RESEARCH ARTICLE

Unbiased K-L estimator for the linear regression model

[version 1; peer review: 2 approved, 1 approved with reservations]

Benedicta Aladeitan ¹, Adewale F Lukman ², Esther Davids¹, Ebele H Oranye³, Golam B M Kibria⁴

¹Physical Sciences, Landmark University, Omu-Aran., Kwara State, +234, Nigeria

²Biostatistics and Epidemiology, University of Medical Sciences., Ondo, Ondo State, +234, Nigeria

³Statistics, University of Nigeria, Nsukka, Enugu State, +234, Nigeria

⁴Mathematics and Statistics, Florida International University, Florida, USA, USA

V1 First published: 19 Aug 2021, 10:832
<https://doi.org/10.12688/f1000research.54990.1>
 Latest published: 19 Aug 2021, 10:832
<https://doi.org/10.12688/f1000research.54990.1>

Abstract

Background: In the linear regression model, the ordinary least square (OLS) estimator performance drops when multicollinearity is present. According to the Gauss-Markov theorem, the estimator remains unbiased when there is multicollinearity, but the variance of its regression estimates become inflated. Estimators such as the ridge regression estimator and the K-L estimators were adopted as substitutes to the OLS estimator to overcome the problem of multicollinearity in the linear regression model. However, the estimators are biased, though they possess a smaller mean squared error when compared to the OLS estimator.

Methods: In this study, we developed a new unbiased estimator using the K-L estimator and compared its performance with some existing estimators theoretically, simulation wise and by adopting real-life data.

Results: Theoretically, the estimator even though unbiased also possesses a minimum variance when compared with other estimators. Results from simulation and real-life study showed that the new estimator produced smaller mean square error (MSE) and had the smallest mean square prediction error (MSPE). This further strengthened the findings of the theoretical comparison using both the MSE and the MSPE as criterion.

Conclusions: By simulation and using a real-life application that focuses on modelling, the high heating values of proximate analysis was conducted to support the theoretical findings. This new method of estimation is recommended for parameter estimation with and without multicollinearity in a linear regression model.

Open Peer Review

Approval Status

	1	2	3
version 1			
19 Aug 2021	view	view	view

1. **Oluwayemisi Oyerinke Alaba** , University of Ibadan, Ibadan, Nigeria
2. **Muhammad Amin** , University of Sargodha, Sargodha, Pakistan
3. **Mohammad Arashi** , Ferdowsi University of Mashhad, Mashhad, Iran

Any reports and responses or comments on the article can be found at the end of the article.

Keywords

Linear regression model, Ordinary Least Square estimator, Ridge regression, K-L estimator, High Heating values, Proximate analysis.

Corresponding author: Benedicta Aladeitan (aladeitan.benedicta@lmu.edu.ng)

Author roles: **Aladeitan B:** Writing – Original Draft Preparation; **Lukman AF:** Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Davids E:** Writing – Review & Editing; **Oranye EH:** Writing – Review & Editing; **Kibria GBM:** Methodology, Resources, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2021 Aladeitan B *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Aladeitan B, Lukman AF, Davids E *et al.* **Unbiased K-L estimator for the linear regression model [version 1; peer review: 2 approved, 1 approved with reservations]** F1000Research 2021, 10:832 <https://doi.org/10.12688/f1000research.54990.1>

First published: 19 Aug 2021, 10:832 <https://doi.org/10.12688/f1000research.54990.1>

Introduction

Considering the general linear regression model

$$y = X\beta + \varepsilon_i \quad (1)$$

such that ε_i is normally distributed with mean 0 and variance $\sigma^2 I$ where I is the identity matrix. y is an $n \times 1$ vector of dependent variable, X is an $n \times p$ matrix of the independent variables, β is a $p \times 1$ vector of unknown regression parameters of interest. The method of ordinary least square (OLS) is well known and generally accepted for estimating the parameters (β s) in the linear regression model. The model is defined as:

$$\hat{\beta}_{OLS} = H^{-1} X' y \quad (2)$$

Where $H = X'X$ and $\hat{\beta}_{OLS}$ is normally distributed that is $\hat{\beta}_{OLS} \sim N(\beta, \sigma^2 H^{-1})$. However, when the OLS estimator is applied to a model where there is correlation between the independent variables, then the variance of the regression estimates becomes inflated^{1,2}. This relationship between the independent variables is referred to as multicollinearity^{3,4}.

In addressing the problem of multicollinearity, various biased estimators with mean square error smaller than the OLS have been developed by different authors²⁻¹⁵. The limitations of these estimators is that they are biased, however the unbiased versions of some of them have been developed. The advantage of these estimators is that they produced estimates that were similar to the OLS estimator with better mean squared error. Crouse *et al.*^{16,17} developed the unbiased ridge and the Liu estimators. Wu¹⁸ developed the unbiased version of the two-parameter estimator by Ozkale and Kaciranlar⁹. Lukman *et al.*¹⁹ developed the unbiased modified ridge-type estimator. Recently, the K-L estimator was proposed to circumvent the problem of multicollinearity in the linear regression model¹³. The K-L estimator is classified as a biased estimator with a single biasing parameter¹³.

In this study, a new unbiased technique is developed based on the K-L estimator and its properties are derived. We compared the unbiased K-L estimator with some existing techniques using the mean square error (MSE) criterion.

