Received: 8-10-2025
Revised: 26-11-2025
Accepted: 7-12-2025
Published: 30-12-2025

Elicit Journal of Science Research
ISSN: 3093-0502

https://elicitpublisher.com

EMPIRICAL PAPER

The Effect of Multivariate Regression Over Multiple Linear Regression Models Using Non-Normal Data

Isiak, K.O[1*], Musa Olatunji Yunusa[2], Ishola Kabir Babatude[3], Sanda Akeem Kunle[4], Abubakar Adam Oluwakayode[4]

Affiliations

[1]Department of Mathematics, Nigerian Army University BIU, Nigeria
[2]Department of Statistics, Kwara State Polytechnic Ilorin, Nigeria
[3]Department of Computer Science, Nigerian Army University BIU, Nigeria
[4]Department of Statistics, University of Ilorin, Ilorin Nigeria

Orcid Identifiers

Isiak, K.O
0009-0000-1767-2709

Musa Olatunji Yunusa
0009-0000-0562-0776

Ishola Kabir Babatude
0009-0000-8410-9772

Sanda Akeem Kunle

Abubakar Adam Oluwakayode
0009-0006-6458-7631

Abstract

Purpose: This study evaluated the performance of multivariate linear regression relative to multiple linear regression when applied to non-normal data, with the objective of identifying the model that offers greater accuracy and reliability under conditions of equal mean vectors and known variance–covariance matrices.

Methodology: Simulated datasets were generated using R software. Prior to estimation, all relevant assumptions for both multivariate and multiple linear regression were tested. Model evaluation involved examining residual symmetry, assessing the influence of independent variables on dependent variables, and testing overall model significance. Multivariate regression performance was assessed using Pillai's Trace, Wilks' Lambda, Hotelling–Lawley Trace, and Roy's Largest Root, while the multiple linear regression model was evaluated using analysis of variance.

Results: The findings revealed that the multivariate linear regression model outperformed the multiple linear regression model under non-normal conditions. The multivariate approach demonstrated more stable residual behavior and stronger statistical evidence of relationships among variables, indicating superior model fit and robustness.

Novelty and contribution: This study provides new empirical evidence on the comparative suitability of regression models under non-normal data conditions, highlighting the superior performance of multivariate linear regression in handling multiple dependent and independent variables simultaneously.

Practical and social implications: The results offer practical guidance for researchers, data analysts, and policymakers in selecting appropriate analytical techniques for non-normal multivariate datasets. Improved model selection enhances the accuracy of empirical findings and supports more reliable decision-making in education, health, and the social sciences.

Keywords: Non – Normal data, Residual, Pillai, Wilks Lambda, Hotelling Lawley Trace, Roy

*Corresponding author

E-mail address: isiaqkamaldeen3@gmail.com

How to cite this article:

Isiak, K. O., Musa, O. Y., Ishola, K. B., Sanda, A. K., & Abubakar, A. O. (2025). The effect of multivariate regression over multiple linear regression models using non-normal data. *Elicit Journal of Science Research, 1*(1), 28–39.

1 Introduction

A statistical method called multivariate regression describes the relationship between the several dependent variables (responses) and two or more independent variables (predictors) at the same time. Multivariate regression extends the framework to account for interdependencies among multiple outcomes, unlike multiple regression, which predicts a single outcome variable. This makes it especially helpful in domains like social sciences, economics, medicine, engineering and business management, where variables are naturally correlated (Rencher & Christensen, 2012).

The method enables researchers to determine how explanatory variables jointly influence several variables, while also considering the correlations among those dependent variables (Johnson & Wichern, 2007). Thereby, multivariate regression provides more efficient parameter estimates and increases statistical power compared to running several separate regressions especially when comparing models using non normal data.

Multivariate regression has a wide range of applications. It can be used, for instance, to examine how socioeconomic status and instructional strategies interact to affect students' reading and math scores. It is used in medicine to investigate how risk variables like age, nutrition, and exercise affect a variety of health outcomes, including body mass index, blood pressure, and cholesterol levels (Tabachnick & Fidell, 2019). To account for shared variance in learning outcomes, for instance, student performance in science, math and reading might be jointly modelled in education research unlike multiple linear regression where dependent variable will be modelled individually with several independent variables.

Compared to performing independent regression for every outcome, multivariate regression offers the advantage regression of modelling outcome variables that may be associated, which lowers error variance and provides a more effective estimation of regression coefficients (Hair et al., 2019). Stronger presumptions, including the multivariate normality of residuals and the homogeneity of variance – covariance matrices, are necessary, nevertheless, and must be verified prior to use.

