A Comparative Study of the Performances of the Principal Component Ridge and Principal Component Liu Estimators using Different Forms of Biasing Parameter

Omokova Mary Attah
Department of Statistics
University of Abuja, Abuja
Nigeria

Samuel Olayemi Olanrewaju
Department of Statistics
University of Abuja, Abuja
Nigeria

ABSTRACT

Multicollinearity is a significant issue in multiple linear regression that occurs when two or more independent variables are highly correlated. This correlation undermines the reliability and stability of regression estimates, making it challenging to isolate and interpret the individual effect of each predictor variable. In the presence of multicollinearity, traditional estimation methods like Ordinary Least Squares (OLS) become less effective, often resulting in inflated standard errors and less reliable statistical inference.

When multicollinearity exists, biased estimation techniques such as Ridge regression, the Liu estimator, and Principal Component-based estimators are frequently used. These estimators provide more stable and interpretable results when independent variables are correlated. Other estimators that are a combination of existing estimators have been formed. These include the Principal Component Ridge (PCRE) and Principal Component Liu (PCLIU) estimators. They further mitigate the adverse effects of multicollinearity. This study evaluates the performance of PCRE and PCLIU estimators under varying degrees of multicollinearity, sample sizes, and error variances. Seven distinct forms of the biasing parameter k, along with their generalized versions, are examined in this analysis. Originally introduced in 2019 by Fayose and Ayinde, these forms include the maximum, minimum, arithmetic mean, geometric mean, harmonic mean, mid-range, and median. Monte Carlo simulations, repeated 1,000 times, were conducted on regression models with four and seven predictors, across five levels of multicollinearity, three error variances, and eight sample sizes. The Mean Square Error (MSE) criterion was used for evaluation. Results indicate that the maximum form of the Principal Component Ridge estimator consistently outperforms others in terms of efficiency.

Keywords

Ordinary Least Squares, Ridge Regression Estimator, Liu Estimator, Principal Component Estimator, Principal Component Ridge Estimator, Principal Component Liu Estimator.

1. INTRODUCTION

Regression analysis is used to study the relationship between two variables, known as the dependent and independent variables. When only one independent variable is used to predict the value of the dependent variable, it is called simple linear regression. However, when two or more independent variables are used to predict the value of the dependent variable—also known as the response—it is called multiple linear regression.

The ordinary least squares (OLS) estimator is the traditional method used to estimate the parameters of a linear regression model. It aims to find a line or hyperplane that best fits the observed data by minimizing the sum of the squared residuals. The residuals are the differences between the predicted values and the actual observed values in the model. Under certain assumptions, known as the Gauss-Markov assumptions, OLS estimates are BLUE (Best Linear Unbiased Estimators) (Gujarati, 2021) [1]. This means that, among all unbiased linear estimators, the OLS estimates are the most efficient and thus the most precise.

One of the Gauss-Markov assumptions that must be satisfied is that the independent variables should not be highly linearly correlated. When this assumption does not hold, multicollinearity is said to exist in the linear regression model.

Multicollinearity frequently occurs in linear regression. When multicollinearity is present, it becomes difficult to distinguish the unique contribution of each independent variable. Although the ordinary least squares (OLS) estimator—the traditional method for estimating parameters in a linear regression model—still produces unbiased estimates in the presence of multicollinearity, these estimates exhibit large variances and covariances and are highly sensitive to small changes in the data set. This often results in wide confidence intervals (Gujarati, 2021; Kibria and Lukman) [1,2].

Multicollinearity can be detected using various metrics such as the variance inflation factor (VIF), condition index, and correlation matrices (Paul, 2006; Montgomery et al. 2021) [3,4].

Multicollinearity may arise due to the choice of the model specification or from the nature of the data collected (Gujarati, 2021) [1]. Several solutions have been proposed over the years to address multicollinearity in linear regression models. Paul (2006) [3] noted that if multicollinearity is caused by the model specification, then re-specifying the regression model may help reduce its impact. Two common approaches to re-specification include redefining the explanatory variables and eliminating one or more variables from the model (Paul, 2006; Khalaf & Iguernane, 2016) [2,5]. However, the removal of explanatory variables can reduce the predictive power of the model, particularly if the excluded variables have significant explanatory power related to the response variable and therefore may not provide a satisfactory solution.

Researchers have developed estimators to address the problem of multicollinearity in linear regression models. These estimators include the Ridge Estimator by Hoerl and Kennard (1970) [6] and the Liu Estimator by Liu (1993) [7]. Although these estimators produce biased estimates, they are often preferred over unbiased ones because they have a higher probability of being closer to the true population parameter value (Muniz & Kibria, 2009) [8].

In the ridge estimator proposed by Hoerl and Kennard (1970) [6], a non-zero value k called the ridge parameter is added to the diagonal of the $X'X$ matrix of the ordinary least squares (OLS) estimator This addition typically reduces the mean squared error (MSE) of the ridge regression estimator compared to that of the OLS estimator. When K=0. The ridge estimator equals the OLS estimator. Traditionally, K is determined graphically using the ridge trace plot developed by Hoerl and Kennard (1970) [6]. However, using the ridge trace plot is subjective.

Over time, different authors have proposed various methods to estimate the ridge parameter kk. Notable contributors include Hoerl and Kennard (1970) [6], Hoerl et al. (1975) [9], McDonald and Galarneau (1975) [10], Lawless and Wang (1976) [11], Dempster et al. (1977) [12], Troskie and Chalton (1996) [13], Firinguetti (1999) [14], Kibria (2003) [15], Khalaf and Shukur (2005) [16], Batah et al. (2008) [17], Kibria and Banik (2016) [18], Lukman and Ayinde (2017) [19], Lukman et al. (2020) [20], Fayose and Kayode (2019) [21], among others.

The ridge regression estimators can be classified into two main types: generalized ridge estimators and ordinary ridge estimators. A major challenge in applying ridge regression is selecting an appropriate value for the ridge parameter k. The solution obtained from the ridge estimator depends critically on the value of k used. Since the ideal value of k depends on the unknown population parameters, it can only be estimated from the data. Currently, there is no consensus on the best method for determining k.