Methods

Unbiased K-L estimator with prior information

Hoerl and Kennard⁵ developed the ridge estimator to mitigate multicollinearity in the linear regression model. The ridge estimator of β with the biasing parameter k is:

$$\hat{\beta}_{RR}(k) = (H + kI)^{-1} X' y, \quad k > 0 \quad (3)$$

The modified ridge technique was proposed with the addition of the prior information⁶. This is expressed as follows:

$$\hat{\beta}_{MRE}(k, b) = (H + kI)^{-1} (X' y + kb) \quad (4)$$

According to 16, the unbiased ridge estimator with the introduction of prior information J is given as

$$\hat{\beta}_{UMRE} = (H + kI)^{-1} (X' y + kJ) \quad (5)$$

where J and $\hat{\beta}_{OLS}$ are uncorrelated and $J \sim N(\beta, D)$ such that $D = (\frac{\sigma^2}{k}) I_p$ and I_p is $p \times p$ identity matrix. J is estimated by $J = \frac{\sum_{i=1}^p \hat{\beta}_i}{p}$

Modified ridge-type method proposed by 3 is given as follows:

$$\hat{\beta}_{MRT}(k, d) = [H + k(1+d)]^{-1} S \hat{\beta}_{OLS} = A_{kd} \hat{\beta}_{OLS} \quad (6)$$

where $A_{kd} = [S + k(1+d)]^{-1} S$.

The unbiased modified ridge type estimator¹⁹ was developed and defined as follows:

$$\hat{\beta}_{UMRT}(A_{kd}, J) = A_{kd} \hat{\beta}_{OLS} + (I - A_{kd})J = \hat{\beta}_{MRT}(k, d) + (I - A_{kd})J \quad (7)$$

where $A_{kd} = [H + k(1+d)]^{-1} H$ such that $D = \frac{\sigma^2}{k(1+d)}$. Consequently, $J \sim N\left(\beta, \frac{\sigma^2}{k(1+d)}\right)$ for

$k > 0, 0 < d < 1$.

Recently¹³, proposed the K-L estimator and found that this estimator generally outperform the ridge regression estimator. The K-L estimator of β is defined as:

$$\hat{\beta}_{KL}(k) = (H + kI)^{-1} (H - kI) \hat{\beta}_{OLS} = A_k \hat{\beta}_{OLS} \quad k > 0 \quad (8)$$

where $A_k = (H + kI)^{-1} (H - kI)$

This research proposes an unbiased K-L estimator following the convex method. The convex method is defined as:

$$\hat{\beta}(G, J) = G \hat{\beta}_{OLS} + (I - G)J \quad (9)$$

where G is a $p \times p$ matrix and I is an identity matrix of $p \times p$ dimensions. Thus, the MSE of $\hat{\beta}(G, J)$ is

$$MSE(\hat{\beta}(G, J)) = \sigma^2 G H^{-1} G' + (I - G) D (I - G)' \quad (10)$$

Such that,

$$\frac{\partial_{MSE}(\hat{\beta}(G, J))}{\partial G} = 2G(\sigma^2 H^{-1} + D) - 2D = 0 \quad (11)$$

The value of G from (11) is $G = D(\sigma^2 H^{-1} + D)^{-1}$. Accordingly, $D = \sigma^2 (I - G)^{-1} G H^{-1}$. We observed that the convex estimator $\hat{\beta}(G, J)$ is an unbiased estimator of β and possesses minimum MSE for optimal value of G . Consequently, the new unbiased estimator is defined as

$$\hat{\beta}_{UKL}(k, J) = A_k \hat{\beta}_{OLS} + (I - A_k)J = \hat{\beta}_{KL}(k) + (I - A_k)J \quad (12)$$

where $A_k = (H + kI)^{-1} (H - kI)$ and the value of $V = \frac{\sigma^2 (S - kI) S^{-1}}{2k}$. Therefore, $J \sim N\left(\beta, \frac{\sigma^2 (S - kI) S^{-1}}{2k}\right)$ for $k > 0$.

It can be expressed conveniently that $\hat{\beta}_{UKL}(k, J)$ is unbiased for β . The new estimator has properties defined as follows:

$$\begin{aligned} E(\hat{\beta}_{UKL}(k, J)) &= E(\hat{\beta}_{KL}(k) + (I - A_k)J) \\ &= E(A_k \hat{\beta}_{OLS} + [I - A_k]J) \text{ where } E(\hat{\beta}_{OLS}) = E(J) = \beta \\ &= A_k \beta + (I - A_k) \beta \\ &= \beta \end{aligned} \quad (13)$$

It follows from Equation (13) that the proposed estimator is unbiased. This classified the new estimator into the same class with the OLS estimator. The biasedness of the estimator is also zero. This is proved as follows:

$$Bias(\hat{\beta}_{UKL}(k, J)) = E(\hat{\beta}_{UKL}(k, J)) - \beta = \beta - \beta = 0 \quad (14)$$

$$MSE(\hat{\beta}_{UKL}(k, J)) = D(\hat{\beta}_{KL}(k) + (I - A_k)J) = \frac{\sigma^2 (H - kI)}{H(H + kI)} \quad (15)$$

