International Journal of Science and Technology Research Archive

ISSN: 0799-6632 (Online)

Journal homepage: https://sciresjournals.com/ijstra/

(RESEARCH ARTICLE)

Check for updates

Generalized kibria-lukman estimator for multicollinearity in linear regression models: theoretical insights and comparative analysis

Ayanlowo E.A [1], Oladapo D.I [2], Odeyemi A.S [3] and Obadina G.O [4, *]

[1] *Department of Basic Sciences, Babcock University. Ilishan-Remo, Ogun State. Nigeria*
[2] *Department of Mathematical Sciences, Adeleke University, Ede, Osun State. Nigeria.*
[3] *Department of Statistics, University of Fort Hare Alice, Eastern Cape, South Africa.*
[4] *Department of Statistics, Olabisi Onabanjo University, Ago-Iwoye, Ogun State, Nigeria*

International Journal of Science and Technology Research Archive, 2024, 07(02), 114-119

Publication history: Received on 13 November 2024; revised on 22 December 2024; accepted on 24 December 2024

Article DOI: https://doi.org/10.53771/ijstra.2024.7.2.0073

Abstract

Multicollinearity, a common issue in regression models caused by high correlations among explanatory variables, undermines the stability and reliability of traditional estimators like Ordinary Least Squares (OLS). This study investigates the Generalized Kibria-Lukman (GKL) estimator, introduced by Dawoud et al. (2022), which uses a flexible biasing parameter to address the inflated variances typical in multicollinear datasets. Through comprehensive simulation studies and empirical testing, we compare the GKL estimator's performance with other biased estimators, including ridge regression and the Liu estimator, focusing on Mean Squared Error (MSE) as the primary evaluation metric. The results demonstrate that the GKL estimator consistently achieves lower MSE values, particularly in highly multicollinear conditions, underscoring its effectiveness as a robust alternative for improving accuracy in regression models where traditional methods struggle. These findings highlight the GKL estimator's potential as a superior choice in complex, multicollinear regression environments.

Keywords: Multicollinearity; Generalized Kibria-Lukman Estimator; Regression Models; Biasing Parameter; Mean Squared Error (MSE)

1. Introduction

Multicollinearity is a common issue in linear regression analysis where explanatory variables exhibit high correlation, potentially leading to large variances of coefficient estimates. This issue poses significant challenges for Ordinary Least Squares (OLS) estimation, as high multicollinearity inflates the standard errors of estimated coefficients, causing the model to produce unreliable and even misleading inferences. As such, estimators may yield theoretically inconsistent results, where coefficient signs may contradict theoretical expectations.

Various methods have been developed to address multicollinearity, including the ridge regression estimator proposed by Hoerl and Kennard (1970), the Liu estimator (1993), and more recently, the Kibria-Lukman (KL) estimator, which introduces a biasing parameter to stabilize estimates. Each of these biased estimators improves upon the OLS by controlling the variance of coefficients at the expense of introducing a small bias, thus providing more reliable estimates in the presence of multicollinearity. Dawoud et al. (2022) introduced a generalization of the KL estimator, termed the Generalized Kibria-Lukman (GKL) estimator, which features a flexible biasing parameter that adapts across observations. The GKL estimator aims to improve the efficiency of regression models particularly in scenarios of severe multicollinearity by optimizing the mean squared error (MSE).

* Corresponding author: Obadina G.O

This study investigates the theoretical properties of the GKL estimator, conducts simulation studies to evaluate its performance in comparison with OLS, generalized ridge, and Liu estimators, and applies the GKL estimator to a real-world dataset to validate its practical applicability.

2. Materials and Methods

2.1. Linear Regression Model and Multicollinearity

Consider a standard linear regression model:

$$y = X\beta + \epsilon$$

where y is the vector of observations on the dependent variable, X is the matrix of explanatory variables, β is the vector of regression coefficients, and ϵ is the vector of normally distributed errors with mean zero and variance $\sigma^2 I_n$.

When multicollinearity is present, the columns of X are highly correlated, making the estimation of β problematic. This multicollinearity inflates the variances of the OLS estimates of β, which can produce unreliable estimates with large standard errors. As a result, confidence intervals for coefficients widen, and significance tests lose reliability, often leading to erroneous inferences.