The aim of this article is to show the effect of the multivariate linear regression model on non normal data over the multiple linear regression model in a situation where the mean vectors are equal and variance – covariance matrices are known.

2 Literature Review

(Minhui, 2006) creates a new intelligent data mining and knowledge discovery method that is computationally possible to choose the optimal subset of predictors for multivariate regression (MR) models, assuming that the model's random error terms belong to a generic non - normal family of distributions. By combining intelligent statistical modeling techniques based on the information-theoretic measure of complexity (ICOMP) criterion with genetic algorithms (GA) and multivariate non-normal regression models with Power Exponential (PE) and family of elliptically contoured (EC) error distributions, our method creates an intuitive three-way hybrid approach. EC assumptions is demonstrated using both real and simulated data. The new method is advised for intelligent data mining where the data do not match the conventional normal assumption and with multivariate skewed PE regression models that can handle skewness and kurtosis simultaneously, as well as model selection issues.

(AbuElgasim, 2022) provided a logistic regression model and a multiple linear regression model based on the presumptions of both models. Because the dependent variable is nominal, the study relied on a logistic regression model. To determine the impact of student grades and gender, which are independent variables, on the dependent variable of student status, it also used data from the previous year's preparatory year that was gathered from Qassim University's College of Business and Economics. While gender has a major impact on student status, grades do not significantly affect it, according to the study. When the logistic regression model and multiple regression model were compared, it became clear that the logistic regression model was best suited to ascertain the relationship between the students' grades and gender as independent variables and their status as a dependent variable. The use of logistic regression is advised throughout the study, particularly when dealing with nominal dependent variables.

(Varsha, 2023) focuses on multivariate analysis of variance for the dataset of iris flowers, which comprises four dependent variables and three species. Using a combination of four dependent variables, the objective is to determine whether the flower morphology of three iris species varies. And used the Pillai's trace test since it is the most effective choice in the event that the homogeneity of variance assumptions of the MANOVA are broken. When

there are numerous independent and dependent variables, multivariate analysis of variance is employed. It creates a blend that divides the independent variable groups by linearly combining several dependent variables.

(Yiming, 2023) investigates the theoretical development and model applications of multiple regression to demonstrate the flexibility and broadness of the adoption of multiple regression analysis. Four different kinds of regression are explored individually. Four kinds of regressions are multivariate/multiple linear regression, multivariate multiple linear regression, multinomial logistic regression, and multivariate non-linear regression. Multivariate multiple linear regression is more accurate than multivariate/multiple linear regression when dealing with more than a variable. Multinomial logistic regression is relatively mature and accurate to solve the problem of non-linearity and multiple independent variables. It does not require the variable to obey multivariate normal distribution.

(Rosa, 2015) explains that, it is typical for a single study to have several outcomes of relevance in health – related research. Theses results are frequently examined independently, disregarding their relationship. A multivariate technique out to be a more effective substitute for separate studies of every result. This is not always the case, which is surprising. In the study, several settings of linear model were covered and compare the multivariate and univariate approaches using non normal data. Demonstrates that for linear regression models, the multivariate and univariate models' estimates of the regression with covariates shared across the outcomes are identical, but the multivariate model outperforms the univariate model in terms of efficiency for outcome – specific covariates.

(Johannes, 2022) proposed a new solution which is obtained by modelling the error term distribution through a finite mixture of multi – dimensional Gaussian components. The multivariate linear regression model is studied under this assumption. Identifiability conditions are proved and maximum likelihood estimation of the model parameters is performed using the EM algorithim. Model selection criteria are used to determine the number of mixture components; if this number is equal to one, the classical approach is the outcome of the proposed. Through Monte Carlo trials, the suggested approach's performances are assessed and constrasted with those of alternative methods. Finally, the findings from the examination of an actual dataset are shown.

3 Materials and Methodology

Definition 1. The multivariate linear regression model $y_i = \beta^T x_i + e_i$ for $i = 1,\ldots m$ has $m \geq 2$ response variables $Y_i, \ldots, Y_m$ and predictor variables $X_1, X_2, \ldots, X_p$ where $X_1 = 1$ is the trivial predictor. The i^{th} case is $(x_i^T, y_i^T) = (1, x_{i2,\ldots,} x_{ip,} Y_{i1}, \ldots, Y_{im})$ where the 1 could be omited. The main aim of fitting multivariate model is by looking for joint relationship of $Y_1, Y_2, \ldots, Y_m$ on $X_1, X_2, \ldots, X_p$

$$Y_1 = f(X_{11}, X_{12}, \ldots, X_{1p})$$

$$Y_2 = f(X_{21}, X_{22}, \ldots, X_{2p})$$

.