In a study by Fayose and Ayinde (2019) [21] , seven different forms of the biasing parameter kk that perform more efficiently than the generalized form were proposed and used as kk values. These different forms include the maximum, minimum, arithmetic mean, geometric mean, harmonic mean, mid-range, and median.

Over the years, other researchers have developed combined estimators with the expectation that combining different estimators might inherit the advantages of both components. Baye and Parker (1984) [22] combined the Ridge estimator with Principal Component (PC) estimator to develop the r-k class estimator, which is an alternative method of dealing with multicollinearity. This new combined estimator was shown to perform better than the individual component elements (Baye and Parker,1984) [22]. Nomura and Ohkuba (1985) [23] compared the r-k class estimator with the Ridge and OLS estimators according to scalar mean square error criterion, showing that the r-k class estimator outperforms each of these individual estimators. Additionally, Sarkar (1996) [24] compared the performance of the OLS, Ridge and Principal Component estimator with respect to their Matrix Mean Square Error (MMSE). He derived necessary and sufficient conditions under which the r-k class estimator achieves a smaller MMSE than the OLS, Ridge and the Principal component estimators.

The article by Baye and Parker (1984) [22] motivated other researchers to investigate combinations of estimators to address multicollinearity. Kaciranlar and Sakallioglu (2001) [25] developed the r-d class estimators by combining the Liu and Principal Component (PC) estimators. Using the MSE criterion, they compared this new estimator with the ordinary least squares (OLS), PC and Liu estimators, demonstrating improved performance. Ozkale and Kaciranlar (2007) [26] further combined the PC estimator with the restricted least squares estimator to develop a new estimator, called the Restricted principal components regression (RPCR) estimator. Batah et al. (2009) [27] introduced the modified r–k class ridge regression (MCRR) estimator by combining the unbiased ridge regression (URR) estimator of Crouse et al. (1995) [28] with the PC estimator. Adegoke et al. (2016) [29], combined the Ridge and Liu estimators to improve efficiency in the presence of multicollinearity. Lukman et al. (2020) [20] proposed a new estimator by combining modified ridge type and principal component estimators. More recently Huang and Bai (2023) [30] combined the modified Kibria–Lukman and the principal component regression estimators to develop an estimator effective for use in the presence of multicollinearity.

The Principal Component Ridge (PCRE) estimator (Bayes &Parker, 1984) [22] and the Principal Component Liu (PCLIU) estimator (Kaciranlar & Sakallioglu, 2001) [25] are both designed to handle multicollinearity exists in a linear regression model. They are dimension reduction techniques that produce biased estimates but with reduced variance compared to the ordinary least squares (OLS) estimator. Both estimators combine principal component analysis (PCA) with either ridge regression (for PCRE) or Liu regression (for PCLIU).

Ozbey and Kaçıranlar (2015) [31] using the Mean squared error (MSE) demonstrated that when multicollinearity is present, both estimators provide more precise estimates than the OLS estimator. The shrinkage of regression coefficients for the Principal Component Ridge estimator is achieved through a non-zero constant k (Bayes &Parker, 1984) [22] while the Principal Component Liu estimator uses a shrinkage parameter d that takes a value between zero and one (Kaciranlar & Sakallioglu, 2001) [26].

Furthermore, the PCLIU estimator tends to perform better than the PCRE estimator when severe multicollinearity exists in the data, whereas the PCRE estimator outperforms the PCLIU under moderate multicollinearity conditions (Ozbey & Kaçıranlar, 2015) [31]. A common critique of both PCRE and PCLIU estimators is that since their principal components are linear combinations of the original correlated variables, interpreting the resulting coefficient estimates in terms of the original variables is difficult.

The aim of this study is to evaluate different forms of the biasing parameter used with the Principal Component Ridge (PCRE) and Principal Component Liu (PCLIU) estimators to identify which form yields the most efficient estimates. By determining the most efficient estimator, this research seeks to provide guidance on the optimal choice of biasing parameter form for use with either estimator in the presence of multicollinearity.

1.1 OLS ESTIMATOR

Consider the standard regression model:

$Y = X\beta + U$ (1) (1)

Where X is an n x p matrix with full rank, Y is an n x 1 vector of dependent variable, β is a p x 1 vector of unknown parameters, and U is the error term, such that $E(U) = 0$ and $E(UU^1) = \sigma^2 I_n$ provided

X^1X is invertible (Lukman et al. 2018), the OLS estimator is given as:

$$\hat{\beta} = (X'X)^{-\prime}X'Y \qquad (2)$$
(2)

Consider the regression model in equation (1) and letting Λ= diag $(\lambda_{1,}\lambda_2, \ldots, \lambda_p)$ be a pxp diagonal matrix of the eigenvalues of $X'X$, and T be a p × p orthogonal matrix whose columns are the eigenvectors associated with $\lambda_1, \lambda_2 \ldots . \lambda_p$ such that $TT' = T'T = I_P$.

Then,

$$X'X = T\Lambda T' \qquad (3)$$
(3)

and

$$\Lambda = T'X'XT \qquad (4)$$
(4)

Defining Z = XT, then

$$X=ZT' \qquad (5)$$
(5)

Putting eqn. (5) into eqn. (4)

$$\Lambda = T'X'XT = Z'Z \qquad (6)$$
(6)

Substituting eqn. (6) for X in eqn. (1), we obtain the equivalent model given as:

$$Y=ZT'\beta + U \qquad (7)$$
(7)

Let $T'\beta = \alpha$, then:

$$\beta=T\,\alpha \qquad (8)$$
(8)

and

$$Y = Z\alpha + U \qquad (9)$$
(9)

The ordinary Least Squares estimator of α, α_{OLS} is given as:

$$\hat{\alpha}_{OLS} = (Z'Z)^{-\prime}Z'y = \Lambda^{-\prime}Z'Y \qquad (10)$$
(10)

1.2 RIDGE REGRESSION ESTIMATOR (RE)

To solve the problem of multicollinearity, Hoerl and Kennard (1970) [6], developed the Ridge Regression estimator which works by adding a small value, K, to the diagonal elements of the X^1X matrix before computing the estimates. This value, k, balances the trade-off between fitting the data well and keeping the coefficients small. Using the model in equation (9), this corresponds to adding a small value, K, to the diagonal elements of the $Z'Z$ matrix before computing the estimates. Thus, the ridge regression estimator (RE) becomes:

$$\hat{\alpha}_{RE} = (Z'Z + KI)^{-\prime}Z'Y \qquad (11)$$
(11)

Where $Z'Z$ is a p x p product matrix of explanatory variables, $Z'Y$ is a p x 1 vector of the product of dependent and explanatory variables,

K = diagonal $(K_1, K_2, \ldots, K_P)$, $K_i = \frac{\hat{\sigma}^2}{\hat{\alpha}_i} \geq 0$. i= 1, 2, ..., p.