2.2. Biased Estimators

To address multicollinearity, biased estimators are commonly employed, including the following:

Ridge Regression Estimator: Introduced by Hoerl and Kennard, the ridge estimator addresses multicollinearity by adding a regularization term to the OLS estimator:

$$\hat{\beta}_{ridge} = (X'X + kI)^{-1}X'y$$

where k is a constant that shrinks the estimated coefficients and reduces variance.

Liu Estimator: This estimator introduces a biasing parameter d that also addresses the high variance caused by multicollinearity:

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

where $0 < d < 1$.

Kibria-Lukman (KL) Estimator: This ridge-type estimator was proposed to minimize the adverse effects of multicollinearity by employing a different form of shrinkage. The KL estimator is defined as:

$$\hat{\beta}_{KL} = (X'X + kI)^{-1}(X'X - kI)\hat{\beta}_{OLS}$$

where $k > 0$ serves as the biasing parameter.

2.3. Generalized Kibria-Lukman (GKL) Estimator

The GKL estimator generalizes the KL estimator by introducing observation-specific biasing parameters, allowing the estimator to adapt to varying levels of multicollinearity across observations. This estimator is defined as:

$$\hat{\beta}_{GKL} = (X'X + K)^{-1}(X'X - K)\hat{\beta}_{OLS}$$

where K is a diagonal matrix with observation-specific biasing parameters, optimizing MSE in high-multicollinearity settings. By adjusting biasing parameters for each observation, the GKL estimator aims to achieve a balance between bias and variance, yielding more stable and reliable estimates.

3. Simulation Study

The simulation study evaluates the performance of the GKL estimator against OLS, ridge, and Liu estimators under various levels of multicollinearity and error variance. The following simulation parameters were used:

Parameter	**Values**
Sample sizes	50, 100, 150
Error variance (σ)	1, 5, 10
Correlation among predictors (ρ)	0.8, 0.9, 0.99
Number of predictors (p)	3, 7

For each combination of parameters, 1,000 replications were conducted to estimate the Mean Squared Error (MSE) for each estimator.

Table 1 Simulation Results for Mean Squared Error (MSE) with 3 Predictors ($\boldsymbol{p = 3}$)

Sample Size n	**Error Variance σ**	**Predictor Correlation ρ**	**OLS MSE**	**Ridge MSE**	**Liu MSE**	**KL MSE**	**GKL MSE**
50	1	0.8	0.1249	0.1094	0.1021	0.0918	0.0871
		0.9	0.2260	0.1829	0.1723	0.1604	0.1483
		0.99	2.0641	1.1439	1.1028	1.0842	0.9845
100	5	0.8	3.1235	1.7550	1.5322	1.4590	1.3072
		0.9	5.6491	2.8600	2.6431	2.5102	2.2487
		0.99	51.6036	22.2378	20.9123	18.7532	16.5468

The simulation results displayed in Table 1 reveal the performance of various estimators—Ordinary Least Squares (OLS), Ridge, Liu, Kibria-Lukman (KL), and Generalized Kibria-Lukman (GKL)—under different conditions of multicollinearity and sample sizes. The Mean Squared Error (MSE) serves as the evaluation metric across these varying conditions, providing insight into the relative efficiency and stability of each estimator.

For a sample size of n=50 and low error variance (σ=1), as predictor correlation ρ increases from 0.8 to 0.99, the MSE values for all estimators rise. However, this increase is markedly less pronounced for biased estimators, particularly for KL and GKL. At ρ=0.8, OLS exhibits an MSE of 0.1249, whereas Ridge, Liu, KL, and GKL achieve progressively lower MSEs, with GKL attaining the lowest MSE at 0.0871. This trend persists as ρ intensifies; when ρ=0.9, GKL's MSE (0.1483) remains lower than that of all other estimators. At an extremely high correlation level (ρ=0.99), OLS reaches an MSE of 2.0641, signifying substantial inefficiency, while GKL maintains a comparatively lower MSE of 0.9845. This indicates that GKL handles multicollinearity far better than OLS and even outperforms Ridge, Liu, and KL in controlling the error under high multicollinearity conditions.