Given that there exists an orthogonal matrix Q , such that $Q'X'XQ = E = \text{diag}(e_1, e_2, \dots, e_p)$ where e_i is the i^{th} eigenvalue of $X'X$, E and Q are the matrices of eigenvalues and eigenvectors of $X'X$ respectively. Equation (1) can be expressed canonically as:

$$y = Z\alpha + \varepsilon \quad (16)$$

where $Z = XQ$, $\alpha = Q'\beta$ and $Z'Z = E$. For Equation (15), we get the following representations:

$$\hat{\alpha}_{OLS} = E^{-1}Z'y \quad (17)$$

$$\hat{\alpha}_{RR}(k) = (E + kI)^{-1}Z'y \quad (18)$$

$$\hat{\alpha}_{UMRE}(k) = [E + kI]^{-1}(Z'y + kJ) \quad (19)$$

$$\hat{\alpha}_{KL}(k) = (E + kI)^{-1}(E - kI)Z'y \quad (20)$$

$$\hat{\alpha}_{UKL}(k, J) = \hat{\alpha}_{KL}(k) + (I - A_k)J \quad (21)$$

Lemma 1.1 Let N be an $n \times n$ positive definite matrix and α be some vector, then $N - \alpha\alpha' \geq 0$ if and only if $\alpha'N^{-1}\alpha \leq 1$ ²⁰.

Lemma 1.2 Let $\hat{\alpha}_i = C_i y$ $i = 1, 2$ be two linear estimators of α . Suppose that $D = \text{Cov}(\hat{\alpha}_1) - \text{Cov}(\hat{\alpha}_2) > 0$, where $\text{Cov}(\hat{\alpha}_i)$, $i = 1, 2$ denotes the covariance matrix of $\hat{\alpha}_i$ and $\text{bias}(\hat{\alpha}_i) = b = (C_i X - I)\alpha$, $i = 1, 2$. Consequently,

$$(\hat{\alpha}_1 - \hat{\alpha}_2) = MSEM(\hat{\alpha}_1) - MSEM(\hat{\alpha}_2) = \sigma^2 D + b_1 b_1' - b_2 b_2' > 0 \quad (22)$$

if and only if $b_2'[\sigma^2 D + b_1 b_1']^{-1} b_2 < 1$ where $MSEM(\hat{\alpha}_i) = \text{Cov}(\hat{\alpha}_i) + b_i b_i'$ ²¹.

Theoretical comparison

$\hat{\alpha}_{OLS}$ and $\hat{\alpha}_{UKL}(k, J)$

Theorem 1.1. $\hat{\alpha}_{UKL}(k, J)$ is preferred to $\hat{\alpha}_{OLS}$ by using the matrix mean square error as criteria for $k > 0$.

Proof

Recall that,

$$MSEM(\hat{\alpha}_{OLS}) = \sigma^2 E^{-1} \quad (23)$$

$$MSEM(\hat{\alpha}_{UKL}(k, J)) = \sigma^2 (E + kI)^{-1} (E - kI) E^{-1} \quad (24)$$

The difference between (23) and (24) is as follows:

$$\begin{aligned} MSEM(\hat{\alpha}_{OLS}) - MSEM(\hat{\alpha}_{UKL}(k, J)) &= \sigma^2 E^{-1} - \sigma^2 (E + kI)^{-1} (E - kI) E^{-1} \\ &= \sigma^2 [E^{-1} - (E + kI)^{-1} (E - kI) E^{-1}] \\ &= \sigma^2 \text{diag} \left[\frac{1}{e_i} - \frac{(e_i - k)}{e_i(e_i + k)} \right] \end{aligned} \quad (25)$$

Simplifying (25) further, we observed that $E^{-1} - (E + kI)^{-1} (E - kI) E^{-1}$ will be positive definite since $2k > 0$ for $k > 0$.

$\hat{\alpha}_{RR}(k)$ and $\hat{\alpha}_{UKL}(k, J)$

Theorem 3.2. $\hat{\alpha}_{UKL}(k, J)$ is preferred to $\hat{\alpha}_{RR}(k)$ by using the matrix mean square error as criteria for $k > 0$.

Proof

$$MSEM(\hat{\alpha}_{RR}(k)) = \sigma^2 B_k E B_k' + k^2 B_k \alpha \alpha' B_k' \quad (26)$$

where $B_k = (E + kI)^{-1}$

The difference of Equation (24) and (26) is as follows:

$$\begin{aligned} MSEM(\hat{\alpha}_{RR}(k)) - MSEM(\hat{\alpha}_{UKL}(k, J)) &= \sigma^2 E(E+k)^{-2} - \sigma^2 (E+k)^{-1} (E-k) E^{-1} \\ &= \sigma^2 [E(E+k)^{-2} - (E+k)^{-1} (E-k) E^{-1}] \\ &= \sigma^2 \text{diag} \left[\frac{e_i}{(e_i+k)^2} - \frac{(e_i-k)}{e_i(e_i+k)} \right] \end{aligned} \quad (27)$$

Simplifying (27) further, we observed that $E(E+k)^{-2} - (E+k)^{-1} (E-k) E^{-1}$ will be positive definite since $k^2 > 0$.

$\hat{\alpha}_{URR}(k)$ and $\hat{\alpha}_{UKL}(k, J)$

Theorem 3.3. $\hat{\alpha}_{UKL}(k, J)$ is preferred to $\hat{\alpha}_{URR}(k)$ by using the matrix mean square error as criteria for $k > 0$.