.

.

$$Y_m = f(X_{1i}, X_{2i}, \ldots, X_{ip})$$

And $X \sim Non - Normal_p(\mu, \Sigma)$, X is partitioned into a $a \times 1$ vector known as X_1 and a $(m-p) \times 1$ vector known as X_2

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_m \\ X_{m+1} \\ X_{m+2} \\ \vdots \\ X_{m+r} \end{bmatrix} \qquad \sum = \begin{bmatrix} \Sigma_{11(m \times m)} & \Sigma_{12(m \times p)} \\ \Sigma_{21(p \times m)} & \Sigma_{22(p \times p)} \end{bmatrix} \qquad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \\ \mu_{m+1} \\ \mu_{m+2} \\ \vdots \\ \mu_{m+r} \end{bmatrix}$$

$$E(X_1/X_2 = x_2) = \mu_1 + \Sigma_{x_1 x_2} \Sigma_{x_2 x_2}^{-1} (X_2 - \mu_2) \qquad \text{Equa}^1$$

The equ[1] above is conditional expectation of X_1/X_2 which used to fitting models involving multivariate dependent varaiables $m \geq 2$ and several independent variables jointly. The equ[1] can be decomposed as follows

$$E(X_1/X_2 = x_2) = \overline{X} - S_{12}S_{22}^{-1}\overline{X}_2 + S_{12}S_{22}^{-1}X_2 \qquad \text{Equa}^2$$

Where

$$\overline{X} - S_{12}S_{22}^{-1}\overline{X}_2 = \beta_o \text{ and } S_{12}S_{22}^{-1}X_2 = \beta_1$$

$$\beta = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdot & \cdot & \cdot & \beta_{1r} \\ \beta_{21} & \beta_{22} & \cdot & \cdot & \cdot & \beta_{2r} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \vdots & \vdots & \cdot & \cdot & \cdot & \cdot \\ \beta_{p1} & \beta_{p2} & \cdot & \cdot & \cdot & \beta_{pr} \end{bmatrix} \quad \text{and} \quad X = \begin{bmatrix} X_{m+1} \\ X_{m+2} \\ \cdot \\ \vdots \\ X_{m+r} \end{bmatrix}$$

$$(Y_1 - \overline{Y}_1) = \widehat{\beta}_{11}(X_1 - \overline{X}_1) + \widehat{\beta}_{12}(X_2 - \overline{X}_2) + \ldots + \widehat{\beta}_{1p}(X_{1m} - \overline{X}_{1m})$$

$$(Y_2 - \overline{Y}_2) = \widehat{\beta}_{21}(X_1 - \overline{X}_1) + \widehat{\beta}_{22}(X_2 - \overline{X}_2) + \ldots + \widehat{\beta}_{2p}(X_{2m} - \overline{X}_{2m})$$

.

.

.

$$(Y_1 - \overline{Y}_1) = \widehat{\beta}_{p1}(X_1 - \overline{X}_1) + \widehat{\beta}_{p2}(X_2 - \overline{X}_2) + \ldots + \widehat{\beta}_{pr}(X_{im} - \overline{X}_{im})$$

So, the joint multivariate models above can be redefined as:

$$Y_{i1} = \beta_{o1} + \beta_{11}x_i + \beta_{21}x_{i2} + e_{i1} \qquad \text{for } i = 1, \ldots, m \qquad \text{Equ}^3$$

$$Y_{i2} = \beta_{o2} + \beta_{12}x_i + \beta_{22}x_{i2} + e_{i2} \qquad \text{for } i = 1, \ldots, m \qquad \text{Equ}^4$$

Where $Y_i = \left(\frac{Y_{i1}}{Y_{i2}}\right)$

Then, equation 3 and 4 can be written jointly as

$$Y_i = \left(\frac{\beta_{o1}}{\beta_{o2}}\right) + \left(\frac{\beta_{11}}{\beta_{12}}\right)x_i + \left(\frac{\beta_{21}}{\beta_{22}}\right)x_i^2 + e_i \qquad \text{for } i = 1, \ldots, m \qquad \text{Equ}^{4i}$$

The hypothesis Testing when Σ^2 is known and when Σ^2 is unknown

$H_o : \beta_o = 0$ vs $H_1 \neq 0$ i.e under H_o,.

