International Journal of Development Mathematics Vol 3 Issue 1 Page No. 135 - 149

INTERNATIONAL JOURNAL OF DEVELOPMENT MATHEMATICS

ISSN: 3026-8656 (Print) | 3026-8699 (Online)

journal homepage: https://ijdm.org.ng/index.php/Journals

Performance of Proposed Ridge – PCA Estimators: Simulation Evidence and Real Data Applications

Remilekun E. Alabi[a,*], Olatayo O. Alabi[b] and Oluwadare O. Ojo[b]

[a]Department of Statistics, Federal Polytechnic Ado-Ekiti, Nigeria.
[b]Department of Statistics, Federal University of Technology, Akure, Nigeria.

A R T I C L E I N F O

Article history:
Received 02 January 2026
Received in revised form 11 Mrch 2026
Accepted 20 March 2026

Keywords:
Multicollinearity, Ridge Estimator, Principal Component Estimator, Hybrid Estimators, MSE

MSC 2020 Subject classification:
62P20

A B S T R A C T

This study examines the longstanding problem of multicollinearity in Gaussian linear regression models and introduce novel estimation techniques aimed at improving estimator stability and predictive accuracy. While Ordinary Least Squares (OLS) estimators are efficient under ideal conditions, their performance deteriorates in the presence of high collinearity among explanatory variables. The study proposes hybrid estimators that combine Ridge Regression and Principal Component Analysis (PCA). Four new ridge parameters were developed and integrated with PCA to construct hybrid estimators designed to address multicollinearity. Monte Carlo simulations were conducted under varying levels of multicollinearity, sample sizes, and error variances. The performance of these proposed estimators was compared to existing methods using Mean Squared Error (MSE) as the evaluation criterion. The results consistently indicate that the proposed Ridge PCA hybrid estimators, particularly minimum version of the Chand and Kibria (2024) ridge parameter combined with Principal Component estimator (PCARCK2MIN) and Chand and Kibria (2024) ridge parameter combined with Principal Component estimator (PCARCK1) outperform existing methods. Applications to real – life datasets including Portland cement and Longley data validates the efficiency and practical relevance of the proposed estimators for regression analysis under multicollinearity.

1. Introduction

Linear regression model is a statistical tool that investigates the relationship between an effect variable and one or more explanatory variables (Lukman and Ayinde, 2017). The model is simply defined as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i,\ i = 1,\ldots,n \tag{1}$$

where y_i is the effect variable, $x_{i1},\ldots, x_{ik}$ are the explanatory variables, $\beta_0, \beta_1,\ldots, \beta_k$ are the unknown parameters to be estimated, ε_i denotes the disturbance term. The model is simply a simple regression model when there is one explanatory variable. The parameters in model (1) are mostly estimated by the Ordinary Least Squares (OLS). OLS is generally preferred and possesses some desirable properties when the assumptions of the linear regression models are intact, this makes the model to be classical (Dawoud *et al.,* 2022). These include linear relationship among the explanatory variables; the disturbance terms must come from a Gaussian distribution and has non-scattered variance and others (Chatterjee and Hadi, 1977). In reality most of the aforementioned assumptions are normally violated. For instance, literature has shown that linear relationship often exists among explanatory variables which are termed multicollinearity (Fayose and Ayinde 2019). Multicollinearity is a phenomenon where two or more explanatory variables are highly correlated in a Gaussian linear regression (Fayose *et al,* 2023a and Aladesuyi *et al.,* 2025). There is tendency for perfect, strong or moderate linear dependency among the explanatory variables. Ordinary least squares is unbiased but inefficient when there is linear relationship among the explanatory variables (Gujarati *et al.,* 2012). It yields regression coefficients whose absolute values are too large and whose signs may actually reverse with negligible

* Corresponding author. Tel.: +2348032127090
E-mail address: alabi_re@fedpolyado.edu.ng (Alabi, R. E)
https://doi.org/10.62054/ijdm/0301.11

changes in the data (Buonaccorsi, 1996). If the multicollinearity is not perfect but high, the estimated coefficients can become unstable and highly sensitive to slight changes in the model leading to inflated standard errors and misleading inferences (Belseley *et al.,* 1980; Fayose & Ayinde, 2019). Consequently, reliable interpretation of the model parameters may be compromised undermining the credibility of the results derived from such analysis. The pursuit of effective techniques to mitigate the adverse effects of multicollinearity is of paramount importance in both theoretical and applied statistics.

Among the existing methods or approaches proposed to address or handle multicollinearity are Ridge Regression, Principal Component Analysis (PCA) among others. Both Ridge and PCA have emerged as prominent methodologies (Hoerl & Kennard, 1970; Jolliffe, 1986). Ridge regression offers a regularization technique that modifies the Ordinary least squares (OLSE) estimation process by introducing a penalty term (k) to the loss function thus allowing for the shrinkage of coefficient estimates towards zero (Hoerl and Kennard, 1970, Fayose *et al.,* 2023b). The ridge parameter (k) counteracts the inflation of variances associated with multicollinearity effectively enhancing the stability of the estimates and producing more reliable and responsive predictions.

In contrast, PCA serves as a dimensionality reduction technique that transforms the original correlated variables into a set of uncorrelated variables often called principal components (Jolliffe, 2002). By focusing on the principal components that explain the most variance in the data, PCA can help circumvent multicollinearity problem by ensuring that regression model utilizes orthogonal predictors. While both Ridge regression and PCA have demonstrated utility in addressing multicollinearity, each method presents notable limitations. Ridge regression while effective in controlling for multicollinearity does not completely eliminate correlation among predictors, it merely diminishes the variability of the coefficient estimates. Moreover, the choice of penalty parameter (k) can significantly influence the model's performance necessitating careful cross – validation. Meanwhile, PCA while adept at reducing dimensionality and addressing multicollinearity transforms the original predictors into a new set of components that may lack interpretability in the context of the original variables posing challenges for practical application and insight derivation (Jolliffe, 1986).

To harness the strengths of both Ridge regression and PCA while mitigating their respective limitations, the concept of hybrid or combined estimators has been introduced by different authors in recent literature. These combined estimators integrate Ridge parameter (k) with PCA estimator to form or create a more robust framework or new hybrid estimator to tackle multicollinearity in Gaussian linear regression model. Their approach demonstrated a significant reduction in mean squared error (MSE) compared to standard methods. Among other authors that have ′utilized these combined approaches are Zou and Hastie (2005), Buhlmann and Van de Geer (2011); Chang and Yang (2012); Huang and Wang (2018) and Alabi *et al.,* (2025) among others. The paper intends to comprehensively propose new ridge parameter k's and combine them with PCA to form new hybrid estimators to resolve the problem of multicollinearity within the Gaussian linear regression framework through simulation studies and real – life dataset. We compared the estimators' performance to that of some existing techniques. The existing estimators we compared them with are: Hoerl and Kennard (1975), Fayose and Ayinde (2019), Kibria and Lukman (2020) and Chand and Kibria (2024)

2. Methods

2. 1.1 Existing and Proposed Estimators

Consider the linear equation model in matrix form defined as:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{e} \quad \textbf{(2)}$$

Where y is the vector of response variables, X is $\boldsymbol{n} \times \boldsymbol{p}$ design matrix of explanatory variables or predictor $\boldsymbol{\beta}$ is $\boldsymbol{p} \times \boldsymbol{1}$ true vector or regression coefficients e ~ N$(\boldsymbol{0}, \boldsymbol{\sigma^2})$ is the disturbance term which is normally distributed with mean 0 and variance $\boldsymbol{\sigma^2}$.