Proof

$$MSEM(\hat{\alpha}_{URR}(k)) = \sigma^2 (E + kI)^{-1} \quad (28)$$

$$\begin{aligned} MSEM(\hat{\alpha}_{URR}(k)) - MSEM(\hat{\alpha}_{UKL}(k, J)) &= \sigma^2 (E+k)^{-1} - \sigma^2 (E+k)^{-1} (E-k) E^{-1} \\ &= \sigma^2 \text{diag} \left[\frac{1}{(e_i+k)} - \frac{(e_i-k)}{e_i(e_i+k)} \right] \end{aligned} \quad (29)$$

We observed that $\sigma^2 (E+k)^{-1} - \sigma^2 (E+k)^{-1} (E-k) E^{-1}$ will be positive definite since $k > 0$.

$\hat{\alpha}_{KL}(k)$ and $\hat{\alpha}_{UKL}(k, J)$

Theorem 3.4. $\hat{\alpha}_{UKL}(k, J)$ is preferred to $\hat{\alpha}_{KL}(k)$ by using the matrix mean square error as criteria for $k > 0$.

Proof

$$MSEM \hat{\alpha}_{KL}(k) = \sigma^2 E_k (E-k) \Lambda^{-1} E_k (E-k) + 4k^2 E_k \alpha \alpha' E_k \quad (30)$$

where $E_k = (E+k)^{-1}$. Consequently,

$$\begin{aligned} MSEM(\hat{\alpha}_{KL}(k)) - MSEM(\hat{\alpha}_{UKL}(k, J)) &= \sigma^2 (E-k)^2 E^{-1} (E+k)^{-2} - \sigma^2 (E-k) E^{-1} (E+k)^{-1} \\ &= \sigma^2 \text{diag} \left[\frac{(e_i-k)^2}{e_i(e_i+k)^2} - \frac{(e_i-k)}{e_i(e_i+k)} \right] \end{aligned} \quad (31)$$

We observed that $\sigma^2 (E-k)^2 E^{-1} (E+k)^{-2} - \sigma^2 (E-k) E^{-1} (E+k)^{-1}$ will be positive definite if $k > 2\Lambda$ for $k > 0$.

Selection of the biasing parameters

In this study, we adopt the following biasing parameter for the ridge and the unbiased ridge estimators:

$$\hat{k} = \frac{p\hat{\sigma}^2}{\sum_{i=1}^p \hat{\alpha}_{i,OLS}^2} \quad (32)$$

For the K-L estimator, we adopted:

$$\hat{k}_r = \min \left[\frac{\hat{\sigma}^2}{2\hat{\alpha}_{i,OLS}^2 + \frac{\hat{\sigma}^2}{e_i}} \right] \quad (33)$$

For the proposed estimator, the following biasing parameters were examined:

$$\text{UKL-1; } \hat{k}_r = \min \left[\frac{\hat{\sigma}^2}{2\hat{\alpha}_{i,OLS}^2 + \frac{\hat{\sigma}^2}{e_i}} \right] \quad (34)$$

$$\text{UKL-2}; \hat{k}_1 = \max \left[\frac{1}{\hat{\sigma}_{i,OLS}^2} \right] \quad (35)$$

$$\text{UKL-3}; \hat{k}_2 = \sqrt{\max \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_{i,OLS}^2} \right)} \quad (36)$$

$$\text{UKL-4}; k = \frac{p\sigma^2}{\sum_{i=1}^p \hat{\sigma}_{i,OLS}^2} \quad (37)$$

Results

R Studio was used for both the simulation and real-life analysis. The independent variables were generated by following the study of McDonald and Galarneau²² given as:

$$X_{ij} = (1 - r^2)^{1/2} z_{ij} + r z_{i(p+1)} \quad (38)$$

where Z_{ij} are independent standard normal pseudorandom numbers, r^2 is the relationship between any two independent variables and p is the number of independent variables taken as three and seven in this study. The values of r^2 varies between 0.8, 0.9, 0.99 and 0.999 respectively. For $p=3$, the response variable is defined as:

$$Y_i = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e_i \quad (39)$$

where e_i is normally distributed with mean 0 and variance σ^2 . β is chosen such that $\beta'\beta = 1$ ²³. Samples of size 30, 50, and 100 were used. Values of σ are 1 and 5. The mean square error is calculated as:

$$MSE(\hat{\alpha}) = \frac{1}{2000} \sum_{j=1}^{2000} (\hat{\beta}_{ij} - \beta_i)' (\hat{\beta}_{ij} - \beta_i) \quad (40)$$

where $\hat{\beta}_{ij}$ is the estimate of the i^{th} parameter in j^{th} replication and β_i are the true parameter values. The MSE results are presented in Table 1 and Table 2. We observed the following:

1. All other alternative techniques studied in this work outperforms the OLS estimator at all the levels of multicollinearity.

Table 1. Estimated MSE for the simulation study when sigma=1.