Test Statistic

In multivariate, there are four multivariate test statistic which are Wilks' Lambda Λ, Pillai's Trace Ψ, Hotelling – Lawley Trace T and Roy's Largest Root Θ. Wilks' Lambda test statistics with only Wilks' Lambda defined below.

$$|\Lambda| = \frac{|S_{11} - S_{12}S_{22}^{-1}S_{21}|}{|S_{11}|}$$

$$\frac{(N-r-1)-q+1}{q} * \frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}} \sim F_{2q,\ 2(N-r-q)}$$

The hypothesis will be considered insignificant if the p – value is less than α (0.05), but it will be considered significant if the p – value greater than α (0.05). Or The hypothesis will be considered insignificant if the Test Statistic greater than $F_{2q,\ 2(N-r-q)}$, but it will be considered otherwise if Test Statistic less than $F_{2q,\ 2(N-r-q)}$.

P – value = $P(\text{Test Statistic} \geq \text{Observed value}|H_o) = 1 - f_{cdf}(F, df_1, df_2)$

P – value = $2\min\left[P(F_{n_1-1,\ n_2-1} \leq F_0), P(F_{n_1-1,\ n_2-1} \geq F_0)\right]$ $\qquad \text{Equa}^5$

Definition 2. The model $y = X\beta + e_i$, it is assumed that the errors are normally and independently distributed with constant variance σ^2 or $e_i \sim N(0, \sigma^2 I)$.

The normal density function for the errors is

$f(e_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left[-\frac{1}{2\sigma^2}e_i^2\right]}$ $i = 1, 2, \ldots, n$ Equa[6]

The method of MLE (Maximum Likelihood Estimation) of $e_1, e_2, \ldots, e_n$ is used in estimating the parameter coefficients of the multiple linear regression model $y = X\beta + e_i$ as follows

$$L(\beta, \sigma^2) = \prod_{i=1}^{n} f(e_i)$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n} e_i^2\right]}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n} e'e\right]}$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n} (y - X\beta)'(y - X\beta)\right]} \quad \text{Equa[7]}$$

By taking monotonic log transformation of the Equa[7], we have

$$\text{In } L(\beta, \sigma^2) = -\frac{n}{2} In(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta) \quad \text{Equa[8]}$$

Since, the interest is to obtain the coefficients of the multiple linear regression model, differentiate partially with respect to $\underline{\beta_i}$ and equate to zero i.e $\frac{\partial InL(\beta, \sigma^2)}{\partial \beta} = 0$. Then we have

$$\frac{1}{\sigma^2} X'(y - X\beta) = 0$$

$$\beta = (X'X)^{-1}(X'y) \quad \text{Equa[9]}$$

The hypothesis Testing

$H_o : \beta_o = 0$ vs $H_1: \beta_o \neq 0$ i.e under H_o (Test whether all the regression coefficinent is significantly equal to zero)

Test Statistic

$$F = \frac{R^2/k}{(1-R^2/n-k-1)} \quad \text{Equa[10]}$$

And R^2 is called coefficient if determination which determine how good the model is fit.

$$R^2 = \frac{\hat{\beta}'X'Y - \frac{Y'Y}{n}}{(n-1)\sigma_y^2} \quad \text{Equa[11]}$$

The null hypothesis is rejected, if the P – value less than the significance level $\alpha(0.05)$ otherwise the null hypothesis is not rejected. And the value of R^2 ranges from 0 to 1. The higher R^2 shows a better fit i.e more variance explained.

P – value = $2\min\left[P\left(F_{n_1-1,\ n_2-1} \leq F_0\right), P\left(F_{n_1-1,\ n_2-1} \geq F_0\right)\right]$

Method of Simulation

The non normal data used for this study under multivariate linear regression model were simulated for two dependent variables and three independent variables through gamma distribution using R – Statistical Software. The sample size considered was $n = 200$ and the iteration was done hundred times i.e $i = 1, 2, \ldots, 100$ i.e three independent variables were generated and each was replicated 100 times.

4 Results of Analysis for Multivariate Linear regression

Response Y1 :

Table 1 Result for residuals

Min	1Q	Median	3Q	Max
-5.6219	-1.1777	0.0829	1.2995	5.0789

The table 1 shows that a good model fit (no significant bias) is indicated by residuals (errors) that are approximately symmetric arround zero.

Table 2 Coefficients

	Estimate	Std. Error	F	Pr(>|F|)
(Intercept)	3.43146	0.47113	7.284	7.73e-12
X1	1.16836	0.07301	16.003	< 2e-16
X2	-0.53680	0.04719	-11.376	< 2e-16
X3	1.93644	0.28053	6.903	6.91e-11

The regression equation is

$$\widehat{Y} = 3.43146 + 1.16836X_1 - 0.53680X_2 + 1.93644X_3$$

When All X's = 0, the predicted Y = 3.43146. Highly significant. For every 1 – unit increase in X_1, increase by 1.168 when other variables are constant.Highly significant. For every 1 – unit increase in X_2, Y decrease by 0.537 when other variables are constant. Highly significant. For every 1 – unit increase in X_3, Y increases by 1.936. Highly significant.

