JEAS, Vol. 31 No. 145 (2025) pp (143-162)
P-ISSN 2518-5764 - E-ISSN 2227-703X
DOI :https://doi.org/10.33095/s8543c81

Journal of Economics and Administrative Sciences

Comparison Between the Partial Least Squares Method and Principal Components Using Genetic Algorithm with an Application

Hayder Osman Hussein* **Rabab Abdul-Ridha Saleh**
Department of Statistics, College of Administration and Economics,
University of Baghdad, Iraq.
***Corresponding author**

Received: 30/8/2024 **Accepted:17/11/2024** **Published: 1 /2/ 2025**

Abstract:

The problem of multicollinearity among independent variables in a regression model was addressed in this study using Partial Least Squares (PLS) and Principal Component Analysis (PCA). The influence exerted by multicollinearity can deteriorate the results of regression modeling; that is, it makes the traditional technique OLS less reliable. The research is centered on applying advanced algorithms for Partial Least Squares (SIMPL and O-PLS) as well as for PCA (NIPALS and SVD) on real-life data scenarios, as well as integrating genetic algorithm (GAs) with these algorithms to optimize predictive performance.

The relative efficiency of these methods is evaluated primarily through the amplitude of the Mean Square Error (MSE) used as a criterion for comparison. The results show the effectiveness of PLS-OPLS above PCA in terms of the lowest before and after embedding genetic algorithms into the MSE. All this underpins the effectiveness of PLS in minimizing multicollinearity thereby allowing for the formulation and prediction of very highly predictive models.

More and more indication of subtle class was revealed with regard to fine tuning advanced GA technique in favor of enhancing regression modeling for complex data analysis. The ongoing research will help in opening numerous possibilities to reinforce regression methodology, especially when one factor in an array of applications representing a relatively significant level of prediction accuracy.

Keywords: Multicollinearity, Partial least squares (PLS), Principal Component Analysis (PCA), Genetic Algorithm (GA), multiple linear regression.

1. Introduction:

Regression is defined as one of the statistical methods used to analyze the relationship between an explanatory variable or several explanatory variables and the response variable, and to predict the value of the response variable based on the explanatory variables after finding the estimated regression equation (Al-Bayati, 2012). In general, regression analysis is used to achieve two goals: the first goal is to search for a mathematical function that helps in knowing how the variables are related to each other, while the second goal is to know the accuracy of the prediction and the strength of the relationship between these variables. Linear models are widely used in various fields, and linear models are among the linear models that are widely used to analyze data in many medical, economic, social and other applied scientific research. When applying regression analysis, researchers often face the problem of multicollinearity, which occurs when there is a strong linear relationship between the explanatory variables, which leads to an increase in the variance of the regression model parameters and makes the results inconsistent. However, the accuracy of ordinary least squares (OLS), and one of its basic assumptions is the independence of the explanatory variables (Al-sabaah, Shorouk Abdul Redha and Al-Quraishi, 2018). However, the problem of multicollinearity between explanatory variables is one of the most common problems facing researchers when using multiple linear regression analysis, and it occurs when the regression model contains many variables and there is a complete relationship between two or more explanatory variables or between all variables(Hassan, Mahmoud Mahdi, 2020). Therefore, to solve this problem and reduce the dimensions of multiple variables, this research uses the Partial Least Squares (PLS) method and Principal Component Analysis (PCA) and employs artificial intelligence algorithms (genetic algorithms) in the methods and compares them to achieve higher accuracy. Therefore, these methods are considered among the most important methods in regression and are used to solve multiple linear issues. This research will be divided into several sections as follows: The first section contains the introduction, while the second section includes a review of previous studies related to the research topic. The third section deals with the methodology and methods used in this research. The fourth section deals with the applied results, while the fifth section is devoted to discussing these results. Finally, the sixth section includes the most important conclusions that were reached.

2. Literature review and Hypothesis Development

Several studies have been conducted that addressed these methods, including:

In 2016, where (Al-Badrani,2016) presented a study comparing the principal component regression and partial least squares regression methods applied to Kirkuk Cement Plant. It was concluded that the partial least squares regression method (PLS) succeeded in achieving an ideal regression model for all dependent variables, in addition to its ability to predict future values for all dependent variables.

In 2018 by the researcher (Al-Bayati,2018) titled Comparison Between Partial Least Squares Method and Singular Value Decomposition Algorithm for Estimating Parameters of the Logistic Regression Model in the Presence of Multicollinearity Problem Using Simulation. In this research, a simulation approach was used to compare the estimation methods through the mean squared error of the model. The comparison reveals that the Singular Value Decomposition algorithm is superior in estimating the parameters of the logistic regression model when there is a problem of multicollinearity.