2.1.2 The Ridge Estimator

The Ordinary Least Square estimator is defined as: $\hat{\boldsymbol{\beta}}_{\boldsymbol{OLS}} = \boldsymbol{S^{-1}X'Y}$ (3)

where $\boldsymbol{S} = (\boldsymbol{X'X})$ (4)

The ridge regression estimator is the mostly used estimators in literature for handling multicollinearity problem. The Generalized Ridge Estimator is defined as:

$$\boldsymbol{\beta_{GRRE}} = (\boldsymbol{S} + \boldsymbol{kI})^{-1}\boldsymbol{X}'\boldsymbol{Y} \quad (5)$$

where S is a p x p product matrix of concomitant variables, $X'Y$is a p x 1 vector of the product of effect and explanatory variables, k = diagonal ($k_1, k_2, \ldots, k_n$), $k_i \geq 0$, i = 1, 2,..., n. k is a non – negative constant called biasing or ridge parameter. When k = 0, equation (5) returns to OLS estimator (Fayose and Ayinde, 2019; Kibria and Lukman, 2020; Fayose *et al.,* 2023a). In this paper, we considered these selected 'k' parameters: Hoerl and Kennard (1975), Fayose and Ayinde (2019), Kibria and Lukman (2020), Chand and Kibria (2024) and also proposed four new ridge parameters respectively.

2.1.3 Principal Component Estimator

The study considered PCA method to also handle multicollinearity in the model and also combined both ridge parameter and PCA method to form hybrid estimator to handle multicollinearity in the model.

PCA transforms the original predictors x into a new set of uncorrelated variables called principal components.

$$\boldsymbol{\beta_{PCR}} = \boldsymbol{V}(\boldsymbol{V'SV})^{-1}\boldsymbol{V'X'Y} \quad (6)$$

'S' is defined in equation (4).

Let the covariance matrix of *X* be:

$C = X'X$ and the eigen value decomposition of C gives $C = VDV'$ (7)

where V is $p \times p$ matrix of eigen vectors (Principal Components) and D = diag $(\lambda_1, \lambda_2, \ldots, \lambda_p)$ is the diagonal matrix of the eigen values.

The data matrix X is transformed into principal components $Z = XV$.

where Z is the new transformed data matrix of uncorrelated principal components

By regressing y on the principal components Z instead of X; is given as:

$$y = Z\gamma + e \quad (8)$$

where $\gamma = V'\beta$

The ordinary least square (OLS) estimator for γ in this model is given as;

$$\hat{\gamma} = (Z'Z)^{-1}Z'Y \quad (9)$$

By substituting $Z = XV$, therefore the PCA estimator of β denoted as $\hat{\beta}_{PCA}$ is defined as:

$$\hat{\beta}_{PCA} = V\hat{\gamma} = V(Z'Z)^{-1}Z'y \quad (10)$$

where $Z'Z = V'X'XV = D$

Therefore, the PCA estimator in equation (10) can further be defined as:

$$\hat{\beta}_{PCA} = V(D)^{-1}V'X'y \quad (11)$$

2.1.4 Some Alternative Ridge Estimators to OLSE

The ridge estimator is defined as:

$$\hat{\beta}_{RIDGE} = (Z'Z + kI)^{-1}Z'y) \quad (12)$$

where k is the non – negative tuning parameter. Different means of deriving k exists in the literature. These include:

Following (Hoerl and Kennard, 1975), k is given by:

$$\boldsymbol{KGRHK_{(median)}} = \boldsymbol{\hat{k}_i^M(HK)} = \boldsymbol{median}\left(\frac{\sigma^2}{\hat{\alpha}_i^2}\right), \text{ i = 1, 2, 3, p.} \quad (13)$$

where $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p}$is the mean square error from the OLS, α_i is the i[th] element of the vector, and is the regression coefficient from the OLS.

Following (Fayose and Ayinde, 2019), k is given by:

$$KGRFA = \hat{k}_i^{Min}(FA) = \frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}\left\{\left[\left(\frac{\hat{\alpha}_i^4\lambda_{Min}^2}{4\hat{\sigma}^2}\right) + \left(\frac{6\hat{\alpha}_i^4\lambda_{Min}}{\hat{\sigma}^2}\right)\right]^{\frac{1}{2}} - \left(\frac{\hat{\alpha}_i^2\lambda_{Min}}{2\hat{\sigma}^2}\right)\right\} \quad (14)$$

where $\lambda_{Min} = Min(\lambda_i) = 1,2,3,\ldots,p$.
Following (Kibria and Lukman, 2020), k is given by:

$$\boldsymbol{KGRKL = \hat{k}_i^{Min}(KL) = min\left(\frac{\hat{\sigma}^2}{2\hat{\alpha}_i^2 + (\hat{\sigma}/\lambda_i)}\right)} \quad (15)$$

Following (Chand and Kibria, 2024), k is given by:

$$KGRCK_1 = \hat{k}_i(CK_1) = \hat{\sigma}p^{(1+p/n)} \quad (16)$$

Following (Chand and Kibria, 2024), k is given by:

$$\boldsymbol{KGRCK_2 = \hat{k}_i(CK_2) = \hat{\sigma}\, max(p^{(1+p/n)}, p^{(1+1/p)})} \quad (17)$$

2.1.5 The Liu Estimator

The Liu estimator proposed by Liu (1993) combined the Stein estimator with Ordinary Ridge Regression estimator to handle multicollinearity. The Liu estimator of β is defined below as

$$\hat{\beta}_L = (X'X + I)^{-I}(X'X + dI)\hat{\beta}_{OLS} \qquad 0 < d < 1. \quad (18)$$

Where $d = min\left[\frac{\hat{\alpha}^2}{(\hat{\sigma}^2/\lambda_i) + \hat{\alpha}_i^2}\right]$ (19)

Where d = $(diag(d_i)$ and is a diagonal matrix of the biasing parameter. The Liu estimator can return to OLS when d = 1

2.2 Proposed Ridge Estimators

For the ridge parameter whose estimators are defined in (13), (14), (15), (16) and (17), the concept of different forms by Lukman and Ayinde (2017) and Fayose and Ayinde (2019) was introduced based on minimum (MI), maximum (MA) and Median (MD) of eigen values (λ_i) of X′X of the design matrix of the regression model.
Consequently, in this paper, we proposed some new ridge parameters whose estimators are defined below:
RIDGE ESTIMATOR (PROPOSED 1)

$$\hat{k}_i(CK2) = \hat{\sigma}\, min(p^{(1+p/n)}, p^{(1+1/p)}) \quad (20)$$

i.e. The minimum version of Chand and Kibria (2024)
RIDGE ESTIMATOR (PROPOSED 2)

$$\hat{k}_{KL(PROP)}^{HM} = p\sum_{i=1}^{p}\left[\frac{\sigma^2}{2\alpha_i^2 + \sigma/\lambda_i}\right] \quad (21)$$

i.e. the Harmonic mean version of Kibria and Lukman (2020)
RIDGE ESTIMATOR (PROPOSED 3)

$$\hat{k}_{KL(PROP)}^{Median} = median\left[\frac{\sigma^2}{2\alpha_i^2 + \sigma/\lambda_i}\right] \quad (22)$$

i.e. the median version of Kibria and Lukman (2020)
RIDGE ESTIMATOR (PROPOSED 4)

$$\hat{k}_{KL(PROP)}^{FM} = \left[\frac{\sigma^2}{2\,max(\alpha_i^2) + \sigma/max(\lambda_i)}\right] \quad (23)$$

Fixed maximum of Kibria and Lukman (2020).