Also, from the table all predictors and the intercept are highly significant (P – value < 0.05)

Table 3 Results for R^2 and R^2_{adj}

Std error	df	R^2	R^2_{adj}	F – statistic	df	p – value
1.94	196	0.6803	0.6754	139	196	< 2.2e-16

Table 3: shows the R^2 (0.6803), this indicates that three independent variables account for 68.03% of the variation in the dependent variable. Random error and other variables not included in the model account for the remaining 31.97% of variation. R^2_{adj}(0.6754) is the modified form of R^2 that accounts for the model's predictor count. Its result shows that 67.54% of the variation is still explained, which is still rather strong and indicates that the model fits the data well with little overfitting. The overall significance of the model is indicated by the extremely small P – value (< 2.2e-16) and the relatively high F – value (139). Consequently, a considerable portion of the variance in the dependent variable can be explained by the predictors taken together.

Response Y2:

Table 4 Result for residuals

Min	1Q	Median	3Q	Max
-4.7503	-1.3497	0.1557	1.4814	5.6008

The table 4: shows that a good model fit (no significant bias) is indicated by residuals (errors) that are approximately symmetric arround zero.

Table 5 Coefficients

	Estimate	Std. Error	F	Pr(>|F|)
(Intercept)	-0.99157	0.49903	-1.987	0.0483
X1	0.74716	0.07733	9.661	< 2e-16
X2	0.35483	0.04998	7.099	2.25e-11
X3	-1.66072	0.29715	-5.589	7.58e-08

The regression equation is

$$\widehat{Y} = -0.99157 + 0.74716X_1 + 0.35483X_2 - 1.66072X_3$$

When All X's = 0, the predicted Y = - 0.99157. Highly significant. For every 1 – unit increase in X_1, increase by 0.747 when other variables are constant.Highly significant. For every 1 – unit increase in X_2, Y increases by 0.35483 when other variables are constant. Highly significant. For every 1 – unit increase in X_3, Y decreases by 1.66072. Highly significant.

Table 6 Results for R^2 and R^2_{adj}

Std error	df	R^2	R^2_{adj}	F – statistic	df	p – value
2.055	196	0.4776	0.4696	59.72	196	< 2.2e-16

Table 6: shows the R^2 (0.4776), this indicates that three independent variables account for 47.76% of the variation in the dependent variable. Random error and other variables not included in the model account for the remaining 52.24% of variation. R^2_{adj}(0.4696) is the modified form of R^2 that accounts for the model's predictor count. Its result shows that 46.96% of the variation is still explained, which is still partially strong and indicates that the model partially fits the data well with no overfitting. The overall significance of the model is indicated by the extremely small P – value (< 2.2e-16) and the relatively high F – value (139). Consequently, a considerable portion of the variance in the dependent variable can be explained by the predictors taken together.

Table 7 Result for Wilks Lambda Test Statistic

	df	Wilks	F	df	df	Pr(>|F|)
X1	1	0.37001	166.010	2	195	< 2.2e-16
X2	1	0.52293	88.950	2	195	< 2.2e-16
X3	1	0.70594	40.613	2	195	1.799e-15

Table 7: For X_1 (Wilks = 0.37001 i.e very strong multivariate effect), indicates a significant portion of the multivariate variance may be explained by X_1, For X_2 (Wilks = 0.52293 i.e strong multivariate effect), implies a significant portion of the multivariate variance may be explained by X_2 and For X_3 (Wilks = 0.70594 moderate multivariate effect), indicates a significant portion of the multivariate variance may be explained by X_3. F – values (166.010, 88.950 and 40.613) implies that the group mean differ considerably across the dependent variables. Therefore, the null hypothesis is rejected which indicates that X_1, X_2 and X_3 have no multivariate effect on the dependent variables. A substanstial correlation is shown by the very high F- values (166.010, 88.950, 40.613) and low Wilks Lambda (0.37001, 0.52293 and 0.70594). Therefore, among all the three variables, X_2 and X_3 show the strongest multivariate relationship with X_1 showed the strongest multivariate effect.

Table 8 Result for Pillai Test Statistic

	df	Pilla	F	df	df	Pr(>|F|)
X1	1	0.62999	166.010	2	195	< 2.2e-16
X2	1	0.47707	88.950	2	195	< 2.2e-16
X3	1	0.29406	40.613	2	195	1.799e-15

Table 8: For X_1 (Pillai = 0.62999 indicates strongest multivariate effect), For X_2 (Pillai = 0.47707 indicates stronger multivariate effect) and For X_3 (Pillai = 0.29406 indicates moderate multivariate effect). The multivariate effects of X_1, X_2 and X_3 on the dependent variables are statistically significant as P – values are incredibly small (<0.05). The combination of all the dependent variables are strongly influenced by all three predictors, but X_1 has the strongest overall impact.

