International Journal of Advanced Statistics and Probability, 11 (1) (2024) 20-28

International Journal of Advanced Statistics and Probability

Website: www.sciencepubco.com/index.php/IJASP

Research paper

Comparative analysis of some linear predictive models in the presence of multicollinearity

Owoyemi Q. A. [1] *, Bolakale A. [1]

[1] University of Ilorin, Ilorin, Nigeria
**Corresponding author E-mail: Owoyemiqudus3@gmail.com*

Abstract

This paper investigates the comparative performance of some linear predictive models in the presence of multicollinearity. By examining the efficacy of Ordinary Least Squares (OLS), Ridge Regression, Lasso Regression, and Elastic Net Regression, this study aimed to figure out the best method for building robust and interpretable models under such conditions. The research explores how these models address multicollinearity, focusing on coefficient stability, prediction accuracy, and variable selection. Through a rigorous analysis of simulated and real-world datasets, the study shows the strengths and weaknesses of each model, providing valuable insights for researchers and practitioners looking to mitigate the challenges posed by multicollinearity in selecting the most proper method for regression modeling. This will lead to the creation of a model with increased interpretability of the relationships between variables, less variance, and more dependable coefficient estimations.

Keywords: *Variance Inflation Factor; Multicollinearity; Regression Analysis; Regularization; Predictive Models.*

1. Introduction

Regression analysis is a fundamental part of statistical modeling. It enables us to understand the relationship between a dependent variable (what we are trying to predict) and one or more independent variables. The most general form is linear regression, in which we model the relationship using a linear equation. Multicollinearity, which occurs when the independent variables in a regression model are highly interrelated, is a significant difficulty in regression analysis. In non expert's words, multicollinearity means that the predictor variables "explain" each other to a significant degree. The most common type of regression is linear regression, which represents this relationship as a linear equation. The value of the dependent variable can be computed using the values of the independent variables. I Simply put, the predictor variables become extremely repetitive, greatly "explaining" each other's changes. The presence of multicollinearity poses various issues, including:

1) Unreliable Coefficient Estimates: When independent variables are highly correlated, the regression model has difficulty distinguishing their contributions to the dependent variable. This can produce coefficient estimates (slopes) with large variations and statistical insignificance. As a result, evaluating the true link between variables and their effects on outcomes becomes problematic (Hair et al., [1]).
2) Increased Variance: Multicollinearity can increase the model's total variance, making predictions less trustworthy and applicable to new data (James et al., [2]). The model's estimations are unstable, which diminishes their credibility and trustworthiness.
3) Difficulties in Interpretation: With correlated variables, it is difficult to separate the unique effect of each predictor on the result. Multicollinearity creates ambiguity, making it difficult to understand the genuine causal linkages underlying the data. These concerns have a substantial impact on a regression model's validity and usefulness. Ignoring multicollinearity might lead to misleading and erroneous results.

The prevalence of multicollinearity is not limited to theoretical concerns, it is a frequent occurrence in real-world datasets, particularly in fields like economics, finance, and social sciences, where variables often show inherent interdependencies. This ubiquity needs the development of robust techniques to address the challenges it presents.

1.1. Regularization techniques

Several techniques, known as regularization, can be employed to mitigate the effects of multicollinearity. These methods introduce a penalty term into the regression model, which discourages coefficients from becoming too large. Three prominent regularization techniques used in this research are:
Ridge Regression: Shrinks the coefficients of all predictor variables towards zero,
reducing their variance but not necessarily leading to variable selection.

Lasso Regression: Like Ridge regression, it shrinks coefficients, but with the added benefit of forcing some coefficients to become exactly zero. This effectively performs variable selection, finding the most important predictors.
Elastic Net Regression: Combines elements of Ridge and Lasso, offering a balance between shrinkage and variable selection.

1.2. Aim and goals of the study

This study will compare numerous regression approaches to decide their usefulness in dealing with the difficulty of multicollinearity. Goals are as follows:

1) Assess the performance of some regression techniques, including Ordinary Least Squares (OLS) regression, Ridge Regression, Lasso Regression, and Elastic Net Regression, in the presence of multicollinearity.
2) Investigate how regularization approaches (Ridge Regression, Lasso Regression, and Elastic Net Regression) might reduce the effects of multicollinearity on coefficient estimates and model variance.
3) Explain the best regression strategy for developing robust and interpretable models where multicollinearity is an issue.

2. Literature review

Multicollinearity occurs in linear regression when independent variables (predictors) show a high degree of linear correlation with one another (Menard, [3]). This association violates the concept of error independence, which is needed for accurate coefficient estimates and predictions in regression models (Montgomery & Chatterjee, [4]).
There are several ways for detecting multicollinearity in a regression model:

1) Correlation Matrix: Analyzing the correlation matrix between independent variables can provide an early indication of correlations. High correlations (numbers near 1 or -1) indicate multicollinearity (Gujarati, [5]). However, this strategy might be misleading when dealing with complex connections involving numerous variables.
2) The Variance Inflation Factor (VIF) is a more reliable measure of multicollinearity. Multicollinearity causes an estimated coefficient's variance to be exaggerated significantly. A VIF score larger than 5 or 10 is commonly seen as an indicator of problematic multicollinearity (Fox, [6]). VIF gives a more nuanced picture of the severity of multicollinearity for each independent variable.
3) Eigenvalue Analysis: Examining the correlation matrix's eigenvalues can help identify components with low variance, indicating multicollinearity (Belsley et al., [7]). This method is more complex and necessitates a thorough understanding of linear algebra.