2.2 Derivation of the Properties of PCA with Ridge Estimator

Ayinde *et al*. (2020) derived a new approach of Principal Component Analysis (PCA) estimator as an alternative to:

$$\hat{\beta}_{PCA} = VD^{-1}V'X'Y \quad (24)$$

This is defined as

$$\hat{\beta}_{PCA} = (X'X)^{-1}X'\hat{y}_r \quad (25)$$

Where $\hat{y}_r$ is the predicted variable by regressing y on the r-principal component defined as:

$\hat{y}_r = Z_r(Z_r'Z_r)^{-1}X'y$ (26)

Such that Z_r=XT_r

where T_r is the r – principal component and T is the orthogonal matrix. Therefore, combination of PCA with ridge estimator is defined as:

$$\hat{\beta}_{R-PCA} = (X'X + kI)^{-1}X'Z_r\left(Z_r{}'Z_r\right)^{-1}X'y \quad (27)$$

Where k is the biasing parameter for individual biasing parameter of Hoerl and Kennard (1975), Fayose and Ayinde (2019), Kibria and Lukman (2020), Chand and Kibria (2024a), Chand and Kibria (2024b).

2.2.1 Properties of $\widehat{\boldsymbol{\beta}}_{R-PCA}$

Mean of the $\widehat{\boldsymbol{\beta}}_{R-PCA}$

To compute the mean of $\hat{\beta}_{R-PCA}$, we take the expected value of equation (27)

$$\mathrm{E}(\hat{\beta}_{R-PCA}) = E\left[(X'X + kI)^{-1}X'Z_r\left(Z_r{}'Z_r\right)^{-1}X'y\right] \quad (28)$$

$$y = X\beta + e$$

$$=\mathrm{E}(\hat{\beta}_{R-PCA}) = E\left((X'X + kI)^{-1}X'Z_r\left(Z_r{}'Z_r\right)^{-1}X'(X\beta + e)\right).$$

$$=E\left((X'X + kI)^{-1}X'Z_r\left(Z_r{}'Z_r\right)^{-1}X'X\beta + (X'X + kI)^{-1}X'Z_r\left(Z_r{}'Z_r\right)^{-1}X'e\right)$$

$$= (X'X + kI)^{-1}X'Z_r\left(Z_r{}'Z_r\right)^{-1}X'XE(\beta) + (X'X + kI)^{-1}X'Z_r\left(Z_r{}'Z_r\right)^{-1}X'E(e)$$

$E(e)$ = 0 and $E(\beta) = \beta$

Therefore

$$(E(\hat{\beta}_{R-PCA}) = (X'X + kI)^{-1}X'Z_r\left(Z_r{}'Z_r\right)^{-1}X'X\beta \quad (29)$$

Variance of the $\widehat{\boldsymbol{\beta}}_{R-PCA}$

$$\mathrm{Var}\ (\hat{\beta}_{R-PCA}) = E\left[\left(\hat{\beta}_{R-PCA} - E(\hat{\beta}_{R-PCA})\right)\left(\hat{\beta}_{R-PCA} - E(\hat{\beta}_{R-PCA})\right)'\right] \quad (30)$$

$$=E\begin{bmatrix}\left[(X'X + kI)^{-1}X'Z_r\left(Z_r{}'Z_r\right)^{-1}X'y - (X'X + kI)^{-1}X'Z_r\left(Z_r{}'Z_r\right)^{-1}X'X\beta\right] \\ \left[(X'X + kI)^{-1}X'Z_r\left(Z_r{}'Z_r\right)^{-1}X'y - (X'X + kI)^{-1}X'Z_r\left(Z_r{}'Z_r\right)^{-1}X'X\beta\right]'.\end{bmatrix} \quad (31)$$

Further expansion of (31) gives equation (32)

$$Var(\hat{\beta}_{R-PCA}) = \sigma^2\ (X'X + kI)^{-2}(X'Z_r)^2\left(Z_r{}'Z_r\right)^{-2}X'X \quad (32)$$

Bias of the $\widehat{\boldsymbol{\beta}}_{R-PCA}$

Bias($\hat{\beta}_{R-PCA}$)=$E\left(\hat{\beta}_{R-PCA}\right) - \beta$

$$= (X'X + kI)^{-1}X'Z_r\left(Z_r{}'Z_r\right)^{-1}X'X\beta - \beta$$

$$= \left((X'X + kI)^{-1}X'Z_r\left(Z_r{}'Z_r\right)^{-1}X'X - I\right)\beta \quad (33)$$

Mean Squared Error Matrix of $\widehat{\boldsymbol{\beta}}_{R-PCA}$

MSEM ($\hat{\beta}_{R-PCA}$)= $Var\left(\hat{\beta}_{R-PCA}\right)$) + Bias^2($\hat{\beta}_{R-PCA}$) (34)

MSEM($\hat{\beta}_{R-PCA}$)= $\sigma^2(X'X + kI)^{-2}(X'Z_r)^2(Z_r'Z_r)^{-1}X'X + [(X'X + kI)^{-1}X'Z_r(Z_r'Z_r)^{-1}X'X - I]^2\beta^2$ (35)

Also, equation (35) can still be written as:

$$MSEM\left(\hat{\beta}_{R-PCA}\right) = \sigma^2(X'X + kI)^{-2}(X'Z_r)'(X'Z_r)(Z_r'Z_r)^{-1}X'X + [(X'X + kI)^{-1}X'Z_r(Z_r'Z_r)^{-1}X'X - I]'$$
$$[(X'X + kI)^{-1}X'Z_r(Z_r'Z_r)^{-1}X'X - I]\beta'\beta \quad (36)$$

Recall the general form of linear regression model in matrix form as given in (2)

Therefore, the canonical form of equation (2) can be written as:

$$y = WX + e \quad (37)$$

Where $W = XQ, \alpha = Q'\beta$ and Q is the orthogonal matrix whose columns indicate the eigen vectors of the design matrix $X'X$. Hence, $Q'X'XQ$ $=\wedge=$ $diag\left(\lambda_1, \lambda_2, \ldots, \lambda_p\right)$ where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$ are the ordered eigen values $X'X$. The OLS estimator in equation (2) can be defined as:

$$\hat{\alpha}_{ols} = \wedge^{-1}X'y \quad (38)$$

Thus, the canonical form of equation (26) is

$$\hat{\alpha}_{R-PCA} = (\wedge + kI)^{-1}X'Z_r\left(Z_r{}'Z_r\right)^{-1}X'y \quad (39)$$

For the convenience of establishing the statistical properties of $\hat{\alpha}_{R-PCA}$, the following Lemmas will be useful.
Lemma 1: Let F be a positive definite matrix such that F > 0 and let αbe some vectors then;
$F-\alpha\alpha' \geq 0$ if and only if $\alpha' F^{-1}\alpha \leq 1$ (Trenkler and Tontenbury,1990).
Lemma 11: Let $\hat{\alpha}_j = A_{ij}$ for j=1,2 be two competing estimators of α.. Also suppose that
$D = Cov(\hat{\alpha}_1) - Cov(\hat{\alpha}_2) > 0$ where $Cov(\hat{\alpha}_1)$ *and* $Cov(\hat{\alpha}_2)$ are the covariance matrix of $\hat{\alpha}_1 and \hat{\alpha}_2$
Therefore $D = MSEM(\hat{\alpha}_1) - MSEM(\hat{\alpha}_2) \geq 0$ if and only if $a_2'[\sigma^2 D + a_1 a_1^2]^{-1} a_2 < 1$ where $MSEM(\hat{\alpha}_1)$= $Cov(\hat{\alpha}_1)$+$a_i a_i'$
Such that a_i=Bias($\hat{\alpha}_i$) $= (AiX - I)\alpha$

2.2.2 The Superiority of the Proposed Estimator in the Sense of MSEM Criterion

The proposed estimator is compared with some already existing estimators such as OLS, Ridge estimators in the sense of MSEM.