P	n	p	OLS	Ridge	U-Ridge	KL	UKL1	UKL2	UKL3	UKL4
3	30	0.8	1.2575	1.1783	1.2011	1.2086	1.2235	1.1140	1.0996	1.1589
		0.9	3.7789	3.4543	3.5538	3.5654	3.6322	2.2056	2.9748	3.3636
		0.99	5.6205	3.2848	3.8963	3.9439	4.4144	1.9226	2.0469	2.8841
		0.999	35.7111	12.5735	18.3843	19.5479	24.0513	4.6898	4.5471	8.6644
	50	0.8	3.1817	2.9869	3.0504	3.0704	3.1070	1.8099	2.6533	2.9248
		0.9	3.1387	2.8611	2.9496	2.9555	3.0146	1.8909	2.5800	2.7765
		0.99	4.7483	3.2525	3.6586	3.6869	3.9912	2.0781	2.4117	2.9391
		0.999	21.4035	8.1820	11.4916	12.2398	14.7889	3.5507	3.4045	5.9699
	100	0.8	4.0550	3.9420	3.9793	3.9983	4.0171	2.5400	3.5609	3.9050
		0.9	1.9893	1.8763	1.9128	1.9249	1.9460	1.2788	1.6186	1.8409
		0.99	5.2450	4.1823	4.4893	4.4379	4.6780	2.9884	3.4428	3.9226
		0.999	14.0023	5.2709	7.4834	7.8224	9.5428	1.7324	1.6713	3.7563

P	n	p	OLS	Ridge	U-Ridge	KL	UKL1	UKL2	UKL3	UKL4
7	30	0.8	4.0218	3.0750	3.1889	3.5199	3.5851	1.3614	1.9079	2.6116
		0.9	4.7904	3.2139	3.3907	3.9152	4.0246	1.7000	1.9975	2.5728
		0.99	19.0010	6.7494	7.9171	12.1457	12.9583	8.1601	7.1691	3.8239
		0.999	159.6416	41.2828	52.2229	93.3516	101.1318	76.3306	73.9092	16.4346
	50	0.8	3.5577	2.8785	2.9656	3.2243	3.2697	1.5240	1.7956	2.4915
		0.9	5.4123	4.0309	4.2014	4.6759	4.7732	1.4829	2.1778	3.3137
		0.99	13.4754	4.6632	5.5231	8.6238	9.2025	6.0085	4.6971	2.3589
		0.999	121.0170	33.8018	42.0867	73.0005	78.7071	55.0748	53.4195	13.2499
	100	0.8	3.8136	3.4409	3.4917	3.6804	3.6991	1.1615	2.3239	3.1992
		0.9	4.0491	3.4280	3.5095	3.7816	3.8185	1.3207	2.2128	3.0562
		0.99	7.5441	3.9596	4.3415	5.4878	5.7412	2.9686	2.6924	2.7004
		0.999	42.4513	12.1138	15.0220	24.9592	27.0472	18.0973	16.6656	4.7009

Table 2. Estimated MSE for the simulation study when sigma=5.

p	N	p	OLS	Ridge	U-Ridge	KL	UKL1	UKL2	UKL3	UKL4
3	30	0.8	6.2685	2.9512	3.8065	3.934	4.5876	3.4963	2.1897	2.3464
		0.9	12.7983	6.1772	7.9023	8.2484	9.5299	6.7402	3.9396	4.9341
		0.99	86.9643	28.7447	43.3680	46.5598	57.8349	8.9211	6.2380	18.9046
		0.999	837.2886	261.9639	405.9761	437.9803	549.2998	62.0722	86.7562	165.7144
	50	0.8	5.8175	3.4525	4.1075	4.2205	4.6864	3.7063	3.1453	2.9308
		0.9	7.7474	4.0443	5.0257	5.2374	5.9512	4.2569	3.0837	3.3158
		0.99	48.2308	16.4888	24.4289	26.4166	32.4916	6.0867	4.03304	11.1893
		0.999	464.4477	143.1775	223.0953	243.7469	305.1109	35.2421	47.7284	90.4319
	100	0.8	5.5264	3.8849	4.3618	4.3957	4.7344	4.2461	3.9482	3.4782
		0.9	5.0219	2.6888	3.3142	3.3498	3.8251	3.4881	2.5229	2.2156
		0.99	33.7948	12.2441	17.6493	18.6328	22.8495	9.0504	3.5528	8.6170
		0.999	304.1409	89.4080	142.8193	153.6682	195.4571	12.0858	27.2452	54.1628
7	30	0.8	24.8444	8.5389	10.1199	15.8535	16.9275	7.0383	5.6513	4.3758
		0.9	44.3043	13.6507	16.5797	27.3928	29.4076	11.5369	10.8441	6.2566
		0.99	396.6272	105.4977	132.8158	235.8310	254.8663	125.2039	167.0124	40.2664
		0.999	3912.3412	1018.8512	1289.3604	2316.7446	2505.2394	1816.5141	1944.9425	380.5377
	50	0.8	17.5698	6.4007	7.5074	11.4719	12.2056	5.4997	3.8656	3.3115
		0.9	33.4275	11.1706	13.3419	21.3280	22.7780	9.3004	7.1973	5.3547
		0.99	293.0569	78.6266	98.9722	175.8942	189.8296	67.7918	110.9219	28.3353
		0.999	2915.6689	775.1140	978.3931	1745.1447	1884.4693	1239.4901	1424.9954	271.2904
	100	0.8	8.9550	4.4238	4.9084	6.4437	6.7548	3.4211	2.8158	2.8143
		0.9	13.9441	5.7784	6.6171	9.4022	9.9580	4.9346	3.3490	3.2245
		0.99	102.2963	29.2440	36.3967	61.5713	66.4953	19.4200	26.6929	9.8965
		0.999	981.0459	263.5057	333.4068	580.2188	628.6188	322.2568	448.7248	77.0239

2. The ridge estimator outperforms its unbiased version when the MSE is used as a criterion.
3. The proposed unbiased estimator (UKL) in this study outperform its K-L estimator counterpart.
4. There was a general better performance of the proposed estimator over all the estimators considered in this work though its performance is a function of the choice of biasing parameters.