In 2021, (Alkhafaji & Saleh, 2021) presented a study on the statistical analysis of skewed normal distribution variables using a genetic algorithm based on simulation methods. The problem was solved using the genetic algorithm (GA) and other iterative techniques, including Newton-Raphson, Nelder-Mead, and the iterative reweighting algorithm.

The study concluded that the capabilities of the genetic algorithm using the genetic algorithm for skewed normal distribution parameters are at their best when sample sizes are small or medium, while the best iterative reweighting algorithm was found for large sample sizes.

In 2021, (C. Liu et al., 2022) researched on partial least squares regression and principal component analysis: Similarities and differences between two common methods for variable reduction. This research aims to reduce the number of variables and improve the performance of statistical analyses of the variables under study, especially when some variables are highly correlated. The results showed that partial least squares (PLS) is a better alternative compared to principal component analysis.

In 2022, researchers (Samosir et al., 2022) compared partial least squares regression and principal component regression to overcome multicollinearity in the development index model. This research aims to compare these methods in terms of the modeling factors that affected the Human Development Index (HDI) in 2019. The results indicated that PLS performs better than PCR in terms of coefficient of determination and squared error.

In the same year, a researcher (Kaneko, 2022) addressed research on Partial Least Squares Regression based on a genetic algorithm with only the first component to interpret the model. This research focuses on using the regression coefficients from X to Y for the PLSFC model. Additionally, a set of X can be selected to build a predictive PLSFC model using the Genetic Algorithm (GA), referred to as GA-PLSFC. The two methods, PLSRFC and the proposed GA-PLSFC are compared using simulations, and it was found that the proposed GA-PLSCF method is capable of building highly predictive models.

In this research, the problem of multicollinearity and dimensionality reduction was addressed using Partial Least Squares (PLS) method with different algorithms such as (SIMPLS) algorithm, (O-PLS) algorithm, and the Principal Components Approach (PCA) with (SVD) algorithm and (NIPALS). A comparison was made between the two methods based on the first component, using the mean square error (MSE) as a comparison criterion. We also used the genetic algorithm in both methods to improve the results and determine the method with the highest predictive ability. The results were generalized within the framework of this research.

3. Research methodology:

The descriptive approach was adopted in this research, as it covered the theoretical aspect related to various research concepts, in addition to the applied approach based on real data from hospitalized patients at one of the Ministry of Health hospitals in Dhi-Qar Governorate, specifically Al-Rifai Educational Hospital in 2024.

This paper discusses the methods used to address the multicollinearity problem, with emphasis on the partial least squares method and the principal component analysis method. An alternative method for addressing this problem, namely the ridge regression method, can also be suggested.

3.1 Diagnosing the problem of multicollinearity:

The problem of multicollinearity is defined as a high degree of linear correlation between two or more explanatory variables in a multiple regression model. The reason for its occurrence may be that some explanatory variables may measure the same concepts. In mathematical terms, this problem occurs when the information matrix (X'X) is not fully ranked, and thus its determinant will be equal to zero, which means that it has no inverse, and thus this leads to the failure of the usual least squares method in finding an optimal regression model that represents the data of the phenomenon (Al-Badrani, 2016).

3.2 Regression of partial least squares:

This method is based on the variance and covariance matrix between the explanatory variables and the response variable. It can identify factors that are linear combinations of explanatory variables (X).

These factors are called the latent variables and they in turn give the best model for the response variable (Y) (Al-Safawi,2010). Partial Least Squares (PLS) regression is a powerful technique used in statistical modeling, especially when researchers work with many variables with a small number of observations to extract the most important factors that allow predicting one or more factors (Van Roon et al., 2014) This technique seeks to understand the relationship between the explanatory variables and the response variable. Its characteristics were generalized through principal component analysis (PCA) and multiple regression in the presence of a response variable. Moreover, when there are several response variables, the PLS method is effective when we need to predict many response variables using a large set of explanatory variables (Herv´ & Abdi, 2010). Partial least squares are a multiple linear regression tool developed to relate multiple regressions of one or more response variables to latent variables when the explanatory variables are highly correlated and when the number of these variables exceeds the number of observations.