It is important to highlight that no single method is completely reliable for detecting multicollinearity. A combination of these strategies is frequently employed to acquire a thorough knowledge of the issue's presence and severity in a given dataset (Montgomery & Chatterjee, [4]). Numerous studies have explored the impact of multicollinearity and the effectiveness of alternative regression techniques. Here are some relevant examples

1) Shen et al [8]: Ridge Regression was compared to Ordinary Least Squares (OLS) in the context of high-dimensional data (many predictor variables). Their study highlights the benefit of Ridge Regression in improving coefficient stability, particularly when dealing with multicollinearity.
2) Huang et al. [9]: Finding consumer purchasing drivers using LASSO regression: An empirical study of the mobile app market. Their work looks into the usage of LASSO regression for variable selection in marketing research. The authors demonstrate that LASSO may identify key marketing elements influencing client purchases, even with multicollinearity among variables. 3. Liu et al. [10]: Their study investigates the use of Elastic Net regression to deal with multicollinearity in high-dimensional data. The results imply that Elastic Net can beat Ridge and Lasso regression in terms of prediction accuracy and variable selection, especially when dealing with a large number of correlated variables.
3) Altelbany et al [11]: Evaluation of Ridge, Elastic Net, and Lasso Regression Methods in the Presence of multicollinearity, a simulation study on the multicollinearity problem. His research evaluates the performance of Ridge, Lasso, and Elastic Net regression in dealing with multicollinearity under various conditions.

These works show current academic attempts to better understand and address the issues that multicollinearity presents in regression analysis.

3. Methodology

The methodological approach employed in this study to investigate the impact of multicollinearity, and the effectiveness of alternative regression techniques focused on comparing the performance of Ordinary Least Squares (OLS), Ridge Regression, LASSO Regression, and Elastic Net Regression in the presence of a multi collinear dataset.
This study used two datasets to explore the effects of multicollinearity and the selected regression techniques:

3.1. Simulated datasets

The simulated dataset encompasses three variations in sample size (n = 40, n = 60, and n = 1000) to assess how the number of observations affects the impact of multicollinearity and the performance of different regression techniques. Additionally, the number of independent variables was varied across three levels (2, 3, and 5) to investigate the influence of data complexity on multicollinearity and model performance.
The simulated data includes a dependent variable and a set of independent variables that influence the dependent variable. The specific characteristics of these variables, including their distributions and relationships, were random to avoid any form of bias and introduce varying degrees of multicollinearity within the independent variables.
This simulated data allows for controlled manipulation of multicollinearity by adjusting the correlation coefficient (between 0.1 to 0.4 weak, 0.6 strong or moderate, and 0.9 strong degrees of association) among the independent variables.

3.2. Real-world datasets

To complement and buttress the controlled environment of the simulated data and ensure the generalizability of our findings, a real-world dataset was obtained from UnilorinWater Enterprise. This dataset holds information on the dependent variable (electrical conductivity) and a set of independent variables (PH, Hardness, Chlorides, and Volume) for a collection of 86 observations.

3.3. Correlation matrix

The correlation matrix was examined for high correlations (values close to 1 or -1) between independent variables. This will provide a preliminary sign of multicollinearity.

3.4. Variance inflation factor

VIF was estimated for each independent variable in both datasets. A VIF value greater than 5 or 10 will be considered an indicator of multicollinearity. This method provides a more robust measure of the severity of multicollinearity for each variable.

3.5. Model building and evaluation

Model Building: Regression models will be built using the following techniques for both the simulated and real-world datasets:
Ordinary Least Squares (OLS) Model:

TABLE 1: OLS Model Formulation

Observation number	Response y	Explanatory variables $X_{1...}X_k$
1	Y_1	X_1……X_k
2	Y_2	. .
.	.	. . .
.	.	. . .
.	.	. . .
N	Y_n	X_n……X_{nk}

Let an experiment be conducted n times, and the data obtained be as follows:
Assuming that the model is:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + e \text{ or } y = X\beta + e \quad (1)$$

In general, the model with k explanatory variables can be expressed as y= Xβ + e where y = (y1,y2,y3,…,yn)' is a n×1 vector of n observation on study variable,

$$X = \begin{pmatrix} x_{11} & x_{12} & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{nk} \end{pmatrix} \quad (2)$$

The column vectors $X_1, X_2,\ldots, X_k$ are linearly dependent if there exists a set of constants $\lambda_1, \lambda_2,\ldots, \lambda_k$ not all zero, such that

$$\sum_i^n \lambda_I X_i = 0 \quad (3)$$

If this holds exactly for a subset of the $X_1, X_2,\ldots, X_k$, then rank (X'X) < k Consequently (X 'X)$^{-1}$ does not exist. If the condition

$$\sum_i^n \lambda_I X_i = 0 \quad (4)$$

Is true for some subset of $X_1, X_2,\ldots, X_k$, then their will be a near-linear dependency in X'X.
In such a case, the multicollinearity problem exists. It is also said that X'X becomes ill-conditioned.