In this study, the following trends about the mean square error and the factors in the simulation were observed:

1. There is a decrease in the MSE when there is an increase in the sample size at a particular level of multicollinearity.
2. Increase in the value of σ leads to a corresponding increase in the mean square errors of each of the estimators when other variables are kept constant.
3. An increase in the number of explanatory variables leads to a corresponding increase in the MSE of all estimators at varying level of multicollinearity and σ .

Application: poultry waste data

The poultry waste data adopted in this study was found and analyzed in Qian *et al.*^{24,25} and was also recently employed by Lukman *et al.*¹⁹. The study was aimed at modelling the high heating values of proximate-based model. The response variable is High Heating Values (HHV), while the independent variables are Fixed Carbon (FC), Volatile Matter (VM), and Ash (A). The linear regression model is:

$$HHV = \beta_0 + \beta_1 FC + \beta_2 VM + \beta_3 A + \varepsilon \quad (41)$$

where ε is the normally distributed random error term. In this study, the Jarque-Bera (JB) test was employed to know the distribution of the residual. The test statistic and its p-value are 0.6409 and 0.7258, respectively. The result shows that the residual in the model is normally distributed. We diagnosed if the model has the problem of multicollinearity. According to Lukman *et al.*¹⁴, the model suffered from the problem of multicollinearity because the variance inflation factors ($VIF_{FC} = 997.819$, $VIF_{VM} = 2163.504$, $VIF_{ASH} = 1533.782$) are greater than ten (10). Also, there is evidence of multicollinearity with the use of the condition number (CN).

Following Lukman *et al.*^{3,4}, moderate level of multicollinearity is observed if the CN ranges are between 100 and 1000 but severe multicollinearity is encountered if CN is greater than 1000. For effective modelling, we considered some alternative estimators to the ordinary least squared estimator in this study. These include the ridge estimator, unbiased ridge estimator, K-L estimator and the unbiased K-L estimator. The estimators' performance was examined using the mean square error. We also adopt the leave-one-out cross-validation to validate how well the estimators perform¹⁴. The performance of the estimator is assessed through the mean squared prediction error (MSPE). The estimator with the least MSE and MSPE is considered as the best. The result is available in Table 3.

From Table 3, the regression estimates of the following methods are the same: URR, UKL, and OLS as expected. They also possess a smaller mean squared error when compared with the OLSE. The estimators all exhibit

Table 3. Regression coefficients and MSE.

coef.	$\hat{\alpha}_{OLS}$	$\hat{\alpha}_{RR}$	$\hat{\alpha}_{URR}$	$\hat{\alpha}_{KL}$	$\hat{\alpha}_{UKL-1}$	$\hat{\alpha}_{UKL-2}$	$\hat{\alpha}_{UKL-3}$	$\hat{\alpha}_{UKL-4}$
α_0	167.72	102.099	167.72	144.50	167.72	167.72	167.72	167.72
α_1	-1.2704	-0.6143	-1.2704	-1.038	-1.2704	-1.2704	-1.2704	-1.2704
α_2	-1.5311	-0.8763	-1.5311	-1.299	-1.5311	-1.5311	-1.5311	-1.5311
α_3	-1.6840	-1.0267	-1.6840	-1.451	-1.6840	-1.6840	-1.6840	-1.6840
MSE	4521.1	1675.79	2752.17	3925.5	3757.45	3269.29	983.24	1645.56
MSPE	349.92	154.97	210.41	78.33	282.07	109.08	5.891	17.706

the same regression coefficient signs. The proposed estimator UKL demonstrated the best performance in terms of the MSE and the MSPE. Although its performance is a function of the biasing parameter k .

Conclusion

There is high inconsistency in the performance of the OLS estimator for parameter estimation in the linear regression model with multicollinearity problem. The estimator is unbiased but no longer has minimum variance. Due to this setback, in this study, the unbiased K-L estimator was developed and the properties of this new estimator was derived and established. It was found that the estimator is in the class of the unbiased estimator. An added advantage of this estimator over the OLS estimator is that it possesses minimum variance when multicollinearity is present. The superiority of the proposed estimator over the existing methods was theoretically established. The estimator is preferred to other estimators considered in this study.

Furthermore, the simulation and real-life results strengthened the findings of the theoretical comparison in terms of the mean squared error and the mean square prediction error. We recommend this new estimator for parameter estimation in a linear regression model with and without multicollinearity. In further studies, we will extend the new unbiased estimator to other generalized linear models such as the logistic regression model, Beta regression model, Gamma regression model etc.

Data availability

Underlying data

Zenodo: Regression Model to Predict the Higher Heating Value of Poultry Waste from Proximate Analysis. <http://doi.org/10.5281/zenodo.5078977²⁵>.

This project contains the following underlying data:

- hhv data.txt

Data are available under the terms of the [Creative Commons Attribution 4.0 International license](#) (CC-BY 4.0).