There are many algorithms related to this method, all of which depend on two basic steps. The first step is to find the latent variables between (X) and (Y) by maximizing the variance-covariance matrix, and the second step is to regress (Y) on the Components (t) (Abass, 2020). Among these algorithms that are used to solve the problem of dimensionality reduction and thus get rid of the problem of multicollinearity, the (NIPALS (PLS1, PLS2)) algorithm attributed to (Wold), the (Statistically Inspired Modification of Partial Least Squares (SIMPLS)) algorithm attributed to (De Jong), the (KERNAL) algorithm attributed to (Dayal), the (PLS-F) algorithm attributed to (Manne), the O-PLS algorithm attributed to (Trygg and Wold), and other algorithms (Saleh, n.d.) . In this research, we relied on the SIMPLE A algorithm as well as the Orthogonal-PLS algorithm. Assuming that the matrix $\mathrm{X}_{\mathrm{n,p}} = (\mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_\mathrm{n})'$ where $\mathrm{X}_\mathrm{i} = (\mathrm{X}_{\mathrm{i}1}, \mathrm{X}_{\mathrm{i}2}, \dots, \mathrm{X}_{\mathrm{iq}})$ of observations. And the matrix Y consists of n observations and q response variables, and we denote it with the symbol $\mathrm{q}_{\mathrm{n,p}} = (\mathrm{Y}_1, \mathrm{Y}_2, \dots, \mathrm{Y}_\mathrm{n})'$ and the combined data set

$(\mathrm{X}_{\mathrm{n,p}}, \mathrm{q}_{\mathrm{n,p}})$ denotes it with the symbol $\mathrm{Z}_{\mathrm{n,m}}$ where m=p+q and the linear regression model(Hubert & Branden, 2003):

$$\mathrm{y}_\mathrm{i} = \beta_0 + \beta'_{\mathrm{p,q}}\mathrm{x}_\mathrm{i} + \mathrm{e}_\mathrm{i} \qquad \dots (1)$$

Where:

ei: represents the error term that requires: $\mathrm{cov}(\mathrm{e}_\mathrm{i}) = \sum \mathrm{e}_\mathrm{i}$, $\mathrm{E}(\mathrm{e}_\mathrm{i}) = 0$ for size q

$\beta_{0\mathrm{q}}$, ...,$(\beta_{01} = \beta_0)$ represents the constant term with dimension q , β_{pq}: represents the unknown parameters and is the slope matrix with dimension pxq.

3.2.1 The statistically inspired modification of the partial least squares (SIMPLS) algorithm.

The SIMPLS algorithm assumes that the variables x and y have a relationship through the binary model:(Hussein, Suja Muhammad and Saleh, 2014)

$$x_i = \bar{x} + P_{P,K}\,\tilde{t}_i + g_i \qquad \dots (2)$$

$$y_i = \bar{y} + A'_{q,k}\,\tilde{t}_i + f_i \qquad \dots (3)$$

$\bar{x}, \bar{y}$: mean of variables x,y.

$\tilde{t}_i$: Measurements or scores with a dimension k where k ≤p

$\mathrm{p}_{\mathrm{p,k}}$: Loading Matrix (x-loading)

$\mathrm{A'}_{\mathrm{k,p}}$: The slope matrix in regression $\mathrm{y}_\mathrm{i} on\ \tilde{t}_i$

f_i , g_i : Residues.

The algorithm first assumes the formation or construction of components where h is obtained from the components that are linear combinations of x from the variables that have the greatest shared variance with the linear combinations of the variables y. More precisely, we assume that $(\tilde{Y}_{n,q}, \tilde{X}_{n,P})$refers to the centered data matrix, as:

$$\tilde{x} = x_i - \bar{x} \qquad \dots (4)$$

$$\tilde{y} = y_i - \bar{y} \qquad \dots (5)$$

The natural weight vectors (PLS) where $(\|r_a\| = \|q_a\|) = 1$are defined as the vectors that maximize for each a where a=1…h.

$$cov\left(\tilde{Y}_{n,q}\, q_a\, ,, \tilde{X}_{n,P}\, r_a\right) = q'_a \frac{\tilde{Y}'_{q,n}\, \tilde{X}_{n,P}}{n-1}\, r_a = q'_a\, S_{x,y}\, r_a \qquad \dots (6)$$

This algorithm starts $S_{xy}^{1} = S_{xy}$, which is the covariance and shared variance matrix. The which are placed to (r_a,q_a)maximization of the equation is obtained from the first pair of vectors the left and right of the matrix S_{xy}. The vector r_a is the latent variable of S_{xy}* S_{yx} with dimensions (P x P), and the vector q_a is the eigen vector of S_{yx}*S_{xy} with dimensions (q x q)

