlo evaluation of some ridge-type estimators,” J. Am. Stat. Assoc., vol. 70, no. 350, pp. 407–416, 1975.
13. Myers R. H., Classical and modern regression with applications, vol. 2. Duxbury press Belmont, CA, 1990.
14. Suhail M., Chand S., and Kibria B. M. G., “Quantile-based robust ridge m-estimator for linear regression model in presence of multicollinearity and outliers,” Commun. Stat. Simul. Comput., 2019, https://doi.org/10.1080/03610918.2019.1621339
15. Kibria B. M. G. and Banik S., “Some Ridge Regression Estimators and Their Performances,” J. Mod. Appl. Stat. Methods, vol. 15, no. 1, pp. 206–238, 2017, https://doi.org/10.22237/jmasm/1482075860