Increasing the sample size to n=100 and raising error variance to σ=5 provides further evidence of GKL's robust performance. At ρ=0.8, OLS has an MSE of 3.1235, indicating notable inefficiency compared to the biased estimators, particularly GKL, which records an MSE of 1.3072—the lowest among all estimators tested. As correlation increases to ρ=0.9, OLS's MSE rises to 5.6491, while GKL retains the lowest MSE at 2.2487, showcasing its superior bias-variance trade-off. In extreme multicollinearity (ρ=0.99), OLS's MSE skyrockets to 51.6036, an indicator of its substantial inefficiency under severe multicollinearity. Meanwhile, GKL exhibits a comparatively stable MSE of 16.5468, suggesting that it maintains better accuracy and efficiency under these challenging conditions than the other estimators.

The MSE comparisons in Table 1 illustrate the consistent advantage of the GKL estimator across different levels of predictor correlation and sample sizes. As multicollinearity intensifies, GKL consistently exhibits the lowest MSE values, highlighting its robustness in handling high-dimensional multicollinear data.

Table 2 Simulation Results for MSE with 7 Predictors ($\boldsymbol{p = 7}$)

Sample Size n	Error Variance σ	Predictor Correlation ρ	OLS MSE	Ridge MSE	Liu MSE	KL MSE	GKL MSE
50	10	0.8	41.4272	21.1839	19.7721	18.6023	15.5082
		0.9	77.9186	39.4124	37.0942	34.6021	28.5547
		0.99	738.6690	370.3048	362.0176	334.2045	265.3667

In Table 2, the simulation results reveal the impact of an increased number of predictors (p=7), a high error variance (σ=10), and varying degrees of predictor correlation (ρ) on the Mean Squared Error (MSE) across five estimators: Ordinary Least Squares (OLS), Ridge, Liu, Kibria-Lukman (KL), and Generalized Kibria-Lukman (GKL). These results underscore the GKL estimator's relative efficiency in managing both multicollinearity and a larger predictor set under high variance conditions.

For a sample size of n=50 with moderate multicollinearity (ρ=0.8), we observe substantial differences in MSE values across estimators. The OLS estimator records a notably high MSE of 41.4272, reflecting its poor handling of both multicollinearity and high variance. Ridge, Liu, and KL estimators show progressively better performance, with KL reaching an MSE of 18.6023. However, GKL outperforms all, achieving the lowest MSE of 15.5082—demonstrating its superior ability to balance bias and variance under moderate multicollinearity and substantial predictor variance.

As predictor correlation intensifies to ρ=0.9, the MSE for each estimator rises, although the magnitude of this increase varies significantly. OLS, with an MSE of 77.9186, shows a nearly twofold increase, emphasizing its sensitivity to higher multicollinearity. In comparison, Ridge, Liu, and KL maintain better control over MSE growth, with KL reaching 34.6021. GKL, however, stands out with an MSE of 28.5547, once again demonstrating its robustness by yielding the lowest error even as multicollinearity intensifies. This substantial reduction in MSE for GKL relative to other estimators highlights its efficacy in balancing the bias introduced by multicollinearity with the need for stable variance in parameter estimates.

Under extreme multicollinearity (ρ=0.99), the efficiency of each estimator is further strained. OLS shows a dramatic escalation in MSE, soaring to 738.6690, reflecting its severe inefficiency in highly multicollinear data. Ridge and Liu estimators demonstrate some resilience, but still register very high MSE values of 370.3048 and 362.0176, respectively, indicating that even these biased estimators struggle under such extreme conditions. KL reduces the MSE further to 334.2045, but the GKL estimator achieves the most significant improvement, attaining the lowest MSE of 265.3667. This marked reduction emphasizes GKL's unique strength in managing the compounded complexity of seven highly correlated predictors, high variance, and limited sample size.

Table 2 illustrates the consistent advantage of the GKL estimator across escalating levels of predictor correlation and error variance. Particularly under severe multicollinearity (ρ=0.99), the GKL estimator demonstrates resilience, maintaining the lowest MSE compared to all other estimators tested. The results reinforce the GKL estimator's adaptability in handling data with a large number of correlated predictors and significant error variance, making it a highly effective choice for complex regression models where traditional approaches like OLS and even other biased estimators fall short.

4. Empirical Application

To assess practical performance, we applied the GKL estimator to a real-world dataset with known multicollinearity among predictors, such as the Portland cement dataset. The dependent variable is the heat evolved, with four predictors showing significant correlation.