When $K_1 = K_2 = \cdots . = K_p = K$, K>0, then $\hat{\alpha}_{GRE} = \hat{\alpha}_{RE}$ and if K=0, then $\hat{\alpha}_{RE}$=$\hat{\alpha}_{OLS}$

In equation (11) above, K > 0 and *I* is an identity matrix. Note that if K=0 the ridge estimator (RE) reduces to the ordinary least squares (OLS) estimator. The values of k used in this study are those proposed by Hoerl and Kennard (1970) [6], and Fayose and Ayinde (2019) [21].

1.2.1 Hoerl and Kennard Generalized Form

$$K=\hat{\sigma}^2/\hat{\alpha}_i^2,\ i=1,2,3\ldots, p \qquad (12)$$
(12)

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{u_i{}^2}{n-p} \qquad (13)$$
(13)

is the MSE from the OLS regression

α_i is the ith element of the vector α from the OLS regression

p is the number of regressors, and n is the sample size.

1.2.2 Fayose and Ayinde Generalized Form

$$\frac{\hat{\sigma}}{\alpha_i{}^2}\left\{\left[\left(\frac{\hat{\alpha}_i{}^4\lambda_i{}^2}{4\hat{\sigma}^2}\right) + \left(\frac{6\hat{\alpha}_i{}^4\lambda_i}{\hat{\sigma}^2}\right)\right]^{\frac{1}{2}} - \left(\frac{\hat{\alpha}_i{}^2\lambda_i}{2\hat{\sigma}^2}\right)\right\} \qquad (14)\ (14)$$

$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{u_i{}^2}{n-p}$ is the MSE from the OLS regression

α_i is the ith element of the vector α from the OLS regression

p is the number of regressors, and n is the sample size.

λ_i is the ith eigenvalue of the Z^1Z matrix

Following Fayose and Ayinde (2019) [21], seven different forms of the biasing parameter k that perform more efficiently than the generalized form were also used as k values. These different forms are maximum, minimum, Arithmetic mean, Geometric mean, Harmonic mean, Mid-range and Median.

1.3 LIU ESTIMATOR

Liu (1993) [7], motivated by the interpretation of the ridge estimate developed an alternative biased estimator to overcome multicollinearity for the linear regression model presented in equation (9). This estimator modifies the calculation of the regression coefficients by incorporating a biasing parameter, typically denoted as d, which lies between 0 and 1. This parameter allows the estimator to shrink the regression coefficients towards zero, thereby reducing their variance and improving estimation stability. Using our equivalent model the Liu estimator can be written as:

$$\hat{\alpha}_{LIU} = [(Z'Z + \mathrm{I})^{-\prime}(Z'Z + dI)\ \hat{\alpha}_{OLS} \qquad (15)$$

where d is the Biasing parameter and is defined as:

$$d_{LIU} = 1 - \hat{\sigma}^2\left[\frac{\sum_{i=1}^{p}\frac{1}{\lambda_i(\lambda_i+1)}}{\frac{\sum_{i=1}^{p}\hat{\alpha}_i{}^2}{(\lambda_i+1)^2}}\right] \qquad (16)$$
(16)

1.4 PRINCIPAL COMPONENT ESTIMATOR

The main purpose of Principal Component Regression is to estimate the values of a response variable based on selected Principal Components of the explanatory variables. It is a multivariate technique developed by Hotelling (1933) for explaining a set of correlated variables using a reduced number of uncorrelated variables (principal components, or PCs) with maximum variances (Pongpiachan et al., 2024) [34]. The estimates produced using the Principal Component estimators are biased (Weeraratne et al., 2024) [35]. When using Principal Component Regression, two stages are involved. The first stage reduces the number of predictor variables in the model using principal component analysis. The second stage involves using the reduced variables obtained from principal component analysis in an ordinary least square (OLS) (Ayinde et al., 2012) [36]. From our regression model in equation (9), the columns of Z, which define a new set of orthogonal regressors, such as Z=$(Z_1, Z_2, \ldots Z_p)$=$[Z_r, Z_{p-r}]$ are referred to as principal components. The pxp matrix of eigen vectors T =$(t_1, t_2, \ldots t_p)$ can also be written as $[T_r, T_{p-r}]$ with descending eigen values $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ such that the last of these eigenvalues are approximately equal to zero. Thus equation (1) can be written as:

$$Y = X\beta + U$$

$=XTT'\beta + U$
$= X T_r T_r{}'\beta + X T_{p-r} T_{p-r}{}'\beta + U$
$= Z_r\alpha_r + Z_{p-r}\alpha_{p-r} + U$ (17)
(17)

Where Z_r contains PCs that are used in the regression model and Z_{p-r} contain PCs that are discarded from the model. Thus, the regression equation becomes:

$Y = Z_r\alpha_r + U$ (18)
(18)

The principal component estimator, $\hat{\alpha}_{PC}$ from this transformed equation is given as:

$\hat{\alpha}_{PC} = (Z_r{}'Z_r)^{-\prime} Z_r{}'y$ (19)

1.5 PRINCIPAL COMPONENT RIDGE (PCRE) ESTIMATOR

Baye & Parker (1984) [22] combined the Ridge and Principal Component estimator to form a new estimator called the Principal Component Ridge estimator to address the problem of multicollinearity in linear regression models. This estimator works by first transforming the original correlated predictors into a set of uncorrelated principal components and then applying ridge regression to these components. Using our transformed model in equation (9), the Principal Component estimator in equation (19) and the Ridge estimator in equation (11), the Principal Component Ridge estimator is defined as:

$\hat{\alpha}_{PCRE} = (Z_r'Z_r + KI)^{-\prime} Z_r'Y$ (20)
(20)

The finite sample properties of the Principal Component Ridge estimator are derived as:

1.5.1 Proof of Expected Value of $\boldsymbol{\hat{\alpha}_{PCRE}}$

$E(\hat{\alpha}_{PCRE}) = E((Z_r'Z_r + KI)^{-\prime} Z_r'Y)$ (21)
(21)
$= (Z_r'Z_r + KI)^{-\prime} Z_r'E(Y)$
$= (Z_r'Z_r + KI)^{-\prime} Z_r'Z\alpha$ (22)
(22)

1.5.2 Proof of Biasedness of $\boldsymbol{\hat{\alpha}_{PCRE}}$

$BIAS(\hat{\alpha}_{PCRE}) = E(\hat{\alpha}_{PCRE}) - \alpha$ (23)
(23)
$= (Z_r'Z_r + KI)^{-\prime} Z_r'Z\alpha - \alpha$
$= ((Z_r'Z_r + KI)^{-\prime} Z_r'Z - I)\alpha)$ (24)
(24)

Let $Q(r) = ((Z_r'Z_r + KI)^{-\prime} Z_r'Z - I)\alpha$, then,
$BIAS(\hat{\alpha}_{PCRE}) = Q(r)$ (25)
(25)

1.5.3 Proof of Variance of $\boldsymbol{\hat{\alpha}_{PCRE}}$

$VAR(\hat{\alpha}_{PCRE}) = VAR((Z_r'Z_r + KI)^{-\prime} Z_r'Y)$ (26)
(26)
$= (Z_r'Z_r + KI)^{-\prime} Z_r'VAR(Y) Z_r (Z_r'Z_r + KI)^{-\prime}$
$= \sigma^2 (Z_r'Z_r + KI)^{-\prime} Z_r'Z_r (Z_r'Z_r + KI)^{-\prime}$ (27)

1.5.4 Proof of Mean Square Error of $\boldsymbol{\hat{\alpha}_{PCRE}}$

Let $\varphi(r) = \sigma^2 (Z_r'Z_r + KI)^{-\prime} Z_r'Z_r (Z_r'Z_r + KI)^{-\prime}$, then,
$MSE(\hat{\alpha}_{PCRE}) = VAR(\hat{\alpha}_{PCRE}) +$
$BIAS(\hat{\alpha}_{PCRE})\left(BIAS(\hat{\alpha}_{PCRE})\right)'$ (28)
(28)
$MSE(\hat{\alpha}_{PCRE}) = \varphi(r) + Q(r)Q'(r)$ (29)
(29)

1.6 PRINCIPAL COMPONENT LIU (PCLIU) ESTIMATOR

The Principal Component Liu (PCLIU) estimator is a hybrid of the principal component regression and the Liu estimator. It was developed by Kacıranlar & Sakallıoglu (2001) [26]. The PCLIU estimator provides a regression approach for handling multicollinearity in linear models by combining both principal component regression and Liu estimation techniques. Specifically, it addresses regression bias by first creating an orthogonal set of principal components from the predictors using principal component analysis (PCA), which reduces the dimensionality of the data and removes correlations among the transformed predictors. After obtaining the principal components, the Liu estimator is applied to these components to bias the coefficient estimates toward zero by employing its biasing parameter, typically denoted as d. This biasing reduces the overall variance, thereby improving the stability of the estimates. Using our transformed model in (9), the Principal Component estimator in equation (19) and the Liu estimator in equation (15), the Principal Component Liu estimator is defined as:

$\hat{\alpha}_{PCLIU} = (Z_r'Z_r + \mathrm{I})^{-\prime} (Z_r'Z_r + dI)\, \hat{\alpha}_{PC}$ (30)
(30)

1.6.1 Proof of Expected Value of $\boldsymbol{\hat{\alpha}_{PCLIU}}$

$E(\hat{\alpha}_{PCLIU}) = E[(Z_r'Z_r + \mathrm{I})^{-\prime} (Z_r'Z_r + dI)\, \hat{\alpha}_{PC}]$ (31)
(31)
$= (Z_r'Z_r + \mathrm{I})^{-\prime} (Z_r'Z_r + dI)\, E(\hat{\alpha}_{PC})$
$E(\hat{\alpha}_{PCLIU}) = (Z_r'Z_r + \mathrm{I})^{-\prime} (Z_r'Z_r + dI)\, (Z_r'Z_r)^{-\prime} Z_r'Z\alpha$ (32)
(32)

1.6.2 Proof of Biasedness of $\boldsymbol{\hat{\alpha}_{PCLIU}}$

$BIAS(\hat{\alpha}_{PCLIU}) = (Z_r'Z_r + \mathrm{I})^{-\prime} (Z_r'Z_r + dI)\, (Z_r'Z_r)^{-\prime} Z_r'Z\alpha -$
α
$= [(Z_r'Z_r + \mathrm{I})^{-\prime} (Z_r'Z_r + dI)\, (Z_r'Z_r)^{-\prime} Z_r'Z - I]\, \alpha$ (33)
(33)
Let $J(d) = [(Z_r'Z_r + \mathrm{I})^{-\prime} (Z_r'Z_r + dI)\, (Z_r'Z_r)^{-\prime} Z_r'Z - I]\, \alpha$

Thus, $BIAS(\hat{\alpha}_{PCLIU}) = J(d)$ (34)
(34)

1.6.3 Proof of Variance of $\boldsymbol{\hat{\alpha}_{PCLIU}}$

$VAR(\hat{\alpha}_{PCLIU}) = VAR((Z_r'Z_r + K\mathrm{I})^{-\prime} (Z_r'Z_r + dI)\, \hat{\alpha}_{PC})$ (35)
(35)
$= (Z_r'Z_r + \mathrm{I})^{-\prime} (Z_r'Z_r + dI)\, \mathrm{VAR}(\hat{\alpha}_{PC})$
$= (Z_r'Z_r + \mathrm{I})^{-\prime} (Z_r^1 Z_r + dI)\, \sigma^2\, (Z_r'Z_r)^{-\prime}$
$= \sigma^2[(Z_r'Z_r + \mathrm{I})^{-\prime} (Z_r'Z_r + dI)\, (Z_r'Z_r)^{-\prime}$ (36)