Comparison between $\hat{\alpha}_{ols}$ and $\hat{\alpha}_{R-PCA}$

Recall the MSEM of the OLS estimator as
$\hat{\alpha}_{ols}$=$\Lambda^{-1}X'y$as:

MSEM ($\hat{\alpha}_{ols}$)= $\sigma^2\Lambda^{-1}$ (40)

Equation (35) can be written as:

MSEM ($\hat{\alpha}_{R-PCA}$)=$\sigma^2 N^{-2} N_r' N_r \Lambda_r^{-1}\Lambda + [N^{-1}N_r\Lambda_r^{-1}\Lambda - I][N^{-1}N_r\Lambda_r^{-1}\Lambda - I]'\alpha^1\alpha$ (41)

where N=$(\Lambda + kI)$,N_r'= $X'Z_r$, $\Lambda_r = Z_r' Z_r$, $\Lambda = X'X$
Therefore, the difference between (40) and (41) is given as
MSEM ($\hat{\alpha}_{ols}$)-MSEM ($\hat{\alpha}_{R-PCA}$) =$D_{\alpha_{ols}}^{\alpha_{R-PCA}}$
= $\sigma^2\Lambda^{-1}$- $\sigma^2 N^{-2} N_r' N_r \Lambda_r^{-1}\Lambda + [N^{-1}N_r\Lambda_r^{-1}\Lambda - I]'[N^{-1}N_r\Lambda_r^{-1}\Lambda - I]\alpha'\alpha$
= $\sigma^2\Lambda^{-1}$- $\sigma^2 N^{-2} N_r' N_r \Lambda_r^{-1}\Lambda + [N^{-1}N_r\Lambda_r^{-1}\Lambda - I]'\alpha'\alpha[N^{-1}N_r\Lambda_r^{-1}\Lambda - I]$

= $\sigma^2[\Lambda^{-1} - N^{-2}N_r' N_r \Lambda_r^{-1}\Lambda] + [N^{-1}N_r\Lambda_r^{-1}\Lambda - I]'\alpha'\alpha[N^{-1}N_r\Lambda_r^{-1}\Lambda - I]$ (42)

Where *k>0* is the individual biasing parameter of Hoerl and Kennard (1970), Fayose and Ayinde (2019), Kibria and Lukman (2020) and Chand and Kibria (2024).

The proposed estimator $\hat{\alpha}_{R-PCA}$ is superior to$\hat{\alpha}_{ols}$ if and only if $\alpha'\left[N^{-1}N_r{\Lambda_r}^{-1}\Lambda - I\right]'\sigma^2\left[\Lambda^{-1} - N^{-2}N_r' N_r {\Lambda_r}^{-1}\Lambda\right]^{-1}\left[N^{-1}N_r{\Lambda_r}^{-1}\Lambda - I\right]\alpha < 1$

Proof: By considering the dispersion matrix difference

$D_{\alpha_{ols}}^{\alpha_{R-PCA}}$= $\sigma^2\left[\Lambda^{-1} - N^{-2}N_r' N_r {\Lambda_r}^{-1}\Lambda\right]$ (43)

=trace ($D_{\alpha_{ols}}^{\alpha_{R-PCA}}$)

=$\sum_{i=1}^{p} diag$ ($D_{\alpha OLS}^{\alpha R-PCA}$)

$$=\sigma^2 \sum_{i=1}^{p} diag\left[\Lambda^{-1} - N^{-2}N_r' N_r {\Lambda_r}^{-1}\Lambda\right]$$

$$=\sigma^2 \sum_{i=1}^{p} diag\left[\frac{1}{\lambda_i} - \frac{n_r^2\lambda_i}{(\lambda_i+k)^2\lambda_{ir}}\right]_{i=1}^{p} \quad (44)$$

Where λ_i is the diag (X'X), n_r=diag($X'Z_r$)*and* λ_{ir}=diag($Z'Z_r$)
The difference will be positive definite if and only if $(\lambda_{i+}k)^2\lambda_{ir}$- $n_{ir}^2{\lambda_i}^2$> 0. It can be observed that $(\lambda_i + k)^2\lambda_{ir}$- $n_{ir}^2{\lambda_i}^2$ will be greater than zero if *k* >0. Hence, by Lemma II the proof is completed

Comparison between $\hat{\alpha}_{RE}$ and $\hat{\alpha}_{R-PCA}$

The bias vector covariance matrix and MSEM of $\hat{\alpha}_{RE}$estimator defined as

$\hat{\alpha}_{RE}$= $(\Lambda + kI)^{-1}$W'y (45)

$E(\hat{\alpha}_{RE})$ = $(\Lambda + kI)^{-1}\Lambda\alpha$ (46)

Var($\hat{\alpha}_{RE}$) = σ^2 $(\Lambda + kI)^{-1}\Lambda\alpha$ (47)

MSEM($\hat{\alpha}_{RE}$) = $\sigma^2(\Lambda + kI)^{-1}\Lambda(+kI)^{-1} + k^2(\Lambda + kI)^{-1}\alpha\alpha'(\Lambda + kI)^{-1}$ (48)

MSEM($\hat{\alpha}_{RE}$) = $\sigma^2 N^{-1}\Lambda N^{-1} + k^2 N^{-1}\alpha\alpha^1 N^{-1}$ (49)

Therefore, the difference between (41) and (49)

MSEM ($\hat{\alpha}_{RE}$) – MSEM ($\hat{\alpha}_{R-PCA}$)

$=\sigma^2 N^{-1}\Lambda N^{-1} - \sigma^2 N^{-2} N_r' N_r \Lambda_r^{-1}\Lambda + k^2 N^{-1}\alpha\alpha' N^{-1} - [N^{-1}N_r\Lambda_r^{-1}\Lambda - I]^1[N^{-1}N_r\Lambda_r^{-1}\Lambda - I]\alpha'\alpha$
(50)

Let k>0, the estimator $\hat{\alpha}_{R-PCA}$is superior to $\hat{\alpha}_{RE}$if and only if{ MSEM ($\hat{\alpha}_{RE}$) – MSEM ($\hat{\alpha}_{R-PCA}$)}>0,if and only if $=\sigma^2(N^{-2}\Lambda - N^{-2}N_r' N_r\Lambda_r^{-1}\Lambda) + k^2 N^{-1}\alpha\alpha' N^{-1} - [N^{-1}N_r\Lambda_r^{-1}\Lambda - I]^1[N^{-1}N_r\Lambda_r^{-1}\Lambda - I]\alpha'\alpha < 1$

Proof

Considering the dispersion matrix difference between $\hat{\alpha}_{RE}$and $\hat{\alpha}_{R-PCA}$

$D_{RE}^{\alpha_{R-PCA}}$=$\sigma^2(N^{-2}\Lambda - N^{-2}N_r' N_r\Lambda_r^{-1}\Lambda)$ (51)

$=\sigma^2(\Lambda + kI)^{-2}\Lambda - \sigma^2(\Lambda + kI)^{-2}(X'Z_r)^2\left(Z_r{}'Z_r\right)^{-1}\Lambda$

$=\sigma^2(\Lambda + kI)^{-2}\left[\Lambda - (X'Z_r)^2\left(Z_r{}'Z_r\right)^{-1}\Lambda\right]$

$=\sigma^2 \text{diag} \quad \left[\frac{\lambda_i}{(\lambda_i+k)^2} - \frac{n_r^2\lambda_i}{(\lambda_i+k)^2\lambda_{ir}}\right]_{i=1}^{p}$ (52)

$D_{RE}^{\alpha_{R-PCA}}$will be pdf if and only if $\lambda_i\lambda_{ir} - n_{ir}^2 > 0$fork>0.Hence it can be observed that $\lambda_i\lambda_{ir} - n_{ir}^2 > 0$,therefore by Lemma II the proof is completed.