$$\beta_{OLS} = (X'X)^{-1} X'y \quad (5)$$

Which then decomposes with proper matrix operation to matrix,

$$\beta_{OLS} = \begin{pmatrix} \beta o \\ . \\ . \\ . \\ . \\ \beta p \end{pmatrix} \quad (6)$$

Ridge Regression: Ridge Regression: This technique introduces a penalty term that shrinks the coefficients of highly correlated variables towards zero. By reducing the size of coefficients, Ridge Regression helps to stabilize the variance of the model and improve the condition number, leading to more stable coefficient estimates, even in the presence of multicollinearity (Hoerl & Kennard, 1970). However, Ridge Regression does not perform variable selection and may keep all predictors in the model. The tuning parameter (lambda) controlling the shrinkage of coefficients was selected using cross-validation techniques. This helps to find the best lambda value that balances model fit and coefficient stability.
The problem of multicollinearity arises because some of the eigenvalue's roots of X'X are close to zero or are zero. So, if $\lambda_1, \lambda_{2...}, \lambda_k$ are the characteristic roots, and if X'X =Λ= diag($\lambda_1, \lambda_2,\ldots, \lambda_k$) then,

$$\beta_{ridge} = (1 + \delta\Lambda^{-1})^{-1} b \quad (7)$$

Where b is the OLSE of β given by

$$b = (X'X)^{-1} X'y = \Lambda^{-1} X'y \quad (8)$$

Thus, a particular element will be of the forms

$$\frac{1}{1+\frac{\delta}{\lambda_i}} b_i = \frac{\lambda_i}{\lambda_i+\delta} b_i \quad (9)$$

So, a small quantity δ is added to λ_i so that if $\lambda_i = 0$, even then $\frac{\lambda_i}{\lambda_i+\delta}$ stays meaningful.
Then,

$$\beta_{ridge} = (X'X+\lambda_i)^{-1} X'Y \quad (10)$$

Which then decomposes as pointed out in (6) with proper matrix operation to matrix $\beta_{ridge.}$
LASSO Regression: LASSO regression uses an L1 penalty, which shrinks some coefficients to zero. This effectively performs variable selection, finding the most important predictors and removing those with minimal contribution while mitigating the effects of multicollinearity (Tibshirani, 1996). Unlike Ridge Regression, LASSO promotes sparsity by setting some coefficients to exactly zero, offering a clearer picture of the most relevant variables for the model. This can be particularly helpful for interpreting the model and understanding the true causal relationships between the independent variables and the dependent variable. However, LASSO might not be the most suitable choice when the primary goal is prediction accuracy, as setting coefficients to zero can lead to slightly higher prediction errors compared to Ridge regression. The tuning parameter (lambda) controlling the sparsity of the model (number of coefficients set to zero) will be selected using cross-validation. This ensures the selection of the best lambda value for achieving variable selection while mitigating multicollinearity.
Lassos regression for regularization:

$$L_{lasso}(\beta) = \sum_{i=1}^{n}(y_i - x^s{}_i\beta)^2 + \lambda \sum_{i=1}^{m} |\beta_i| \quad (11)$$

Which then decomposes as pointed out in (6) with proper matrix operation to matrix β_{lasso}
Elastic Net Regression: This method combines L1 and L2 (Ridge) penalties, offering a balance between variable selection and reducing coefficient variance (Zou & Hastie, 2005). Elastic Net penalizes coefficients similarly to LASSO but also incorporates shrinkage from Ridge Regression. This can be particularly useful when dealing with many correlated predictors, allowing for some coefficients to be shrunk to 0 while still providing some stability for the remaining ones. Elastic net penalty tries to combine advantages of both lasso and ridge regression, namely shrinkage and sparsity together.
The elastic net regression minimizes

$$EN(\beta) = \sum_{i=1}^{n}(y_i - x^s{}_i\beta)^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} |\beta_j|^2 \quad (12)$$

Which then decomposes as pointed out in (6) with proper matrix operation to matrix $\beta_{Elastic\ net}$
Due to the ridge regularization, the elastic net can handle correlations between the predictors better than Lasso and due to the L_1 regularization, sparsity is obtained. However, the bias issue present for Lasso is still present for elastic net.