Table 9 Result for Hotelling Lawlet Test Statistic

	df	Hotelling-Lawley	F	df	df	Pr(>|F|)
X1	1	1.70266	166.010	2	195	< 2.2e-16
X2	1	0.91231	88.950	2	195	< 2.2e-16
X3	1	0.41654	40.613	2	195	1.799e-15

Table 9: For X_1 (Hotelling-Lawley = 1.70266 indicates very strong multivariate effect), For X_2 (Hotelling-Lawley = 0.91231 indicates strong multivariate effect) and For X_3 (Hotelling-Lawley = 0.41654 indicates moderate multivariate effect). The multivariate effects of X_1, X_2 and X_3 on the dependent variables are statistically significant as P – values are incredibly small (<0.05). The combination of all the dependent variables are strongly influenced by all three predictors, but X_1 has the strongest overall impact.

Table 10 Result for Roy Test Statistic

	df	Roy	F	df	df	Pr(>|F|)
X1	1	1.70266	166.010	2	195	< 2.2e-16
X2	1	0.91231	88.950	2	195	< 2.2e-16
X3	1	0.41654	40.613	2	195	1.799e-15

Table 10: For X_1 (Roy = 1.70266 indicates very weak multivariate effect), For X_2 (Roy = 0.91231 indicates weak multivariate effect) and For X_3 (Roy = 0.41654 indicates strong multivariate effect). The multivariate effects of X_1, X_2 and X_3 on the dependent variables are statistically significant as P – values are incredibly small (<0.05). The combination of all the dependent variables are strongly influenced by all three predictors, but X_3 has the strongest overall impact.

All multivariate tests (Pillai, Wilks Lambda, Hotelling Lawley Trace and Roy) confirm that X_1, X_2 and X_3 are significantly influence the dependent variables, with X_3 showed the strongest overall effect and X_1 showed the weakest effect. The table 10 demonstrates that all three predictors are statistically significant, influencing the multivariate outcome.

Result for Multiple Linear Regression Analysis

The regression equation is

$\widehat{Y}_1 = 5.82 + 1.01X_1 - 0.751X_2 - 1.68X_3$

Table 11 Coefficients

Predictor	Estimate	Std. Error	F	Pr (>F)
(Intercept)	5.823	1.057	5.51	0.031
X1	1.0105	0.1597	6.33	0.024
X2	-0.7509	0.1251	-6.00	0.027
X3	-1.6777	0.7995	-2.10	0.171

Table 11: When All X's = 0, the predicted Y = 5.823. Highly significant. For every 1 – unit increase in X_1, Y increase by 1.0105 when other variables are constant.Highly significant. For every 1 – unit increase in X_2, Y decrease by 0.7509 when other variables are constant. Highly significant. For every 1 – unit increase in X_3, Y deccreases by

1.6777. Highly significant. Also, from the table all predictors and the intercept except X_3 are highly significant (P – value < 0.05).

Table 12 Results for R^2 and R^2_{adj}

STD Error	R^2	R^2_{adj}	F	P - value
0.617269	0.974	0.936	1.68909	0.0456

The table 12 shows the regression model that explains a very large proportion (97.4%) of the variation in the outcome variable, and the model is still robust even after adjustment (93.6%). Because the model is statistically significant (p = 0.0456), the dependent variable is significantly impacted by the predictors taken together. Reasonably reliable forecasts are indicated by the standard error.

Table 13 Results for ANOVA (Analysis of Variance)

Source	df	SS	MS	F	Pr(>F)
Regression	3	28.9050	9.6350	2440.36	0.0001
Residual Error	193	0.7620	0.003948		
Total	196	29.6671			

Table 13 shows the result of ANOVA. P = 0.0001 < 0.05, this indicates that there is statistcal significance in the regression model. And this implies that variation in the dependent variable is significantly explained by the independent factors when considered collectively. i.e a considerable amount of the variability in the dependent variable may be jointly explained by the predictors, according to the statistically significant regression model F(3, 193) = 2440.36, P – value = 0.0001.