References

- Lukman AF, Ayinde K: **Review and classifications of the ridge parameter estimation techniques**. *Hacet J Math Stat*. 2017; **46**(5): 953–967. [Publisher Full Text](#)
- Ahmad S, Aslam M: **Another proposal about the new two-parameter estimator for linear regression model with correlated regressors**. *Commun Stat Simul Comput*. 2020. [Publisher Full Text](#)
- Lukman AF, Ayinde K, Binuomote S, et al.: **Modified ridge-type estimator to combat multicollinearity: Application to chemical data**. *J Chemom*. 2019a; **33**(5): e3125. [Publisher Full Text](#)
- Lukman AF, Adewuyi E, Oladejo N, et al.: **Modified Almost Unbiased Two-Parameter Estimator in linear regression model**. *IOP Conf Ser Mater Sci Eng*. 2019b; **640**: 012119. [Publisher Full Text](#)
- Hoerl AE, Kennard RW: **Ridge Regression: Applications to Nonorthogonal Problems**. *Technometrics*. 1970; **12**(1): 69–82. [Publisher Full Text](#)
- Swindel BF: **Good ridge estimators based on prior information**. *Commun Stat Theory Methods*. 1976; **5**(11): 1065–1075. [Publisher Full Text](#)
- Kejian L: **A new class of biased estimate in linear regression**. *Commun Stat Theory Methods*. 1993; **22**(2): 393–402. [Publisher Full Text](#)
- Yang H, Chang X: **A new two-parameter estimator in linear regression**. *Commun Stat Theory Methods*. 2010; **39**(6): 923–934. [Publisher Full Text](#)
- Özkale MR, Kaçiranlar S: **The restricted and unrestricted two-parameter estimators**. *Commun Stat Theory Methods*. 2007; **36**(15): 2707–2725. [Publisher Full Text](#)
- Sakallioğlu S, Kaçiranlar S: **A new biased estimator based on ridge estimation**. *Stat Papers*. 2008; **49**(4): 669–689. [Publisher Full Text](#)
- Dorugade AV: **A Modified Two-Parameter Estimator in Linear Regression**. *Statistics in Transition New Series*. 2014; **15**(1): 23–36. [Reference Source](#)
- Ayinde K, Lukman AF, Olarenwaju SO, et al.: **Some new adjusted ridge estimators of linear regression model**. *Int J Civ Eng Technol*. 2018; **9**(11): 2838–2852. [Reference Source](#)
- Kibria BMG, Lukman AF: **A new ridge-type estimator for the linear regression model: Simulations and applications**. *Scientifica (Cairo)*. 2020; **2020**: 9758378. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Lukman AF, Kibria GBM, Ayinde K, et al.: **Modified One-Parameter Liu Estimator for the Linear Regression Model**. *Model Simul Eng*. 2020b; **2020**: 1–17. [Publisher Full Text](#)
- Aslam M, Ahmad S: **The modified Liu-ridge-type estimator : a new class of biased estimators to address multicollinearity**. *Commun Stat Simul Comput*. 2020. [Publisher Full Text](#)
- Crouse RH, Jin C, Hanumara RC: **Unbiased ridge estimation with prior information and ridge trace**. *Commun Stat Theory Methods*. 1995; **24**: 2341–2354. [Publisher Full Text](#)
- Sakallioğlu S, Akdeniz F: **Unbiased Liu estimation with prior information**. *Int J Math Sci*. 2003; **2**(1): 205–217. [Reference Source](#)
- Wu J: **An unbiased two-parameter estimation with prior information in linear regression model**. *ScientificWorldJournal*. 2014; **2014**: 206943. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Lukman AF, Ayinde K, Aladeitan B, et al.: **An unbiased estimator**

- with prior information. *Arab J Basic Appl Sci.* 2020a; **27**(1): 45–55.
[Publisher Full Text](#)
20. Farebrother RW: **Further results on the mean square error of ridge regression.** *J R Stat Soc Ser B Methodol.* 1976; **38**(3): 248–250.
[Publisher Full Text](#)
 21. Trenkler G, Toutenburg H: **Mean squared error matrix comparisons between biased estimators an overview of recent results.** *Stat Pap.* 1990; **31**(1): 165–179.
[Publisher Full Text](#)
 22. McDonald G, Galarneau DI: **A Monte Carlo evaluation of some ridge-type estimators.** *J Am Stat Assoc.* 1975; **70**(350): 407–416.
[Publisher Full Text](#)
 23. Newhouse JP, Oman SD: **An evaluation of ridge estimators.** A report prepared for United States Air Force project RAND. 1971.
[Reference Source](#)
 24. Qian X, Lee S, Soto A, *et al.*: **Regression model to predict the higher heating value of poultry waste from proximate analysis.** *Resources.* 2018; **7**: 39.
[Publisher Full Text](#)
 25. Xuejun Q, Seong L, Ana-maria S, *et al.*: **Regression Model to Predict the Higher Heating Value of Poultry Waste from Proximate Analysis.** [Data set]. *Zenodo.* 2018.
<http://www.doi.org/10.5281/zenodo.5078977>

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 16 September 2021

<https://doi.org/10.5256/f1000research.58523.r92410>

© 2021 Arashi M. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Mohammad Arashi

Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Iran

Report on the paper “Unbiased K-L estimator for the linear regression model”

In this study, an unbiased estimator is developed for multicollinear situations. Analytical comparisons are carried out to demonstrate the superiority of the proposed estimator over some existing ones in the literature using the MSE of estimation and MSE of prediction. Numerical analyses, including Monte Carlo simulation and read data, are conducted to support the findings. The mathematical results sound correct, and the topic is eye-catching. Advancing estimation strategies to combat multicollinearity is improving; however, they are mostly biased. In this study, an unbiased version is proposed. I recommend making a minor revision addressing the following comments before acceptance.