1.6.4 Proof of Mean Square Error of $\boldsymbol{\hat{\alpha}_{PCLIU}}$

Let $K(d) = \sigma^2[(Z_r'Z_r + \mathrm{I})^{-\prime} (Z_r'Z_r + dI)\, (Z_r'Z_r)^{-\prime}]$, then,
$MSE(\hat{\alpha}_{PCLIU}) = VAR(\hat{\alpha}_{PCLIU}) +$
$BIAS(\hat{\alpha}_{PCLIU})\left(BIAS(\hat{\alpha}_{PCLIU})\right)'$ (37)
(37)
$MSE(\hat{\alpha}_{PCLIU}) = K(d) + J(d)J'(d)$ (38)
(38)

2.1 MATERIALS AND METHODS

2.1.1 Model Formulation for Monte Carlo Study

To investigate the performance of our proposed and existing estimators when multicollinearity is present, this study considers a multiple linear regression model of the form:

$$Y = X\beta + \varepsilon \quad (39)$$

(39)

where Y is an n × 1 vector of response variable, X is a known n × p full rank matrix of explanatory variables, β is an p × 1 vector of unknown regression parameters, ε is an n × 1 vector of errors such that $E(\varepsilon) = 0$ and $Var(\varepsilon) = \sigma^2 I_n$. The number of explanatory variables (p) used were four (4) and seven (7). The sample sizes considered vary to assess estimator performance across small and large datasets. Specifically, the sample sizes considered are: 10, 15, 20, 30, 40, 50, 100, and 250.

2.1.2 Procedure for Generating Explanatory Variables

The simulation procedure used by McDonald and Galarneau (1975) [10], Wichern and Churchill (1978) [37], Gibbons (1981) [38], Kibria (2003) [15], Dorugade and Kashid (2010) [39], Dorugade (2016) [40], and Lukman et al (2018) [32], Amalare et al. (2023) [41], is also be used to generate explanatory variables in this study. This is given as:

$$X_{ti} = (1-\rho^2)^{\frac{1}{2}} Z_{ti} + \rho Z_{tp} \quad (40)$$

(40)

t=1, 2, 3…, n. i=1, 2,…,p.

Where Z_{ti}'s are independent standard normal random variables with mean zero and unit variance, ρ is the correlation between any two explanatory variables and p is the number of explanatory variables. The ρ is specified so that the correlation between any two regressors is given as ρ^2. These explanatory variables are then standardized so that the matrix X^1X is in correlation form.

2.1.3 Criterion for Investigation of the Performance of Estimators

The quality and efficiency of each estimator are evaluated using the Mean Squared Error (MSE) criterion, a widely accepted metric to quantify the accuracy and variability of parameter estimates. The MSE of an estimator $\hat{\alpha}$ for the true parameter α is given as

$$MSE(\hat{\alpha}) = \frac{1}{1000}\sum_{l=1}^{1000}(\hat{\alpha}-\alpha)^1(\hat{\alpha}-\alpha) \quad (41)$$

(41)

Where $\hat{\alpha}$ is the estimates of any of the estimators being studied. For each simulated sample generated at given levels of multicollinearity, error variance, and sample size, the estimators are applied, and estimates $\hat{\alpha}$ are obtained. The estimated MSE is computed by averaging the squared deviations of $\hat{\alpha}$ from α across replications. The estimator achieving the smallest estimated MSE under given simulation conditions is considered the best-performing estimator.

All simulation routines, including the generation of explanatory variables, estimation of parameters for each estimator considered, and computation of MSEs, were implemented in the statistical software R (version 4.5.0). To ensure the MSE estimates were stable and reliable, 1000 simulation replications for each combination of sample size, number of explanatory variables, multicollinearity level, and error variance, were conducted. The Statistical Package for the Social Sciences (SPSS version 29.0) was used after simulation to perform ranking of estimators based on their computed MSE values. This facilitated a clear identification of the most efficient estimators across varying scenarios.

3. RESULTS

TABLE 1: Frequency of the Principal Component Ridge and Principal Component Liu estimators Over the Levels of Multicollinearity and Error Variance at Each Sample Size When There is Multicollinearity for p=4

	Sample size (n)									
Estimator	10	15	20	30	40	50	100	250	Total	Rank
RHKMAWPC	**7**	**4**	**4**	**2**	**3**	**2**	**1**	**0**	**23**	**1**
RFAMAWPC	**3**	**3**	**2**	**2**	**3**	**2**	**2**	**0**	**17**	**2**
LUHMWPC	**0**	**3**	**4**	**2**	**0**	**2**	**1**	**1**	**13**	**3**
LUMIWPC	**5**	**1**	**0**	**3**	**1**	**1**	**0**	**0**	**11**	**4**
RFAMIWPC	0	0	0	2	2	1	3	1	9	5.5
RHKMIWPC	0	0	0	0	2	0	2	5	9	5.5
RHKGNWPC	0	0	0	0	2	1	1	3	7	7
RFAGMWPC	0	0	1	1	1	1	1	1	6	8
RFAAMWPC	0	1	1	1	0	0	1	0	4	9.5
RHKMDWPC	0	0	2	1	0	0	0	1	4	9.5
RFAHMWPC	0	0	0	1	0	2	0	0	3	11
RFAGNWPC	0	0	0	0	0	1	0	1	2	12
LUGMWPC	0	0	0	0	0	0	0	1	1	16
RHKAMWPC	0	1	0	0	0	0	0	0	1	16
RHKMRWPC	0	1	0	0	0	0	0	0	1	16
RFAMDWPC	0	0	0	0	1	0	0	0	1	16
RHKGMWPC	0	0	0	0	0	1	0	0	1	16
RHKHMWPC	0	0	0	0	0	0	1	0	1	16
LUGNWPC	0	0	0	0	0	0	0	1	1	16
Total	15	14	14	15	15	14	13	15		

NOTE: Estimator with highest frequency is bolded.