The Mean Squared Error (MSE) results for the Portland cement dataset presented in Table 3 provide a clear comparison of estimator performance under real-world multicollinearity. This dataset, known for its high correlation among explanatory variables, offers an opportunity to assess the efficacy of different estimators—Ordinary Least Squares (OLS), Ridge, Liu, Kibria-Lukman (KL), and Generalized Kibria-Lukman (GKL)—in producing stable and reliable parameter estimates.

Table 3 Comparison of MSE for Cement Dataset

Estimator	MSE
OLS	0.0638
Ridge	0.0581
Liu	0.0554
KL	0.0522
GKL	0.0486

The OLS estimator, despite being a standard approach in linear regression, exhibits the highest MSE at 0.0638. This relatively elevated MSE reflects OLS's well-documented limitations in handling multicollinearity. The high correlation among predictors inflates variances, leading to less stable estimates and greater error, which is evident from the higher MSE compared to other estimators specifically designed to manage multicollinearity.

The Ridge estimator, which introduces a penalty term to control large coefficient estimates, achieves a lower MSE of 0.0581. This decrease from the OLS MSE suggests that Ridge is able to provide more reliable estimates by balancing the trade-off between variance reduction and bias. By shrinking coefficients, Ridge mitigates some of the instability caused by multicollinearity, resulting in a tangible improvement in estimator efficiency.

The Liu estimator further reduces MSE to 0.0554, showing a slight yet notable improvement over Ridge. The Liu estimator's unique biasing parameter helps to control multicollinearity's effects more effectively than Ridge. This improvement highlights the estimator's ability to fine-tune biasing to stabilize coefficient estimates, achieving a better balance than Ridge between bias and variance.

The KL estimator, specifically formulated to address multicollinearity through a tailored biasing approach, demonstrates an even greater reduction in MSE, achieving a value of 0.0522. The lower MSE here indicates that KL is more effective at managing the impact of high correlations among predictors, which is critical in the Portland cement dataset. By introducing an optimal bias, KL provides a more stable estimation than both Ridge and Liu, thus yielding more precise parameter estimates and reducing error further.

The GKL estimator, the most advanced of the five tested, achieves the lowest MSE at 0.0486. This result confirms GKL's theoretical advantages over its predecessors, as it incorporates observation-specific biasing parameters that adapt more closely to the structure of multicollinearity within the dataset. By allowing each observation's contribution to vary, GKL is able to optimally control variance while minimizing error, outperforming OLS, Ridge, Liu, and KL estimators. The MSE reduction with GKL—by over 23% compared to OLS—underscores its effectiveness in delivering robust and reliable estimates under real-world multicollinearity conditions.

The MSE values in Table 3 illustrate a clear hierarchy among the estimators, with GKL emerging as the most efficient. Each successive estimator builds upon the previous, introducing more sophisticated techniques to manage the adverse effects of multicollinearity. GKL's adaptable biasing approach demonstrates the greatest accuracy and stability, making it the preferred choice for regression models with complex multicollinearity, such as in the Portland cement dataset. This progression from OLS to GKL not only highlights the limitations of traditional estimators in the face of multicollinearity but also underscores the value of targeted modifications that improve estimator performance in complex real-world applications.

5. Conclusion

The Generalised Kibria-Lukman (GKL) estimator is an advanced regression tool capable of effectively addressing multicollinearity by flexibly adjusting its biasing parameters. This study confirms that the GKL estimator offers superior accuracy, evidenced by consistently lower MSE across simulation scenarios and empirical validation. The GKL estimator shows promise as a reliable alternative to traditional OLS and biased estimators, particularly in high-dimensional and multicollinear settings.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Dawoud, I., Abonazel, M.R., & Awwad, F.A. (2022). Generalized Kibria-Lukman Estimator: Method, Simulation, and Application. *Frontiers in Applied Mathematics and Statistics*, 8, 880086.

[2] Hoerl, A.E., & Kennard, R.W. (1970). Ridge Regression: Biased Estimation for Non-Orthogonal Problems. *Technometrics*, 12, 55–67.

[3] Liu, K. (1993). A New Class of Biased Estimate in Linear Regression. *Communications in Statistics - Theory and Methods*, 22, 393–402.