2.2.3 Determination of Biasing Parameter ‘k’.

Finding appropriate ridge shrinkage or biasing parameter has been the bone of contention in the study of ridge regression. This is because the parameter may either be non – stochastic or may depend on the observed or real – life data set. Therefore, the shrinkage parameter k employed in this study are stated in equation (13), (14), (15), (16) and (17) as well as the proposed ones (20) to (23) respectively.

For practical purpose the MSE of $\hat{\alpha}_{R-PCA}$ can be written as

MSE($\hat{\alpha}_{R-PCA}$)= trace {MSEM ($\hat{\alpha}_{R-PCA}$)} (53)

= $\sum_{I=1}^{P} diag$ (MSEM ($\hat{\alpha}_{R-PCA}$)) (54)

$$=\sum_{i=1}^{P} diag\,[\sigma^2 N^{-2}N_1 N_r\Lambda_r^{-1}\Lambda + [N^{-2}N_r^2\Lambda_r^{-2}\Lambda^2 - 2N^{-1}N_r\Lambda_r^{-1}\Lambda + I]\hat{\alpha}_i^2]$$

MSE ($\hat{\alpha}_{R-PCA}$) = $\sigma^2\sum_{i=1}^{p}\left[\frac{n_{ir}{}^2\lambda_i}{(\lambda_i+k)^2\lambda_{ir}}\right] + \sum_{i=1}^{p}\left[\frac{n_{ir}{}^2\lambda^2{}_i}{(\lambda_i+k)^2\lambda^2{}_{ir}} - \frac{n_{ir}\lambda_i}{(\lambda_i+k)\lambda_{ir}} + 1\right]\hat{\alpha}_i^2$

$$\text{MSE}(\hat{\alpha}_{R-PCA})=\sigma^2\sum_{i=1}^{p}\left[\frac{{n_{ir}}^2\lambda_i}{(\lambda_i+k)^2\lambda_{ir}}\right]+\sum_{i=1}^{p}\left[\frac{{n_{ir}}^2\lambda^2{}_i-(\lambda_i+k)\lambda_{ir}(n_{ir}\lambda_i+(\lambda_i+k)\lambda_{ir}}{(\lambda_i+k)^2\lambda^2{}_{ir}}\right]\hat{\alpha}_i^2 \quad (55)$$

Where n_{ir} is the eigen value of $\boldsymbol{N_r = XZ_r}$, $\boldsymbol{\lambda_{ir}}$ is the eigen value of matrix $\boldsymbol{Z_r' Z_r}$, $\boldsymbol{\lambda_r}$ is the eigen value of the design matrix X^1X, p is the number of the explanatory variables and k is the generalized biasing parameter which is the individual biasing parameter which were earlier defined in (13) to (17), (19) to (23) respectively.

Simulation Procedure and Design

A Monte Carlo simulation study is performed in the study to show the performance of the proposed estimator over some existing estimators in literature.

Consider the linear regression of the form:

$$y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \ldots + \beta_p X_{tp} + U_t \quad (56)$$

where $t = 1, 2, \ldots, n$; $p = 3, 6$, $U_t \approx N(0, \sigma^2)$, X_{ti}, t=1,2,…,n; i= 1,2,...,p are fixed explanatory variables. The explanatory variables are generated using the following procedure (Fayose and Ayinde, 2023b):

$$X_{ti} = (1-\rho^2)^{\frac{1}{2}} Z_{ti} + \rho Z_{tp} \quad (57)$$

where Z_{ti} is independent standard normal distribution with mean zero and constant variance, ρ is the correlation between any two explanatory variables and p is the number of explanatory variables. The error terms U_t were generated to be normally distributed with mean zero and variance σ^2. U_t is the error term. The study used Monte Carlo simulation to conduct the experiment with varying parameters such as sample sizes (n = 10, 20, 30, 50, 100 and 250); level of multicollinearity (ρ = 0, 0.8, 0.9, 0.95, 0.99, 0.999). In the study, σ^2 values were 1, 25 and 100. $E(y_i)$ = expected value of the regression under consideration. $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u_i$ for p = 3 and $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + u_i$ for p = 6. When p = 3; β_0 = 0.1550494, β_1 = 0.6162552, β_2 = 0.5311015, β_3 = 0.5604644. When p = 6; β_0 = 0.09867566, β_1 = 0.47436618, β_2 = 0.35373615, β_3 = 0.41280825, β_4 = 0.32653190,
β_5 = 0.40968387, β_6 = 0.44185515. The experiment was repeated 1000 times (number of replication). The performances of the estimators were compared using the Mean Square Error criterion. For any estimator $\hat{\beta}$, MSE is defined as follows:

$$MSE(\hat{\beta}) = \frac{1}{1000}\sum_{i=1}^{p}\sum_{j=1}^{1000}\left(\hat{\beta}_{ij} - \beta_i\right)^2 \quad (58)$$

where $\hat{\beta}_{ij}$ is i^{th} element of the estimator β in the j^{th} replication which gives the estimate of β_i. β_i are the true value of the parameter previously mentioned. Estimator with the minimum MSE was considered best. The statistical package R Studio was used to write the program that accommodated Twenty – One (21) estimators (OLS, PCA estimator, Generalized Ridge estimators, Liu estimator, Ridge –PCA estimators). Out of the Twenty – One (21) estimators, thirteen (13) are proposed estimators, eight (8) are existing estimators. At a particular level of error variance, multicollinearity and sample size, R studio package gave MSE values. These were recorded 180 times (Multicollinearity levels x Error Variance x Sample Sizes x Number of Regressors types = 5 x 3 x 6 x 2) accordingly. Statistical Package for the Social Sciences (SPSS 25.0) was further used to rank the estimators on the basis of their MSE values. Estimators with high MSE were sorted and removed using SPSS software. The MSE obtained by each estimator was ranked for each degree of multicollinearity and error variance. The degrees of multicollinearity and error variance were tallied to determine the number of times each estimator had the smallest MSE (rank 1 and 5). An estimator is optimal or most efficient if it has the most counts; the mode.

3. Results

We represented the outcomes visually and in Tables.