3.6. Model evaluation

The performance of each model was evaluated using the following criteria for both datasets:
Coefficient estimates and significance: The stability and significance of coefficients were assessed across models. This will help to understand how multicollinearity and the different regression techniques impact coefficient estimates.
Model fit statistics: Goodness-of-fit metrics like R-squared and adjusted R-squared were calculated for each model. These statistics show how well the model explains the variance in the dependent variable.
Prediction accuracy: The performance of models in terms of predicting the dependent variable on unseen data was compared.
Variable selection: The variables identified as important predictors by these techniques were analyzed. This will provide insights into which variables are most relevant in the presence of multicollinearity.
Mean Squared Error (MSE) and Root Mean Squared Error (RMSE): These statistics measure the average squared difference between the predicted values and the actual values of the dependent variable. Lower MSE and RMSE indicate better model fit, with

$$MSE\ (\beta_{OLS}) = \sum_{i}^{n} \frac{(y_i-\hat{y})^2}{n} \quad (13)$$

$$MSE\ (\beta_{ridge}) = \sum_{i}^{n} \frac{(y_i-\hat{y})^2}{n} + (\lambda \sum \beta^2) \quad (14)$$

$$MSE\ (\beta_{Lasso}) = \sum_{i}^{n} \frac{(y_i-\hat{y})^2}{n} + (\lambda \sum |\beta|) \quad (15)$$

$$MSE\ (\beta_{Elastic}) = \sum_{i}^{n} \frac{(y_i-\hat{y})^2}{n} + (\lambda_1 \sum \beta^2) + (\lambda_2 \sum |\beta|) \quad (16)$$

Where:
n is the number of data points
k is the number of independent variables in the model

y_i is the actual value for the ith data point
$\hat{y}$ is the predicted value for the i-th data point
$\bar{y}$ is the average value of the response variable
λ (lambda) is the regularization parameter controlling the penalty strength
|β| stands for the L_1 norm of the coefficient vector β (sum of absolute values of coefficients)
β^2 stands for the L_2 norm of the coefficient vector β (sum of squared values of coefficients)

Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC): These are information criteria used for model selection. Lower AIC and BIC values show a better balance between model fit and complexity (number of parameters). Therefore, the model with least AIC and BIC was chosen in this research. It was calculated for all the techniques by,

$$AIC = n\log(MSE) + 2k \tag{17}$$

R-squared and Adjusted R-squared: R-squared measures the proportion of variance in the dependent variable explained by the model. Adjusted R-squared penalizes R-squared for the number of predictors, providing a more reliable estimate of model fit for comparing models with different numbers of variables. Therefore, the model with high R-squared and Adjusted R-squared was chosen in this research. It was calculated for all the techniques by,

$$R^2 = 1 - \frac{\sum_i^n (y_i - \hat{y})^2}{\sum_i^n (y_i - \bar{y})^2} \tag{18}$$

$$\text{Adjusted } R^2 = 1 - (1 - R^2)(n-1/n-k-1) \tag{19}$$

Cross-validated R-squared: These metric addresses over fitting by evaluating the model's performance on unseen data. A high cross-validated R-squared shows good generalizability of the model. Therefore, the model with high Cross-validated R-squared was chosen in this research. It was calculated for all the techniques by,

$$R^2 = 1 - \frac{\sum_i^n (y_i - \hat{y})^2}{\sum_i^n (y_i - \bar{y}\boldsymbol{train})^2} \tag{20}$$

3.7. Comparative analysis

The performance of different regression techniques was compared based on the analysis of both datasets. Here is how the comparison was done:
Coefficient estimates and significance: Differences in coefficient estimates and their significance across models (OLS, Ridge, Lasso, and Elastic Net) were estimated. This will help to understand how multicollinearity and the chosen technique affect the interpretation of coefficients.
Model fit and prediction accuracy: The goodness-of-fit statistics (R-squared, adjusted R-squared) and prediction accuracy (assessed through cross-validation) of each model were compared. This will show how multicollinearity and the different techniques influence the model's ability to fit the data and make the right predictions.

4. Discussion of results

This section discusses the heart of the study, presenting the findings from the analysis of both the simulated and the real-world datasets. Our primary goal is to investigate the impact of multicollinearity on regression analysis and to compare the effectiveness of Ordinary Least Squares (OLS), Ridge Regression, LASSO Regression, and Elastic Net Regression in mitigating its effect.

4.1. The best methods from all scenarios using monte carlo simulations

Table 2: Results from Analysis of Various Scenarios

Sample size	rho	predictors	MSE	r^2	cross validated r^2	AIC	RMSE	ADJ R2	VIF BEFORE	VIF AFTER	selected predictors	best model
40	-0.4	2	0	0	1	4	0	1	1.2577	1.2344	X1, X2	OLS
60	-0.4	2	0	0	1	4	0	1	1.15	1.04	X1, X2	OLS
40	-0.4	3	0	0	1	6	0	1	1.152	1.2316	X1, X2, X3	OLS
60	-0.4	3	0	0	1	6	0	1	1.073	1.0044	X1, X2, X3	OLS
60	0.6	2	1.3	0.8944	-7.3567	6.62	1.1435	0.8944	1.52	1.37	X1, X2	OLS
40	0.6	2	0.9	0.9356	-0.644	7.75	0.9345	0.9195	1.25	1.19	X1, X2	OLS
40	0.6	3	0.8	0.9348	-21.0889	7.59	0.892	0.925	2.497	2.1614	X1, X2, X3	OLS
60	0.6	3	1.19	0.9552	0.9266	12.39	1.0939	0.9427	2.567	2.29	X1, X2, X3	ELASTIC-NET
60	0.6	5	0.9	0.9725	0.9587	11.9	0.9898	0.9725	2.87	3.3055	X1, X2, X3, X4, X5	OLS
1000	0.6	5	0.9	0.9801	0.9799	11.9	0.9688	0.9801	1.42	2.53	X2, X5	ELASTIC-NET
40	0.9	2	1.1	0.8759	-1.1666	6.19	1.0479	0.8568	8.11	7.12	X1, X2	OLS
60	0.9	2	1.3	0.9329	0.9266	6.51	1.119	0.926	4.233	2.94	X2	RIDGE