1. Everywhere applies, change “ordinary least square” to “ordinary least squares”.
2. What is K-L in the Abstract? Whenever you use an abbreviation for the first time, define it completely.
3. The authors claim, “However, the estimators are biased, though they possess a smaller mean squared error when compared to the OLS estimator” in the Abstract, which is not correct.
4. Equation (2) exists if H is invertible.
5. To improve the literature and cover some existing results, the authors may refer to a couple of related studies such as Arashi et al. (2021a, 2021b), Norouzirad, and Arashi (2019), and Saleh et al. (2019). Since the core of this research is oriented on the multicollinearity problem and its defects in estimation, knowing remedies and strategies to combat multicollinearity comes to be important. The ridge regression (RR) estimation and other estimation strategies built upon it such as shrinkage ridge or scaled lasso thus are

important strategies to be used in this context.

6. Provide more details for obtaining G in Eq. (11).

7. After Eq. (16), what is $Z'Z = E = E_?$

8. For selecting ridge parameter k in Eqs (32)-(37), provide a reference.

9. In Eq. (37), k should read $\hat{k}$.

10. In Eq. (38), the pseudo normal variables are denoted by small z ; however, capital Z has defined the line after.

11. It is helpful to provide a table for the real data used in the Application, although a data availability statement is added.

References

1. Arashi M, Asar Y, Yüzbaşı B: SLASSO: a scaled LASSO for multicollinear situations. *Journal of Statistical Computation and Simulation*. 2021. 1-14 [Publisher Full Text](#)
2. Arashi M, Roozbeh M, Hamzah NA, Gasparini M: Ridge regression and its applications in genetic studies. *PLoS One*. 2021; **16** (4): e0245376 [PubMed Abstract](#) | [Publisher Full Text](#)
3. Norouzirad M, Arashi M: Preliminary test and Stein-type shrinkage ridge estimators in robust regression. *Statistical Papers*. 2019; **60** (6): 1849-1882 [Publisher Full Text](#)
4. Saleh AKME, Arashi M, Kibria BMG: Theory of Ridge Regression Estimation with Applications. *Wiley*. 2019.

Is the work clearly and accurately presented and does it cite the current literature?

Partly

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Partly

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Shrinkage estimation; High-dimensional analysis; Penalized regression

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 13 September 2021

<https://doi.org/10.5256/f1000research.58523.r92412>

© 2021 Amin M. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Muhammad Amin

Department of Statistics, University of Sargodha, Sargodha, Pakistan

Report on manuscript "Unbiased K-L estimator for the linear regression model"

In this paper, the authors introduced a new unbiased KL estimator for the linear regression model to overcome the effect of multicollinearity. The paper is original and deals with a topic of interest. This paper can be indexed after incorporating the following points:

1. Page 1, paragraph 1, line 3, the estimator should be the OLS estimator
2. In the abstract, the conclusion is not appropriate, I request the authors to rewrite the conclusion
3. Page 3, above equation 2, "the model is defined as" should be "the OLS estimator is defined as"
4. Page 3, change "of the two-parameter estimator by Ozkale and Kaciranlar9." to "of the two-parameter estimator based on the work of Ozkale and Kaciranlar9."
5. Page 3, introduction the last paragraph, change "a new unbiased technique is developed based on the K-L estimator" to "a new unbiased K-L estimator is developed for the linear regression model"
6. Below equation 4, "According to 16, the unbiased ridge estimator with the introduction", the authors should be clear is 16 an equation or a reference number?
7. Above equation 6, "Modified ridge-type method proposed by 3 is given as follows", here, indicate the author's name with reference number
8. Write the reason for your proposed estimator over other estimators in the last paragraph of

the introduction section

9. The expression below equation 7, change 1 to I
10. Above equation 8, "Recently13, proposed the K-L estimator", indicate here the author's names and reference numbers
11. In the results section line 1, change "The independent variables were generated" to "The correlated explanatory variables were generated"
12. Equation 19 is not correct, authors should correct this equation
13. All simulation results behaviour is not indicated which should be indicated - when and which estimator over the other gives better performance?
14. There are some grammatical issues that should be corrected

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Partly

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Linear and generalized linear models, Biased Estimation Methods, Outlier Analysis

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 03 September 2021

<https://doi.org/10.5256/f1000research.58523.r92414>

© 2021 Oyeronke Alaba O. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Oluwayemisi Oyeronke Alaba

Department of Statistics, University of Ibadan, Ibadan, Nigeria

In the abstract, inefficiency is better used than performance drops when multicollinearity is present. The new unbiased estimator should be clearly stated and the reason why there is a need to modify it, which was clearly stated in the body of the work.

Why did you choose the biasing parameter - did you check the unbiasedness of KL estimator, what were the limitations in previous studies before the need to modify it? You only stated that you classified it as a biased estimator with a single biasing parameter. The results of the simulation did not display well, but it will be better if few figures could be picked to discuss the results

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Statistical Modelling, Econometrics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research