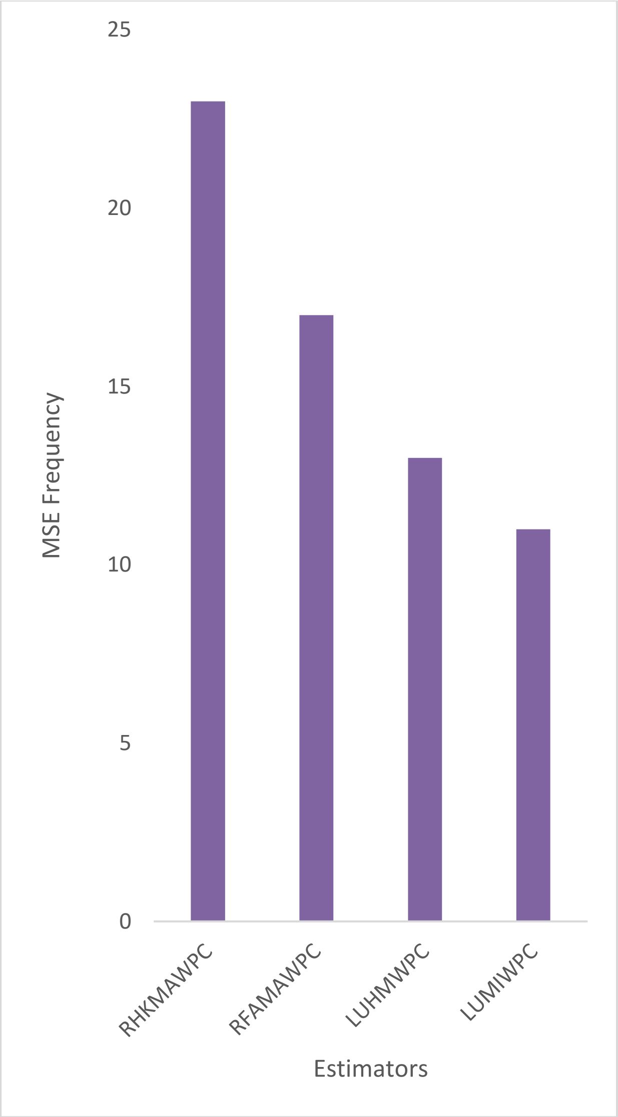

Figure 1: Graphical Representation of the Frequency of the Most Efficient Estimators Under Mean Square Error Criterion at Different Sample Sizes When There is Multicollinearity and p=4.

TABLE 2: Frequency of the Principal Component Ridge and Principal Component Liu estimators Over the Levels of Multicollinearity and Error Variance at Each Sample Size When There is Multicollinearity for p=7

	Sample Size (n)									
Estimator	10	15	20	30	40	50	100	250	Total	Ra
RHKMAWPC	**5**	**5**	**5**	**2**	**2**	**0**	**0**	**0**	**19**	**1**
RFAMAWPC	**3**	**2**	**2**	**4**	**2**	**2**	**1**	**0**	**16**	**2**
LUHMWPC	**0**	**0**	**2**	**1**	**4**	**3**	**3**	**1**	**14**	**3**
LUMIWPC	**4**	**4**	**0**	**0**	**0**	**0**	**1**	**1**	**10**	**4**
RHKGMWPC	0	0	0	2	1	2	3	1	9	5
RFAGNWPC	0	0	1	2	1	2	0	1	7	6.5
RHKMIWPC	0	0	0	0	0	0	2	5	7	6.5
RFAGMWPC	0	1	0	0	2	0	2	1	6	8.5
RHKGNWPC	0	0	1	0	1	1	1	2	6	8.5
LUGMWPC	0	0	0	4	0	0	0	1	5	10.5
RFAMIWPC	0	0	0	0	1	1	2	1	5	10.5
RHKAMWPC	2	1	1	0	0	0	0	0	4	12.5
RFAAMWPC	0	1	1	0	1	1	0	0	4	12.5
LUMAWPC	0	0	0	0	0	1	0	1	2	14
RFAMRWPC	0	0	1	0	0	0	0	0	1	16
RFAHMWPC	0	0	0	0	0	1	0	0	1	16
RHKHMWPC	0	0	0	0	0	1	0	0	1	16
Total	14	14	14	15	15	15	15	15	117	

NOTE : Estimators with highest frequencies are bolded.

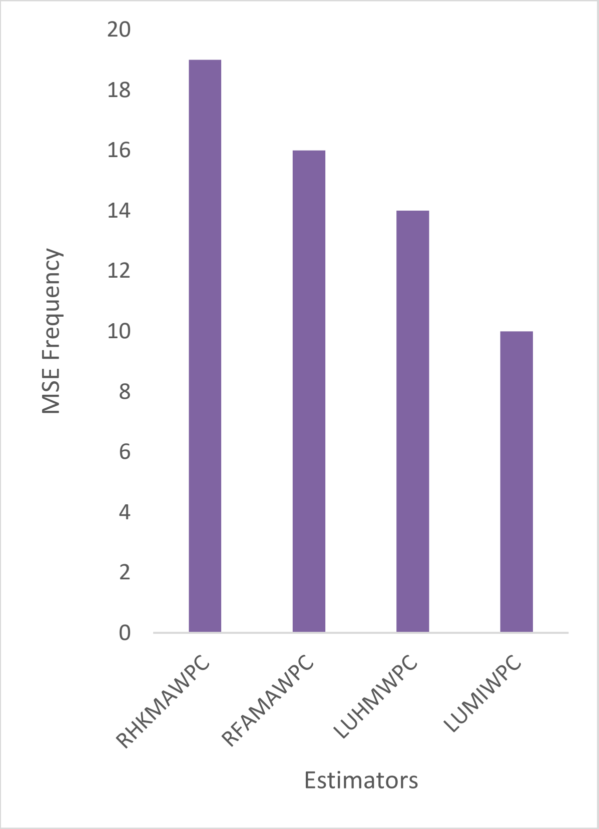

Figure 2: Graphical Representation of the Frequency of the Most Efficient Estimators Under Mean Square Error Criterion at Different Sample Sizes When There is Multicollinearity and p=7

4. DISCUSSION

Tables 1 and 2 present a summary of the frequency with which each estimator produces the minimum mean squared error (MSE) when counted across varying levels of multicollinearity and error variance at each examined sample size. Table 1 corresponds to the case when the number of explanatory variables p=4, while Table 2 corresponds to the case when p=7.