The regression equation is

$\widehat{Y}_2 = - 0.76 + 0.470X_1 + 0.546X_2 - 0.56X_3$

Table 14 Coefficients

	Estimate	Std. Error	F	Pr (>F)
(Intercept)	-0.755	4.530	-0.17	0.860
X1	0.4697	0.6843	0.69	0.410
X2	0.5462	0.5363	1.02	0.310
X3	-0.563	3.426	-0.16	0.870

Table 14: When All X's = 0, the predicted Y = -0.755. Highly significant. For every 1 – unit increase in X_1, Y increase by 0.4697 when other variables are constant.Highly significant. For every 1 – unit increase in X_2, Y increase by 0.5462 when other variables are constant. Highly significant. For every 1 – unit increase in X_3, Y deccreases by 0.563. Highly significant. Also, from the table all predictors and the intercept are highly not significant (P – value > 0.05).

Table 15 Results for R^2 and R^2_{adj}

STD Error	R^2	R^2_{adj}	F
2.64545	0.634	0.085	2.90563

The table 15 shows the R^2 (0.634), this indicates that three independent variables account for 63.4% of the variation in the dependent variable. Random error and other variables not included in the model account for the remaining 36.6% of variation. R^2_{adj}(0.085) is the modified form of R^2 that accounts for the model's predictor count. R^2_{adj} implies that 8.5% of the variation is still explained, which implies a strong evidence that the model fits the data well with no overfitting. But the overall significance of the model is not indicated by the P – value (0.49817) and the relatively

high F – value (2.90563). Consequently, a considerable portion of the variance in the dependent variable can be explained by the predictor X_3.

Table 16 Results for ANOVA (Analysis of Variance)

Source	df	SS	MS	F	Pr(>F)
Regression	3	24.247	8.082	21514.84	0.000
Residual Error	193	0.0725	0.00038		
Total	196	24.3195			

Table 16 shows the result of ANOVA. P = 0.000 > 0.05, this indicates that there is no statistcal significance in the regression model. And this implies that variation in the dependent variable is not significantly explained by the independent factors when considered collectively. i.e a considerable amount of the variability in the dependent variable may be individually explained by the predictors, according to the statistically significant regression model F(3, 193) = 21514.84, P – value = 0.000.

5 Discussions of Findings

The findings reveals that the multivariate linear regression exhibited a strong effect on non normal data compared to the multiple linear regression model, as it captures multiple dependent and independent variables jointly and enhance statistical power. The methodology accounts for non normal data under the assumption of equal variance – covariance matrices and equal mean vectors. Conversely, scenarios involving unequal variance – covariance matrices and unequal mean vectors are not considered.

Tables 1, 4 showed the goodness of the fitted mode (no significant biass) by errors that approximately symmetric arround zero. Table 2, 5 indicate the significance of the coefficient of the fitted models under multivariate regression, table 3 and table 6 showed the results of coefficient of determination and adjusted coefficient of dertermination. R2 and R2adj are the determinant for the determine the significance of the fitted model and the results for the two tables i.e tables 3 and 6 indicated that the two models under multivariate regression are fitted. Table 7, 8, 9 and 10 showed the results obtained by all multivariate tests statistic (Pillai, Wilks Lambda, Hotelling Lawley Trace and Roy) confirm that all the three independent variables (X1, X2 and X3) are significantly influenced the dependent variable with X1 showed the strongest overall effect. The results of table 11 shows that there is significant decrease in predicted value for X1 and X3 and same as that of table 14. Table 12 and 15 showed that the model fit the simulated data well with a no overfitting but with a considerable portion of variance in the dependent variable which is explained by X3. Table 13 and 16 indicate the results obtained for ANOVA which explained that there is no significant difference in the regression models. Therefore, the results obtained in tables 7, 8, 9 and 10 flow with (Varsha D, 2023) that focusses on multivariate analysis of variance for the dataset of iris flower which comprises four dependent variables and three independent variables using Pillai trace test statistic. And the results of table 11 and table 14 flow with (Yiming S, 2023), multivariate/multiple linear regression, multivariate non – linear regression performed outrightly than multiple linear regression and multinomial logistic regression. Similarly, (Rosa O, 2015), demonstrated that for linear regression models, the Multivariate and univariate models' estimates of the regression parameters associated with covariates shared across outcomes are identical, but multivariate regression models outperformed the univariate models in terms of efficiency for outcome – specific covariate. This result flow along with the result obtained in tables 1 – 10.

6 Conclusions

This study compared the performance of multivariate linear regression and multiple linear regression models using non-normal data under the conditions of equal mean vectors and equal variance-cavariance matrices. Both models were fitted, and goodness-of-fit assessments, including significance testing, were conducted to determine their effectiveness. The findings consistently showed that the multivariate linear regression model demonstrated a stronger capacity to handle non-normal data, providing more reliable parameter estimates and capturing interdependencies

among variables more effectively than the multiple linear regression model. This result reinforces the analytical advantage of multivariate approaches in complex data environments where tradtional assumptions are violated.