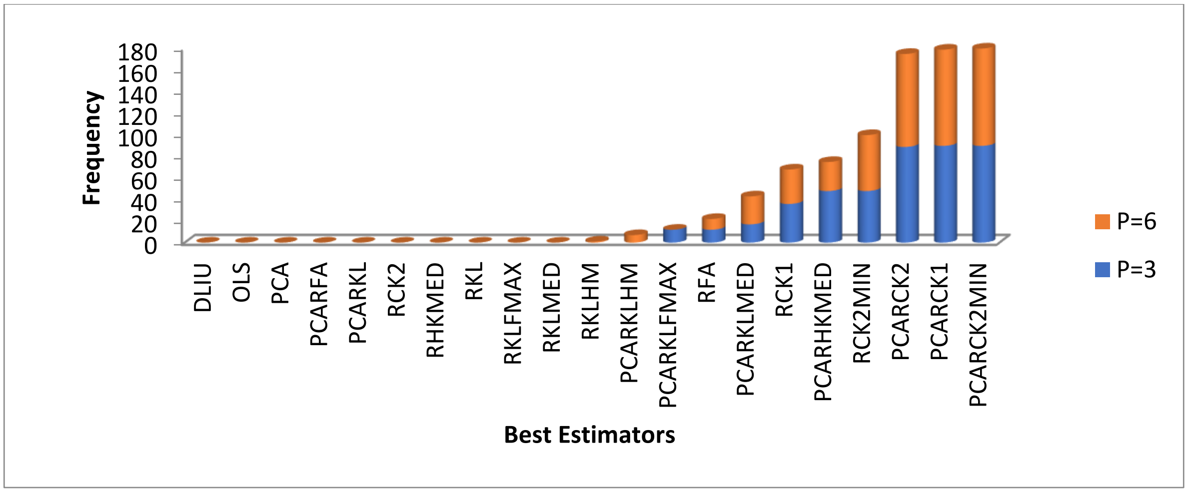

Figure 1: ***Component Bar Chart showing frequency of counts at which MSE is minimum at p = 3 and p = 6 for OLS and PCA estimation methods***

Figure 1 shows that PCARCK2MIN is the most efficient estimator when dealing with multicollinearity in the model. That is, proposed one–parameter ridge estimator (minimum version of Chand and Kibria 2 (2024) with Principal Component estimator) followed by combined estimator nicknamed (PCARCK1) that is. proposed one – parameter ridge estimator with PCA (Chand and Kibria 1 (2024) with Principal Component estimator) and combined estimator denoted as (PCARCK2) that is, proposed one–parameter ridge estimator with PCA (Chand and Kibria 2 (2024) with Principal Component estimator) respectively. In contrast, the most efficient estimator without Principal Component estimator is the proposed one – parameter ridge estimator that is, minimum version of Chand and Kibria 2 (2024) followed by existing one–parameter ridge estimator of Chand and Kibria 1 (2024) and existing one–parameter ridge estimator of Fayose and Ayinde (2019).

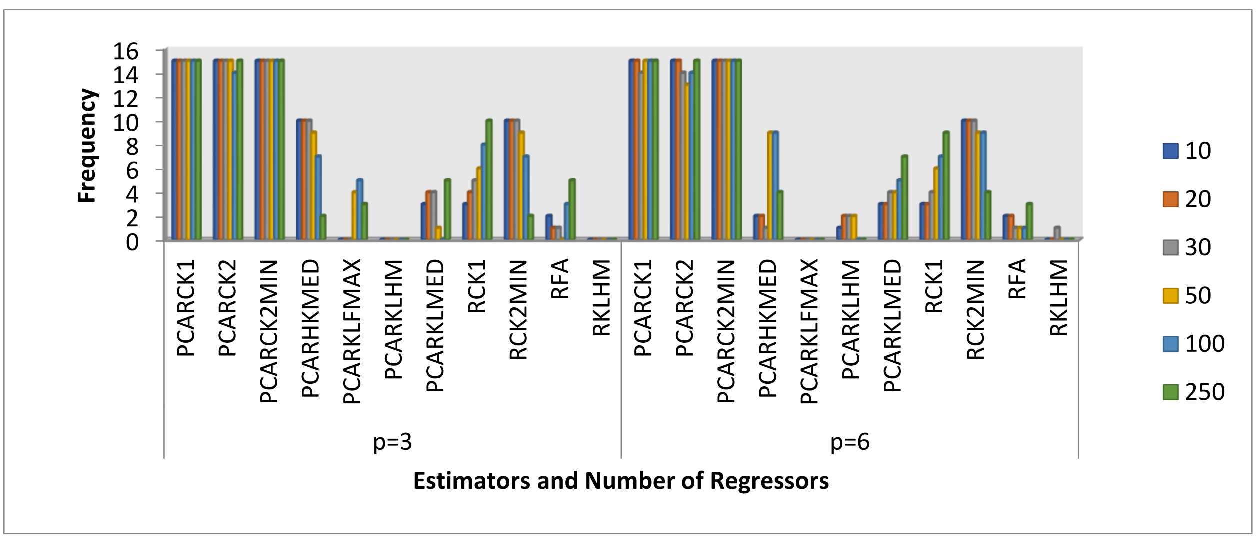

Figure 2: *Multiple Bar chart showing performance of the estimators at different sample sizes when p = 3 and p = 6.*

From Figure 2, proposed (PCARCK2MIN) i.e. minimum version of Chand and Kibria 2 (2024) with Principal Component estimator performed efficiently across all sample sizes at p = 3 and at p = 6 respectively. Similarly, proposed (PCARCK1) i.e. Chand and Kibria 1 (2024) with Principal Component estimator performed efficiently across all sample sizes at p = 3 but when p = 6, it performed across all sample sizes expect at sample size 30.

The simulation results are available on request but for ease of comparison the results are summarized in Table 1.

TABLE 1: Number of Times Each Estimator Produced Minimum MSE when counted over levels of Multicollinearity and Error Variance

P	Estimators	Sample Size (n)							
		10	20	30	50	100	250	TOTAL	RANK
3	**PCARCK1**	15	15	15	15	15	15	**90**	**1st**
	PCARCK2	15	15	15	15	**14**	15	**89**	**2nd**
	PCARCK2MIN	15	15	15	15	15	15	**90**	**1st**
	PCARHKMED	**10**	**10**	**10**	**9**	**7**	**2**	**48**	**3rd**
	PCARKLFMAX	**0**	**0**	**0**	**4**	**5**	**3**	**12**	**6th**
	PCARKLHM	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**7th**
	PCARKLMED	**3**	**4**	**4**	**1**	**0**	**5**	**17**	**5th**
	RCK1	**3**	**4**	**5**	**6**	**8**	**10**	**36**	**4th**
	RCK2MIN	**10**	**10**	**10**	**9**	**7**	**2**	**48**	**3rd**
	RFA	**2**	**1**	**1**	**0**	**3**	**5**	**12**	**6th**
	RKLHM	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**7th**
6	**PCARCK1**	15	15	**14**	15	15	15	**89**	**2nd**
	PCARCK2	15	15	**14**	**13**	**14**	15	**86**	**3rd**
	PCARCK2MIN	15	15	15	15	15	15	**90**	**1st**
	PCARHKMED	**2**	**2**	**1**	**9**	**9**	**4**	**27**	**6th**
	PCARKLFMAX	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**11th**
	PCARKLHM	**1**	**2**	**2**	**2**	**0**	**0**	**7**	**9th**
	PCARKLMED	**3**	**3**	**4**	**4**	**5**	**7**	**26**	**7th**
	RCK1	**3**	**3**	**4**	**6**	**7**	**9**	**32**	**5th**
	RCK2MIN	**10**	**10**	**10**	**9**	**9**	**4**	**52**	**4th**
	RFA	**2**	**2**	**1**	**1**	**1**	**3**	**10**	**8th**
	RKLHM	**0**	**0**	**1**	**0**	**0**	**0**	**1**	**10th**

NOTE: **Estimator with highest frequency at each sample size is bolded at p = 3 and p = 6.**

From Table 1, when p = 3, it is observed that the best or most efficient estimator are PCARCK1 i.e. proposed one –parameter ridge estimator with PCA (Chand and Kibria 1 (2024) with Principal Component estimator) and PCARCK2MIN i.e. proposed one–parameter ridge estimator with PCA (minimum version of Chand and Kibria 2

(2024) with Principal Component estimator) followed by PCARCK2 i.e. proposed one – parameter ridge estimator with PCA (Chand and Kibria 2 (2024) with Principal Component estimator) respectively. But when p = 6, the most efficient estimator is PCARCK2MIN i.e. proposed one–parameter ridge estimator with PCA (minimum version of Chand and Kibria 2 (2024) with Principal Component estimator) followed by PCARCK1 i.e. proposed one – parameter ridge estimator with PCA (Chand and Kibria 1 (2024) with Principal Component estimator) respectively. The top best or most efficient estimators are the proposed ones.