In Table 1, the frequencies of estimators achieving the minimum MSE across combinations of multicollinearity levels and error variances at different sample sizes are shown for p=4. The results indicate that the estimator RHKMAWPC, which is the principal component ridge estimator with the maximum variant of the biasing parameter proposed by Hoerl and Kennard (1970), achieves the lowest MSE most frequently among all estimators considered. This makes RHKMAWPC the most consistently efficient estimator under these settings. The second most frequent efficient estimator is RFAMAWPC, which is the principal component ridge estimator with the maximum variant of the biasing parameter proposed by Fayose and Ayinde (2019). Within the family of Principal Component Liu estimators, the harmonic mean variant LUHMWPC demonstrates superior efficiency, having the highest frequency for the number of times it achieves the lowest mean squared error. It is followed in lowest MSE performance frequency by the minimum variant LUMIWPC. These results illustrate the relative performance of different biasing parameter forms used with the principal component ridge and Liu estimators when p=4.

Figure 1 is a graphical representation of the results in Table 1. It represents the frequency of the four most efficient estimators under the mean squared error criterion at different levels of multicollinearity and error variance at each sample size when multicollinearity exists and p=4. From the graph, the estimator that had the lowest mean squared error the most times is RHKMAWPC.

In Table 2, the frequencies of estimators achieving the minimum MSE across combinations of multicollinearity levels and error variances at different sample sizes are shown for p=7. The pattern here is consistent with that observed for p=4; specifically, for the principal component ridge estimators, RHKMAWPC remains the estimator that most frequently attains the minimum MSE compared with the other estimators under consideration. It is again followed by RFAMAWPC in terms of frequency of achieving minimum MSE. Among the Principal Component Liu estimator variants, the harmonic mean form LUHMWPC continues to hold the highest efficiency, with the minimum variant LUMIWPC ranking second.

Figure 2 is a graphical representation of the findings of Table 2. It shows the frequency of each estimator being the most efficient at different multicollinearity levels, error variances, and sample sizes when p=7. This figure visualizes the sustained superior performance of RHKMAWPC as the sample size and parameter count increase.

These tables and figures demonstrate that the choice of biasing parameter form significantly affects estimator efficiency. RHKMAWPC, employing the maximum variant of the biasing parameter with the principal component ridge estimator, consistently provides the most efficient estimates in terms of minimum MSE among all estimators for different numbers of explanatory variables considered, under various levels of multicollinearity, error variance, and sample sizes. The harmonic mean variant LUHMWPC remains the most efficient among Principal Component Liu estimators for the settings considered.

5. CONCLUSION

Multicollinearity is a common problem in linear regression. It occurs when the explanatory variables are highly correlated. This causes traditional estimation methods like OLS to produce estimates that are inefficient and unstable. The use of biasing parameters in estimators yields estimates with less variance at the expense of a small increase in bias. Biasing estimators are incorporated in both the PCRE and PCLIU estimators. Research has shown that the performance of a biasing estimator depends on the form and value of its biasing parameters. This study evaluated the efficiency of different forms of the biasing parameter for both the Principal Component Ridge (PCRE) and Principal Component Liu (PCLIU) estimators for linear regression models under different conditions of multicollinearity, sample sizes, and error variances. Among the seven different forms of the biasing parameter considered, the maximum form produced the most efficient estimates when used with the principal component ridge estimator. This estimator, RHKMAWPC, had the smallest MSE for the Principal Component Ridge estimator. The estimator LUHMWPC, which is the harmonic mean form when used with the Principal Component Liu estimator, provided the most efficient estimates for the principal component Liu estimator. Overall, RHKMAWPC was the most efficient estimator for both principal component estimators since it had the lowest MSE across the different model settings that were considered. This study highlights the importance of selecting an appropriate form of the biasing parameter tailored to the estimator and data structure at hand. The results obtained suggest that while both PCRE and PCLIU estimators offer improvements over traditional methods, choosing the correct form of the biasing parameter can further improve performance. Model prediction and interpretability are enhanced when the efficiency of an estimator is improved. Future study can assess the robustness of these estimators when heteroscedasticity is found to exist in the model.

6. ACKNOWLEDGMENTS

Our thanks to the experts who have contributed towards development of the template.

7. REFERENCES

[1] Gujarati, D. N. (2021). Essentials of econometrics. Sage Publications.

[2] Kibria, B. G., & Lukman, A. F. (2020). A new ridge-type estimator for the linear regression model: simulations and applications. Scientifica, 2020(1), 9758378.

[3] Paul, R. K. (2006). Multicollinearity: Causes, effects and remedies. IASRI, New Delhi, 1(1), 58–65.

[4] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). Introduction to linear regression analysis. John Wiley & Sons.

[5] Khalaf, G., & Iguernane, M. (2016). Multicollinearity and a ridge parameter estimation approach. Journal of Modern Applied Statistical Methods, 15(2), 25. https://doi.org/10.22237/jmasm/1478002980

[6] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

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

[8] Muniz, G., & Kibria, B. M. G. (2009). On Some Ridge Regression Estimators: An Empirical Comparisons. Communications in Statistics - Simulation and Computation, 38(3), 621–630. https://doi.org/10.1080/03610910802592838

[9] Hoerl, A. E., Kannard, R. W., & Baldwin, K. F. (1975). Ridge regression: some simulations. Communications in Statistics, 4(2), 105–123. https://doi.org/10.1080/03610927508827232

[10] McDonald, G. C., & Galarneau, D. I. (1975). A Monte Carlo Evaluation of Some Ridge-Type Estimators. Journal of the American Statistical Association, 70(350), 407–416. https://doi.org/10.1080/01621459.1975.10479882

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

[12] Dempster, A. P., Schatzoff, M., & Wermuth, N. (1977). A Simulation Study of Alternatives to Ordinary Least Squares. Journal of the American Statistical Association, 72(357), 77–91. https://doi.org/10.1080/01621459.1977.10479910

[13] Troskie, CG* & Chalton, DO. "A Bayesian estimate for the constants in ridge regression." South African Statistical Journal 30, no. 2 (1996): 119-137.