Implication for Theory Development

The study contributes to statistical modelling theory by providing empirical evidence that strengthens the theoretical justification for using multivariate linear regression in situations characterized by non-normality. It supports the expansion of existing modelling frameworks to incorporate robustness considerations, particularly when dealing with correlated outcomes and multidimensional datasets. These findings encourage further refinement of theoretical models that account for real-world data irregularities, promoting more flexible and inclusive statistical assumptions in applied research.

Limitations and Future Research Direction

Despite its contributions, the study is limited by its reliance on simulated datasets, which may not capture all complexities of real-world data. The assumption of known and equal variance-cavariance matrices may also restrict generalizability. Future research should examine the performance of these models using empirical datasets from different fields, explore scenarios with unequal covariance structures, and test the robustness of multivariate models under varying sample sizes and degrees of non-normality. Additionally, comparing alternative robust regression techniques may offer deeper insight into optimal model selection.

Ethical Consideration

In this work, a multivariate regression model with two dependent variables and three independent variables is fitted using simulated data produced by the statistical program R. Issues with informed consent, confidentiality, or privacy do not apply because the dataset is artificially generated and does not involve any human.

However, the assumption of non - normal data was considered. This guarantees reproducibility and makes it possible for further researchers to confirm and expand on the findings. Additionally, the study stays away from any kind of data modification that can provide skewed results or inaccurate model performance representations.

Moreover, the results obtained are based on simulated data and that any inferences made are merely methodological demonstrations rather than practical applications. In compliance with academicals and ethical standards, all software tools and R packages utilized in the analysis are appropriately acknowledged. As a result, the study complies with the ethical, transparent, and responsible data usage guidelines for computational research.

Acknowledgement

The authors would like to thank the Falculty of Natural and Applied Science of Nigerian Army University Biu and Department of Mathematics of Nigerian Army University Biu for providing the support and facilities.

Funding

The authors declare that no financial support was received for the research, authorship, and publication of this article.

Data Availability Statement

Simulated data were used for the research described on the article.

Conflict of Interest

The authors declare no conflicts of interest.

References

AbuElgasim, A. (2022). Important Comparison About the Use of Multiple Linear Regression and Logistic Regression with Applications. International Journal of Mathematics and Statistics Studies. https://www.eajournals.org/

Hair, J.F., Babin, B.J., Anderson, R.E., & Black, W.C., (2019). Multivariate Data Analysis (8th ed.). England Pearson Prentice. DOI: 10.4236/jhrss.2019.134027

Johanneses, H., Dang-Phuong-Lan, Nguyen., Jean-Francois, Aujol., Dominique, B., & Abdullatif, S., (2022). American Institute of Mathematical Sciences. PCA Reduced Gaussian Mixture Models with Applications in Superresolution.Volume 16, Issue 2: 341-366 https://doi:10.3934/ipi.2021053

Johnson, R.A., & Wichern, D.W. (2007) Applied Multivariate Statistical Analysis. 6th Edition, Pearson Prentice Hall, Upper Saddle River. DOI: 10.4236/ojs.2015.57069

Minhui, Liu., (2006). Multivariate Nonnormal Regression Models, Information Complexity, and Genetic Algorithms: A Three Way Hybrid for Intelligent Data Mining. Doctoral Dissertations Submited to University of Tennessee – Knoxville.https://trace.tennessee.edu/itk_graddiss/2007

Rencher, A. C., & Christensen, W.F. (2012). Methods of Multivariate Analysis, 3rd Edition, Wiley, Hoboken. https://doi.org/10.1002/9781118391686

Rosa, Oliveira., & Armando Teixeira-Pinto. (2015). Analyzing Multiple Outcomes: Is it Really Worth the use of Multivariate Linear Regression?. J Biom Biostat. An open access journal. http://dx.doi.org/10.4172/2155-6180.1000256

Tabachnick, B.G., & Fidell, L.S. (2019). Using Multivariate Statistics (7th ed.). Pearson. DOI: 10.4236/jfrm.2021.104025

Varsha D., & Dhananjay, R. (2023). Multivariant Analysis of Variance to Discriminate Groups Based on Dependent Variables. Research Square. https://doi.org/10.21203/rs.3.rs-3071840/v1

Yiming, S., Xinyuan W., Chi Zhang., & Mingkai, Z. (2023). Multiple Regression: Methodology and Applications. International Conference on Applied Mathematics, Modeling Simulation and Automatic Control. https://doi.org/10.54097/hset.v49i.8611