3.1 Application to real – life datasets

3.1.1 Description of the Data used

Two different datasets were used to prove the efficacy of the proposed estimators. They are Portland cement data and Longley data respectively.

3.12 Portland Cement Data

We use the Portland cement data which was originally adopted by Woods *et al.,* (1932) to explain their theoretical results. The data were analyzed by various researchers: to mention a few, Li and Yang (2012), Lukman *et al.*, (2019), Kibria and Lukman (2020) and recently Dawoud and Kibria (2020) among others. The regression model for these data is defined as: $y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon_i$ (58)

For more details about these data, see Woods *et al.,* (1932). According to Dawoud and Kibria (2020), The variance inflation factors are $VIF_1 = 38.50$, $VIF_2 = 254.42$, $VIF_3 = 46.87$ and $VIF_4 = 282.51$. Eigenvalues of S are $\lambda_1 = 44676.206$, $\lambda_2 = 5965.422$, $\lambda_3 = 809.952$, and $\lambda_4 = 105.419$ and the condition number of S is approximately 20.58. The VIFs, the eigenvalues, and the condition number all indicate that severe multicollinearity exists. The estimated parameters and the MSE values of the estimators are presented in Table 2.

TABLE 2: Results of Regression Coefficients and the Corresponding MSE Values

Coef.	**Estimators**								
	$\hat{\alpha}$	$\hat{\alpha}(k)$	$\hat{\alpha}_{CK1}$	$\hat{\alpha}_{CK2}$	$\hat{\alpha}_{FA}$	$\hat{\alpha}_{KL}$	$\hat{\alpha}_{proposed}$	$\hat{\alpha}_{proposed}$	$\hat{\alpha}_{proposed}$
		$\hat{k}$	$\hat{k}$	$\hat{k}_{max}$	$\hat{k}_{min}$	$\hat{k}_{min}$	$\hat{k}_{CK2(\text{min})}$	$\hat{k}_{KL(\text{median})}$	$\hat{k}_{KL(\text{HM})}$
β_o	**62.405**	**8.5871**	**0.0441**	**0.0441**	**0.4095**	**38.2886**	**0.0446**	**0.1081**	**0.0317**
β_1	**1.5511**	**2.1046**	**1.8099**	**1.8099**	**2.1860**	**1.7992**	**1.8326**	**2.1749**	**0.9399**
β_2	**0.5102**	**1.0649**	**1.2394**	**1.2394**	**1.1498**	**0.7587**	**1.2344**	**1.1562**	**1.3976**
β_3	**0.1019**	**0.6681**	**0.4935**	**0.4935**	**0.7521**	**0.3557**	**0.5084**	**0.7451**	**0.0872**
β_4	**-0.1441**	**0.3996**	**0.5420**	**0.5420**	**0.4826**	**0.0996**	**0.5388**	**0.4878**	**0.6393**
K	**-**	**0.0077**	**36.6627**	**36.6627**	**0.2085**	**0.0008**	**33.8447**	**1.2431**	**359.0090**
MSE	**4912.090**	**67.0785**	0.0612	0.0612	**0.2512**	**1847.381**	**0.0627**	**0.0894**	0.4637

****Best Estimators: RCK1 and RCK2MAX***

From Table 2, the top five most efficient estimators based on MSE values are RCK1, RCK2MAX, RCK2MIN, RKLMED and RFA respectively. The performance of one of the proposed efficient estimators in Portland cement data agreed with the simulation results. It appears from Table 2 that the existing RCK estimators perform best among the mentioned estimators as it gives the smallest MSE value.

3.1.3 Longley Data

Longley data were originally used by Longley (1967) and then by other authors (Lukman and Ayinde, 2017 and Badawaire *et al.,* 2023).

The regression model of this data is defined as:

$$y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \varepsilon_i \quad (59)$$

where y is the total derived employment, x_1 is the gross national product implicit price deflator, x_2 is the gross national product, x_3 is unemployment, x_4 is the size of armed forces, x_5 is the non-institutional population 14 years of age and over and x_6 is the time. For more details about these data, see Longley (1967). The variance inflation factors are VIF_1 = 135.53, VIF_2 = 1788.51, VIF_3 = 33.62, VIF_4 = 3.59, VIF_5 = 399.15 and VIF_6 = 758.98. Eigenvalues of S are as follows: 2.76779×10^{12}, 7,039, 139,179, 11,608, 993.96, 2,504, 761.021, 1738.356, 13.309 and the condition number of S is approximately 456,070. The VIFs, the eigenvalues and the condition number all indicate that severe multicollinearity exists (Dawoud and Kibria, 2020). The estimated parameters and the MSE values of the estimators are presented in Table 3.

TABLE 3: Results of Regression Coefficients and the Corresponding MSE Values

Coef.								
	$\hat{\alpha}$	$\hat{\alpha}(k)$	$\hat{\alpha}_{Proposed}$	$\hat{\alpha}_{Proposed}$	$\hat{\alpha}_{Proposed}$	$\hat{\alpha}_{proposed}$	$\hat{\alpha}_{proposed}$	$\hat{\alpha}_{PCA}$
		$\hat{k}$	$\hat{k}_{PCARFA}$	$\hat{k}_{PCARHK(\text{med})}$	$\hat{k}_{PCARKL}$	$\hat{k}_{PCARKL(\text{fmax})}$	$\hat{k}_{PCARCK2(\text{min})}$	
β_0	**-3482.26**	**-2388.90**	**-1.5962**	**-2388.898**	**-3372.02**	**-3372.02**	**-0.0006**	**65.317**
β_1	**0.0151**	**-0.0063**	**-0.0529**	**-0.0063**	**0.0129**	**0.0129**	**-0.0236**	**1.4495**
β_2	**-0.0358**	**-0.0023**	**0.0710**	**-0.0023**	**-0.0324**	**-0.0324**	**0.0613**	**0.3252**
β_3	**-0.0202**	**-0.0152**	**-0.0042**	**-0.0152**	**-0.0197**	**-0.0197**	**-0.0054**	
β_4	**-0.0103**	**-0.0089**	**-0.0057**	**-0.0088**	**-0.0102**	**-0.0102**	**-0.0057**	
β_5	**-0.0511**	**-0.1651**	**-0.4139**	**-0.1651**	**-0.0626**	**-0.0626**	**-0.3047**	
β_6	**1.8292**	**1.2700**	**0.0492**	**1.2700**	**1.7728**	**1.7728**	**0.0425**	
K	**-**	**5.36488E-08**	**0.0003**	**5.36488E-08**	**3.83206E-09**	**3.83206E-09**	**0.7517**	
MSE	792848.81	**341784.46**	**0.135316**	**0.135319841**	**0.135319842**	**0.135319842**	0.773322622	**0.822323712**

Best Estimator: PCARFA

From Table 3, the top five most efficient estimators based on MSE values are PCARFA, PCARHKMED, PCARKL, PCARKLFMAX and PCARCK2MIN respectively.

The performance of two of the proposed efficient estimators in Longley data agreed with the simulation results. It appears from Table 3 that the proposed PCARFA estimator performs best among the mentioned estimators as it gives the smallest MSE value.