[14] Firinguetti, L. (1999). A generalized ridge regression estimator and its finite sample properties: A generalized ridge regression estimator. Communications in Statistics - Theory and Methods, 28(5), 1217–1229. https://doi.org/10.1080/03610929908832353

[15] Kibria, B. M. G. (2003). Performance of some new ridge regression estimators. Communications in Statistics - Simulation and Computation, 32(2), 419–435. https://doi.org/10.1081/SAC-120017499

[16] Khalaf, G., & Shukur, G. (2005). Choosing Ridge Parameter for Regression Problems. Communications in Statistics - Theory and Methods, 34(5), 1177–1182. https://doi.org/10.1081/STA-200056836

[17] Batah, F. S. M., Ramanathan, T. V., & Gore, S. D. (2008). The efficiency of modified jackknife and ridge type regression estimators: a comparison. Surveys in Mathematics and its Applications, 3, 111-122.

[18] Kibria, B. G., & Banik, S. (2016). Some ridge regression estimators and their performances. Journal of Modern Applied statistical methods, 15, 206-238.

[19] Lukman, A. F., & Ayinde, K. (2017). Review and classifications of the ridge parameter estimation techniques. Hacettepe Journal of Mathematics and Statistics, 46(5), 953-967.

[20] Lukman, A. F., Ayinde, K., Oludoun, O., & Onate, C. A. (2020). Combining modified ridge-type and principal component regression estimators. Scientific African, 9, e00536.

[21] Fayose, T. S., & Ayinde, K. (2019). Different forms biasing parameter for generalized ridge regression estimator. International Journal of Computer Applications, 181(37), 2-29.

[22] Baye, M. R., & Parker, D. F. (1984). Combining ridge and principal component regression:a money demand illustration. Communications in Statistics - Theory and Methods, 13(2), 197–205. https://doi.org/10.1080/03610928408828675

[23] Nomura, M., & Ohkubo, T. (1985). A note on combining ridge and principal component regression. Communications in Statistics - Theory and Methods, 14(10), 2489–2493. https://doi.org/10.1080/03610928508828905

[24] Sarkar, N. (1996). Mean square error matrix comparison of some estimators in linear regressions with multicollinearity. Statist. Probab. Lett. 30:133–138.

[25] Kaçıranlar, S., & Sakallıoğlu, S. (2001). Combining the Liu Estimator and The Principal Component Regression Estimator. Communications in Statistics - Theory and Methods, 30(12), 2699–2705. https://doi.org/10.1081/STA-100108454

[26] Ozkale, M. R., & Kaciranlar, S. (2007). Comparisons of the unbiased ridge estimation to the other estimations. Communications in Statistics Theory and Methods, 36(4), 707.

[27] Batah, F. Sh. M., Özkale, M. R., & Gore, S. D. (2009). Combining Unbiased Ridge and Principal Component Regression Estimators. Communications in Statistics - Theory and Methods, 38(13), 2201–2209. https://doi.org/10.1080/03610920802503396

[28] Crouse, R. H., Chun Jin, & Hanumara, R. C. (1995). Unbiased ridge estimation with prior information and ridge trace. Communications in Statistics - Theory and Methods, 24(9), 2341–2354. https://doi.org/10.1080/03610929508831620

[29] Adegoke, A. S., Adewuyi, E., Ayinde, K., & Lukman, A. F. (2016). A comparative study of some robust ridge and liu estimators. Science World Journal, 11(4), 16-20.

[30] Huang, D., Huang, J., & Bai, D. (2023). Combination of the modified Kibria–Lukman and the principal component regression estimators. Communications in Statistics - Simulation and Computation, 1–16. https://doi.org/10.1080/03610918.2023.2292970

[31] Özbey, F., & Kaçiranlar, S. (2015). Evaluation of the predictive performance of the Liu estimator. Communications in Statistics-Theory and Methods, 44(10), 1981-1993.

[32] Lukman, A. F., Haadi, A., Ayinde, K., Onate, C. A., Gbadamosi, B., & Oladejo, N. K. (2018). Some improved generalized ridge estimators and their comparison. WSEAS Transactions on Mathematics, 17, 369–376.

[33] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24(6), 417–441. https://doi.org/10.1037/h0071325

[34] Pongpiachan, S., Wang, Q., Apiratikul, R., Tipmanee, D., Li, L., Xing, L., ... & Poshyachinda, S. (2024). Combined use of principal component analysis/multiple linear regression analysis and artificial neural network to assess the impact of meteorological parameters on fluctuation of selected PM2.5-bound elements. Plos one, 19(3), e0287187.

[35] Weeraratne, N., Hunt, L., & Kurz, J. (2024). Challenges of principal component analysis in high-dimensional settings when n < p. Preprint. https://doi.org/10.21203/rs.3.rs-4033858/v1

[36] Ayinde, K., Apata, E. O., & Alaba, O. O. (2012). Estimators of linear regression model and prediction under some assumptions violation. https://doi.org/ 10.4236/ojs.2012.25069

[37] Wichern, D. W., & Churchill, G. A. (1978). A comparison of ridge estimators. Technometrics, 20(3), 301–311. https://doi.org/10.1080/00401706.1978.10489675

[38] Gibbons, D. G. (1981). A simulation study of some ridge estimators. Journal of the American Statistical Association, 76(373), 131-139

[39] Dorugade, A. V., & Kashid, D. N. (2010). Alternative method for choosing ridge parameter for regression. Applied Mathematical Sciences, 4(9), 447–456

[40] Dorugade, A. V. (2016). Adjusted ridge estimator and comparison with Kibria's method in linear regression. Journal of the Association of Arab Universities for Basic and Applied Sciences, 21, 96–102.

[41] Amalare, A. A., Ayinde, K., & Onanuga, K. (2023). Parameter Estimation of Linear Regression Model with Multicollinearity and Heteroscedasticity Problems. The Journals of the Nigerian Association of Mathematical Physics, 65, 207-216.