60	0.9	3	0.8	0.9743	0.9311	7.61	0.8956	0.9704	5.256	4.134	X2, X3	RIDGE
40	0.9	3	1.7	0.938	0.7448	9.41	1.3061	0.9226	13.02	17.5	X3	RIDGE
40	0.9	5	1.40	0.9664	0.737	12.80	1.184	0.9496	10.91	13.97	X3, X4	ELASTIC-NET
60	0.9	5	0.8	0.934	0.8541	5.52	0.8718	0.934	15.49	15.75	X1, X3, X4	RIDGE
1000	0.9	5	1.04	0.9817	0.9805	12.08	1.0209	0.9815	4.96	4.88	X5	ELASTIC-NET
1000	0.9	3	0.9	0.9652	0.9617	7.92	0.9809	0.9652	5.3644	4.961	X1, X2, X3	LASSO
1000	0.9	2	0.9	0.2887	0.218	5.83	0.9574	0.2851	5.8074	5.5928	X1,X2	LASSO

4.2. Interpretations

The table 2 above suggests that there is a significant difference in RMSE and Adjusted R^2 between Ordinary Least Squares (OLS) regression and other regression methods (Ridge Regression, Lasso Regression, and Elastic Net Regression) for sample sizes of 40 and 60 with low to moderate multicollinearity (rho between 0.1 and 0.4), knowing that the "best model" within each method was chosen, it suggests that even OLS might be performing reasonably well in these scenarios with low to moderate multicollinearity.

The result in table above reveals again that when multicollinearity increases (rho = 0.6) with 2 or 3 predictors, OLS has a higher RMSE compared to the best model (Ridge in these cases). This suggests that regularization techniques can outperform OLS in terms of prediction accuracy when multicollinearity is a concern.

However, in other cases with a larger number of predictors (like sample: 60, rho=0.6, predictors =5), even the "best model" (Elastic Net) might not show a significant improvement over OLS in terms of RMSE. This highlights the complexity of how multicollinearity and the number of predictors interact with different regression techniques. As multicollinearity increases, OLS performance tends to deteriorate in terms of prediction accuracy (RMSE).

However, the impact can vary depending on the number of predictors and the specific data structure. Regularization techniques can outperform OLS in these scenarios.

Table 2 also revealed that we can infer that OLS leads to higher VIF values compared to regularization techniques (Ridge and Elastic Net) in most cases (samples with rho = 0.6 or 0.9 and 3 or more predictors). This aligns with the expectation that regularization techniques address multicollinearity by shrinking coefficients, reducing VIF values. However, there are exceptions (samples: 60, rho = 0.9, 2 independent variable, and n =40, rho = 0.9, 3 independent variables), this highlights that the effectiveness of regularization in reducing VIF might also depend on specific data characteristics.

Regularization techniques are more effective than OLS in reducing VIF values, showing a mitigation of multicollinearity. However, the effectiveness might vary depending on the data structure.

The table also shows that regularization techniques (Lasso, or Elastic Net) selected a smaller subset of variables compared to OLS (e.g., samples n = 60 and n=40, rho = 0.6, 3 independent variables, and n = 60 and n = 40, rho= 0.9, 5 independent variables). This suggests that regularization techniques can perform variable selection while mitigating multicollinearity.

In other cases, with low to moderate multicollinearity (samples with rho between 0.1 and 0.4), OLS also selects the same predictors as the best model. This suggests that when multicollinearity is low or moderate, OLS might not suffer significantly from variable selection issues in these specific scenarios.

Therefore, when dealing with high multicollinearity and a larger number of predictors, regularization techniques can be valuable for selecting a smaller, more relevant subset of variables compared to OLS. However, in scenarios with low to moderate multicollinearity, , OLS might not necessarily perform poorly in terms of variable selection.

The result highlights the importance of considering multicollinearity when choosing a regression technique. While OLS might be a starting point, regularization techniques like Ridge, Lasso and Elastic Net can offer advantages in terms of handling multicollinearity, reducing VIF values, and performing variable selection, leading to more interpretable and robust models, particularly in scenarios with high multicollinearity and a larger number of predictors.