4. Discussions

This study proposed innovative hybrid estimators that effectively combine Ridge Regression with Principal Component Analysis to combat multicollinearity in Gaussian linear regression models. The integration of newly developed ridge parameters with PCA produced estimators that demonstrated remarkable improvements in estimation accuracy and model stability. Simulation results revealed that the proposed estimators, especially PCARCK2MIN and PCARCK1, consistently yielded lower Mean Squared Errors compared to conventional methods such as OLS and standard ridge regression. The estimators-maintained robustness across various sample sizes, multicollinearity levels, and error variances, confirming their adaptability in practical statistical modeling scenarios. Furthermore, application to real – life datasets (Portland cement and Longley) aligned with simulation findings, confirming the empirical superiority of the proposed methods. These results reinforce the need for continued development of hybrid estimation techniques in the presence of multicollinearity. These proposed hybrid estimators can be extended to generalized linear models.

5. Conclusion

The study developed and evaluated new hybrid ridge–PCA estimators designed to mitigate the adverse effects of multicollinearity in Gaussian linear regression models. Findings from the Monte Carlo simulation demonstrated clear superiority of the proposed estimators over conventional methods. In particular, **PCARCK2MIN** consistently produced the smallest Mean Squared Error (MSE) across varying sample sizes, multicollinearity levels and error variances, followed closely by **PCARCK1**. This confirms that integrating optimized ridge parameters with principal component transformation substantially improves estimation efficiency and stability.

Comparison with the two real – life datasets (Portland cement and Longley) further validates the simulation outcomes. For the Portland cement data, the existing ridge–based estimators (notably **RCK1** and **RCK2MAX**) achieved the smallest MSE, although one of the proposed estimators still ranked among the most efficient, partially supporting the simulation evidence. In contrast, results from the Longley dataset strongly aligned with the simulation findings where the proposed **PCARFA** estimator outperformed all competitors by producing the minimum MSE.

Overall, both empirical applications corroborate the robustness of the proposed hybrid estimators, particularly under severe multicollinearity. While simulation results favored PCARCK2MIN as the most stable and efficient estimator, evidence from real – life data confirmed that the proposed estimators generally dominate existing estimators and hereby recommended.

Acknowledgments: The authors wish to express their gratitude to TETFUND for their support in this research.

Author Contributions: R.E.: Conceptualization, Writing, Software, Review. O.O.: Conceptualization, Writing, Supervision and Review. O.O.: Supervision, Editing, Writing, Supervision and Review.

Declaration of Competing Interest: The authors declared that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this journal. Finally, the authors agreed that all information supplied here are real and original.

REFERENCES

Alabi, R. E., Alabi, O. O. and Ojo, O. O. (2025). Development of hybrid ridge–PCA estimators for addressing multicollinearity in Gaussian linear regression models. *World Journal of Advanced Research and Reviews, 27(1), 942–957.* https://doi.org/10.30574/wjarr.2025.27.1.2559

Aladesuyi, A., Ayinde, K. and Fayose, T. S. (2025). Assessing the role of significant roots in parameter estimation of linear regression models under multicollinearity. *Tech–Sphere Journal of Pure and Applied Sciences, 2(1), 1–16.* https://doi.org/10.5281/zenodo.15470100

Ayinde, K., Lukman, A. F., Alabi, O. O. and Bello, H. A. (2020). A new approach of principal component regression estimator with applications to collinear data. *International Journal of Engineering Research and Technology, 13(7), 1616–1622.*

Badawaire, A. B., Dawoud, I., Lukman, A. F., Laoye, V., and Arowolo, O. (2023). Biasing estimator to mitigate multicollinearity in linear regression model. *Al-Bahir Journal for Engineering and Pure Sciences, 2(1), 1–9. https://doi.org/10.55810/2313-0083.1011*

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). Regression diagnostics: Identifying influential data and sources of collinearity. John Wiley and Sons.

Bühlmann, P. and van de Geer, S. (2011). Statistics for high-dimensional data: Methods, theory and applications. Springer.

Buonaccorsi, J. P. (1996). A modified estimating equation approach to correcting for measurement error in regression. Biometrika, 83, 433–440.

Chand, S. and Kibria, B. M. G. (2024). A new ridge type estimator and its performance for the linear regression model: Simulation and application. *Hacettepe Journal of Mathematics and Statistics, 53(3), 837–850. https://doi.org/10.15672/hujms.1359446*

Chang, Y. and Yang, Y. (2012). Analysis of multicollinearity in regression models with an application to a real data example. *Journal of Statistical Theory and Practice, 6(3), 455–470*.

Chatterjee, S. and Hadi, A. S. (1977). Regression analysis by example. Wiley.

Chatterjee, S., Hadi, A. S., and Price, B. (2000). Regression analysis by example (3rd ed.). Wiley.

Dawoud, I. and Kibria, B. M. G. (2020). A new biased estimator to combat multicollinearity of the Gaussian linear regression model. Stats, 3, 526–541. https://doi.org/10.3390/stats3030032

Dawoud, I., Lukman, A. F., and Haadi, A. (2022). A new biased regression estimator: Theory, simulation and application. Scientific African, 15, e01100. https://doi.org/10.1016/j.sciaf.2022.e01100

Fayose, T. S., and Ayinde, K. (2019). Different forms of biasing parameter for generalized ridge regression estimator. *International Journal of Computer Applications, 181, 21–29.*

Fayose, T. S., Ayinde, K. and Alabi, O. O. (2023a). M-robust weighted ridge estimator in linear regression model. African Scientific Reports, 2(123), 1–28.

Fayose, T. S., Ayinde, K., Alabi, O. O. and Bello, A. H. (2023b). Robust weighted ridge regression based on S-estimator. African Scientific Reports, 2(126), 1–28.

Gujarati, D. N., Porter, D. C. and Gunasekar, S. (2012). Basic econometrics (5th ed.). Tata McGraw-Hill.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for non-orthogonal problems. Technometrics, 12(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

Hoerl, A. E. and Kennard, R. W. (1975). Ridge regression: Iterative estimation of the biasing parameter. Communications in Statistics – Theory and Methods, 4(1), 77–88. https://doi.org/10.1080/03610927508827232

Huang, J. and Wang, H. (2018). A regularized estimator for high-dimensional generalized linear models with PCA. Journal of Statistical Computation and Simulation, 88(4), 675–686. https://doi.org/10.1080/00949655.2017.1417839

Jolliffe, I. T. (1986). Principal component analysis. Springer.

Jolliffe, I. T. (2002). Principal component analysis (2nd ed.). Springer.

Kibria, B. M. G. and Lukman, A. F. (2020). A new ridge-type estimator for the linear regression model: Simulations and applications. Scientifica. https://doi.org/10.1155/2020/9758378

Li, Y. and Yang, H. (2012). A new Liu-type estimator in linear regression model. Statistical Papers, 53, 427–437. https://doi.org/10.1007/s00362-010-0336-4

Liu, K. (1993). A new class of biased estimate in linear regression. Communications in Statistics – Theory and Methods, 22, 393–402. https://doi.org/10.1080/03610929308831027

Longley, J. W. (1967). An appraisal of least squares programs for the electronic computer. *Journal of the American Statistical Association, 62, 819–841. https://doi.org/10.1080/01621459.1967.10482997*

Trenkler,G.(1984). *On the performance of Biased Estimators in the Linear Regression Model with Correlated or Heteroscedastic Errors. Journal of Econometrics, 25, 179 -190*

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B, 67(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503*.