4.3. Presentation of the Real-world data

Table 3: Descriptive Statistics of Ph, Hardness, Electrical Conductivity, Chlorides and Volume Dataset

	N	Range	Mean	Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
PH	86	483.9549	2.023050	104.1483564	10846.880	.044	.260	.559	.514
Hardness	86	307.0823	1.538587	53.7570030	2889.815	.349	.260	.513	.514
Electrical Conductivity	86	692.4671	3.916719	149.4314795	22329.767	.099	.260	.282	.514
Chlorides	86	615.4001	7.581249	109.3379133	11954.779	.136	.260	.571	.514
Volume	86	518.0675	2.385268	106.2640763	11292.054	.181	.260	.359	.514

Table 3.2 above show the descriptive statistics of five different properties measured across 86 samples, the descriptive statistics provide a comprehensive overview of the central tendencies and variability of the measured water properties. Electrical Conductivity exhibited the most significant spread, while PH, Hardness, Chlorides, and Volume showed moderate variability around their respective averages. Additionally, the analysis revealed potential biases in the distribution of Hardness and Chlorides and reveals that the dataset exhibits moderate spread around the averages with flatter than normal distributions.

4.4. Validity of assumptions

4.4.1. Normality assumption

The normality assumption of the set of data was examined using Kolmogrov-Smirnov and Shapiro-Wilk's test and the result obtained were presented in the table 3.1.2 below

H_0: The Water quality parameters are not normally distributed

H_1: The Water quality parameters are normally distributed

Table 4: Test of Normality of the Distribution of Ph, Electrical Conductivity, Chlorides, Volume and Hardness

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	Df	Sig.	Statistic	Df	Sig.
PH	.076	86	.200*	.980	86	.203
Electrical Conductivity	.077	86	.200*	.984	86	.386
Chlorides	.054	86	.200*	.992	86	.890
Volume	.066	86	.200*	.985	86	.450
Hardness	.062	86	.200*	.986	86	.469

Based on the normality tests, the data for PH, Electrical Conductivity, Chlorides, and Volume appears to be normally distributed. The results for these properties show significance values greater than 0.05 (showed by *.200) in both the Kolmogorov-Smirnov and Shapiro-Wilk tests, which suggests we cannot reject the null hypothesis of normality. However, the normality of Hardness data stays inconclusive due to similarly high significance values in both tests.

4.4.2. Multicollinearity tests

The multicollinearity test was investigated using the following techniques:

- Tolerance and Variance inflation Factor
- Correlation Matrix

4.4.3. Tolerance and Variance Inflation Factor

Table 5: Multicollinearity test of the Unilorin Water Enterprise Dataset

Model	Unstandardized Coefficients	Standardized Coefficients	T	Sig.	Collinearity Statistics		
	B	Std. Error	Beta			Tolerance	VIF
	(Constant)	-.993	5.948		-.167	.868	
	PH	.933	.148	.650	6.317	.000	.150 6.672
	Hardness	.727	.212	.261	3.433	.001	.274 3.653
	Chlorides	-.055	.094	-.040	-.589	.557	.339 2.949
	Volume	.141	.095	.100	1.485	.142	.348 2.871

4.4.4. Interpretation

While no strict thresholds exist, tolerance values below 0.10 and VIF (Variance Inflation Factor) values above 10 show significant multi-collinearity. In this table, none of the tolerances fall below 0.10, but PH has a VIF of 6.67, exceeding a common benchmark for concern. This suggests PH might be highly correlated with other independent variables.

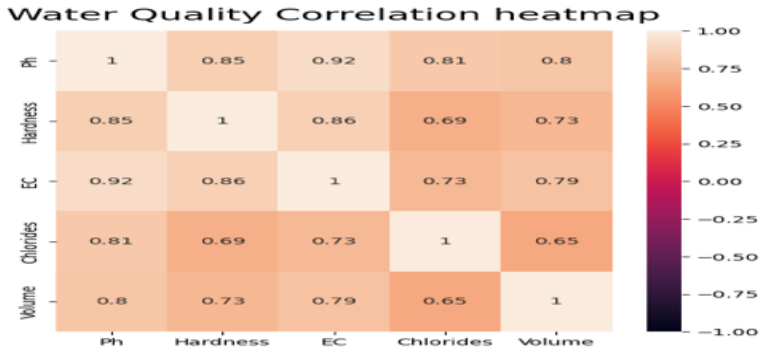

Fig. 1: Heat Map of Correlation Matrix for the Water Quality Dataset Obtained from Unilorin Water Enterprise.

Fig 1 above is the correlation matrix of the Water quality data which shows the strength of the linear relationship between different water quality parameters. It used a heat map where warmer colors showed a stronger positive correlation, cooler colors showed a stronger negative correlation, and white showed no correlation. Values closer to 1 show a strong correlation, while values closer to 0 show a weak correlation. The result suggests that minerals (hardness, EC) increase with higher PH, while volume tends to decrease with increasing PH and mineral content.

The water quality data suggests multicollinearity issues, particularly between hardness, EC, and PH. Their strong positive correlations mean they might influence each other in a way that makes it difficult to isolate their individual effects on another parameter, like volume. The multicollinearity test highlights a concern with PH being highly correlated with other variables. This can make it difficult to isolate the true effect of PH on electrical conductivity and might affect the model's reliability.

4.5. Multiple regression analysis results

The multiple regression analysis results obtained from Table 4 above for Electrical Conductivity (dependent variable) reveal multicollinearity concerns among the independent variables (PH, Hardness, Chlorides, and Volume). Notably, the Variance Inflation Factor (VIF) values for PH (around 6.672) are above the typical threshold of 5, showing inflation of its variance due to its correlation with other factors influencing conductivity. This makes it challenging to isolate the unique effect of PH on Electrical Conductivity.

4.5.1. Regression equation for electrical conductivity

Electrical Conductivity = -0.993 (Constant) + 0.933 * PH + 0.727 * Hardness - 0.055 * Chlorides + 0.141 * Volume
The regression equation offers a prediction for Electrical Conductivity based on these variables, interpreting the individual coefficients with absolute certainty might be difficult due to multicollinearity.

4.5.2. Model summary

The model summary in the table below suggests the model can predict Electrical Conductivity with moderate accuracy (MSE: 3013.809, RMSE: 54.898), explaining a substantial part of the variation (R-squared: 0.871, Adjusted R-squared: 0.865). Although the model complexity seems reasonable based on AIC (935.9) and BIC (945.7), the identified multicollinearity might limit our ability to isolate the unique influence of each factor (PH, Hardness, Chlorides, Volume) on Electrical conductivity.

Table 6: Summary of the Unilorin Water Enterprise Dataset

Mse	Rmse	R-Squared	Adj. R-Squared	Aic	Bic
3013.809	54.898	0.871	0.865	935.9	945.7

4.5.3. Data analysis and results for the UNILORIN WATER ENTERPRISE dataset

The Table below shows the comparative analysis results of Linear regression, Ridge regression, LASSO regression and Elastic net regression for Electrical conductivity, in testing for the best model fit and in handling the problem of multicollinearity using the SK learn function on Python.

Table 7: Results From Regression Techniques with Respect to the Criterions

Regression Techniques Criterion	Linear (Ols)	Ridge	Lasso	Elastic Net
Mse	3340.536638	3340.5367079	3340.5896	3340.5858
Rmse	57.7974	57.7978	57.825	57.823
R^2	0.8628	0.8628	0.8627	0.8627
R^2 Adjusted	0.84456	0.84456	0.844553	0.844553
Aic	6691.07327	6691.073	6691.1792	6691.1717
Bic	177.3256	177.3261	177.32615	177.3261
Cross Validated R^2	0.72050	0.720508	0.72051	0.720516

4.5.4. Interpretations

Based on the information provided in the table, we can tentatively deduce that while all the regression models achieved a high explanatory power (similar R-squared), the table suggests a significant difference in their suitability for handling multicollinearity when predicting electrical conductivity.
The table provides evidence that Ridge Regression performs better than the other models when dealing with multicollinearity in predicting Electrical Conductivity. This is supported by the balance between explaining the data (similar R-squared) and avoiding over fitting (lower AIC and BIC).
Based on the understanding of these regression techniques, linear regression is more susceptible to multicollinearity compared to Ridge, Lasso, and Elastic Net regression, leading to higher variance in coefficients and lower prediction accuracy.
We can also suggest that Ridge regression will improve coefficient stability compared to linear regression, especially when multicollinearity is present, although it might introduce some bias, the fact that Ridge Regression achieves similar prediction accuracy (similar MSE to Linear Regression) with lower model complexity (lower AIC and BIC) suggests it might be achieving more stable coefficients. This is because reducing coefficient variance can sometimes lead to simpler models with similar prediction accuracy.
Based on the information provided, we can see that, with coefficients close to zero, Lasso could lead to a more interpretable model by focusing on the variables with the largest coefficients.
Lasso might not have achieved strict variable selection in this specific case, the provided information suggests it achieved reasonable prediction accuracy and offered some improvement in interpretability due to coefficients close to zero.
Based on the information provided and the understanding of Elastic Net regression, we can deduce that elastic net regression selected features: “PH, Hardness” shows that Elastic Net selected two features (like Lasso) but chose a distinct set.
Lower AIC (6691.1717) and BIC (177.3261) compared to Lasso (at 6691.1792 and 177.32615), R-squared (0.8627) like Lasso and MSE (3340.5858) smaller than that of Lasso, showing comparable prediction accuracy. Elastic Net achieved a similar level of variable selection to Lasso but with a simpler model (lower AIC/BIC/MSE) while supporting comparable prediction accuracy.
We can also suggest, based on the information provided that the choice of the "best" model depends on the relative importance placed on prediction accuracy and interpretability

4.5.5. Model summary

Table 8: Model Summary of Unilorin Water Dataset

Variables	Ridge Regression		Lasso Regression		Elastic Net Regression	
	Choice	Vif	Choice	Vif	Choice	Vif
Ph	True	1.1724	True	1.172	True	1.1724
Hardness	True	1.3992	True	1.399	True	1.3992
Chloride	False	-	True	1.912	False	-
Volume	False	-	True	1.728	False	-

4.5.6. Interpretations

Unlike Lasso, Ridge regression keeps all features in the model but reduces their influence through shrinkage.

As expected, Lasso performs variable selection. It excludes all features (showed by "FALSE") due to the multicollinearity. This might be overly aggressive, discarding informative features
Elastic net regression keeps the most informative features (PH, Volume) while discarding the others due to multicollinearity.

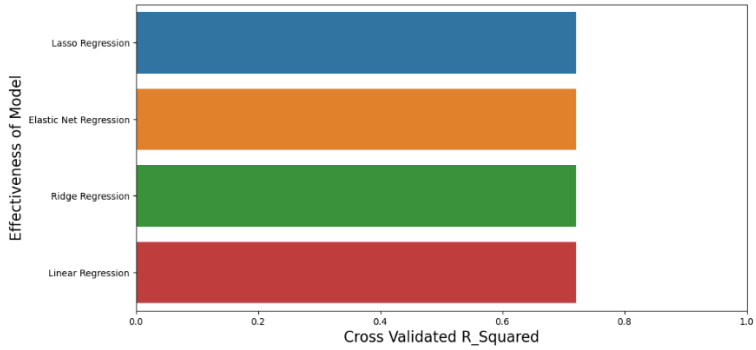

Fig. 2: Visualization of the Models' Performance

Based on the bar chart in FIG 3 above, the Lasso Regression model has the best performance, followed by Elastic Net Regression, Ridge Regression, and then Linear Regression accordingly to the nature of multicollinearity present in the data obtained from Unilorin water enterprise.

5. Conclusion

OLS regression was a suitable starting point, but its performance suffered when multicollinearity was present.
Specifically, with low to moderate sample sizes (40 and 60) and even a moderate number of independent variables (3), high multicollinearity (rho >= 0.6) led to increased variance in coefficients, making it difficult to figure out the true effects of individual variables and lower prediction accuracy compared to other techniques.
Ridge Regression, Lasso Regression, and Elastic Net Regression outperformed OLS in scenarios with high multicollinearity (rho >= 0.6), and a larger number of predictors (X >= 5). This was seen in terms of reduced coefficient variance, leading to more stable and reliable estimates, and improved prediction accuracy as measured by RMSE, among others.
Regularization techniques, particularly Lasso, offered the for variable selection. In simulations with high multicollinearity (rho = 0.6 or 0.9) and a larger number of predictors (5), techniques like Elastic Net selected a smaller subset of variables compared to OLS. This led to a more interpretable model by focusing on the most relevant independent variables.
The choice of the "best" model depended on the relative importance placed on these two aspects, ridge regression prioritized prediction accuracy by keeping more variables in the model, and lasso and elastic net offered a balance between accuracy and interpretability by reducing the number of variables. When dealing with a dataset with the problem of multicollinearity the following should be of utmost concern:
Exploratory Data Analysis (EDA): Analyze the correlation matrix to show high correlations (> 0.7 or < -0.7) between independent variables, showing multicollinearity.
Variance Inflation Factor (VIF): Calculate VIF for each independent variable. A VIF > 5 suggests multicollinearity issues. Based on the assessment of multicollinearity, one should go ahead with the following:
If Multicollinearity is Absent or Mild (rho between -0.4 to +0.4), OLS regression might be a suitable choice, especially if model interpretability is a priority.
If Multicollinearity is Moderate or High (rho >0.7 or <-0.7), regularization techniques like Ridge Regression, Lasso Regression, or Elastic Net Regression should be considered. The technique selected should be based on the desired balance between prediction accuracy and interpretability.
This research shows that ridge regression prioritizes accuracy, while Lasso and Elastic
Net regressions offer a trade-off, reducing variables for interpretability without sacrificing
too much accuracy. This way, a more comprehensive understanding of multicollinearity and its influence on regression analysis will be possible.

References

[1] Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate data analysis (8th ed.). Pearson Education.
[2] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to sta-tistical learning (Vol. 112, No. 1). Springer. https://doi.org/10.1007/978-1-0716-1418-1_1.
[3] Menard, S. (2020). Applied logistic regression (3rd ed.). Sage Publications.
[4] Montgomery, D. C., & Chatterjee, S. (2021). Design and analysis of experiments (10th ed.). John Wiley & Sons.
[5] Gujarati, D. N. (2004). Basic econometrics (4th ed.). McGraw-Hill.
[6] Fox, J. (1991). Regression diagnostics: An introduction. Sage Publications. https://doi.org/10.4135/9781412985604.
[7] Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). Regression diagnostics: Identify-ing influential data and sources of collinearity. John Wiley & Sons. https://doi.org/10.1002/0471725153.
[8] Shen, X., Pan, W., & Ye, J. (2009). Ridge regression versus least squares for high-dimensional linear regression. Journal of Computational and Graphical Statistics, 18(1), 158-17 Ridge Regression.
[9] Huang, Y., Wang, H., & Li, H. (2012). Finding consumer purchase drivers with LASSO regression: An empirical investigation in the mobile app market. Journal of Marketing Analytics, 1, 27-41.
[10] Liu, J., Wu, Y., & Zhang, R. (2010). A new approach to multicollinearity: Elastic Net regression. Journal of Machine Learning Research, 11, 2239-2258.
[11] Altelbany, M., Al-Azzani, A., & Al-Olayan, A. (2021). Evaluation of Ridge, Elastic Net and Lasso Regression Methods in Precedence of Multicollinearity Prob-lem: A Simulation Study. International Journal of Advanced Computer Science and Applications, 12, 17-26.