The elements of $\tilde{t}_i$, which are the linear combinations of the centered data, are obtained from the following formula:

$$\tilde{t}_i = \tilde{x}_i' \; r_a \qquad \text{... (7)}$$

$$\tilde{T}_{n,h} = \tilde{X}_{n,p} \; R_{p,h} \qquad \text{... (8)}$$

$R_{p,h}$=(r_1….r_h)

To obtain more than one solution, it is required that the components $\tilde{T}_{n,h}$ be orthogonal

$T_a' T_j = 0 \;\; \forall j \neq a$

The above constraint imposes the generation of a series of different solutions for the equation (6) after obtaining the first orthogonal component from the equation, through which we can avoid multicollinearity among the explanatory variables in the second step of the algorithm. With the above constraint, the weight vectors for the (SIMPLS) algorithm, namely r_a and q_a ($2 \leq a \leq h$), are obtained using the latent variable of $S_{yx}^{a} S_{xy}^{a}$ and $S_{xy}^{a} S_{yx}^{a}$. Thus, the covariance and shared variance matrix is obtained from the following equation:

$$S_{x,y}^{a} = \left(I_p - V_{a-1} V_{a-1}'\right) S_{x,y}^{a-1} \qquad \text{... (9)}$$

Or

$$S_{x,y}^{a} = S_{x,y}^{a-1} - P_{a-1}(P_{a-1}' \; P_{a-1})^{-1} P_{a-1}' \; S_{x,y}^{a-1} \qquad \text{... (10)}$$

And (V1, ..., Va-1) are represented as orthogonal to the loading matrix X (P1, ..., P_{a-1}), where:

$$P_j = \tilde{X}' T_j / \; T_j' T_j \qquad \text{... (11)}$$

And by substituting T_j, we obtain the following

$$P_j = \frac{\tilde{X}' \tilde{X} r_j}{r_j' \tilde{X}' \; \tilde{X} \; r_j} = \frac{S_x r_j}{(r_j' \; S_x r_j)} \qquad \text{... (12)}$$

which represents the least squares regression coefficient in regressing $\tilde{X}$ on the component T_j.

Where:

S_x is the covariance matrix of the explanatory variables.

The second step in the algorithm is to perform multiple linear regression (MLR), for regressing the extracted components T_1...T_h on the original variables y, and the form of the regression model is as follows:

$$y_i = \alpha_0 + A_{q,h}' \tilde{t}_i + f_i \qquad \text{... (13)}$$

So

$E(fi) = 0$ and $cov\;(fi\;) = \sum f$

Estimation of Multiple Linear Regression (MLR)

$$\hat{A}_{h,q} = (S_t)^{-1} S_{ty} - (R_{h,p}' \; S_x \; R_{p,h})^{-1} R_{h,p}' \; S_{xy} \qquad \text{... (14)}$$

$$\hat{\alpha}_0 = \bar{y} - \hat{A}_{q,h}' \; \bar{\tilde{t}} \qquad \text{... (15)}$$

$$S_f = S_y - \hat{A}_{q,h}' \; S_t \hat{A}_{h,q} \qquad \text{... (16)}$$

S_t and S_y :initial covariance matrix of the variables t and y

We note that multiple linear regression refers to the classic least squares regression of multiple X variables, and when the number of dependent variables is greater than one for the y variables, it is known as multiple linear regression. Due to $\tilde{t}_i = 0$, the constant term $\alpha_0$is estimated by $\bar{y}$.

When substituting $\tilde{t}_i = R_{h,p}'(x_i - \bar{x})$ in equation (3), we obtain the estimators of the parameters. For the original regression model, as follows:

$$\hat{\beta} = R_{p,h}(R_{h,p}' \; S_x \; R_{p,h})^{-1} R_{h,p}' \; S_{xy} \qquad \text{... (17)}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_{q,p}' \; \bar{x} \qquad \text{... (18)}$$

And finally, the estimation of S_e is
$S_e = S_y - \hat{\beta}\, S_x\, \hat{\beta}'$ ... (19)
In the case of a single response (q=1), the estimation of the parameters is $\hat{\beta}_{p,1}$and is expressed as a vector, while the variance of the error S_e is calculated as follows $\widehat{\sigma}_e^2 = S_e^2$

3.2.2 Orthogonal projections to latent structures (O-PLS) Algorithm

Trygg and Wold (2002) modified the original NIPALS algorithm to a systematic variance from the variables in a dataset that are not mutually related to the response variable, a method referred to as Orthogonal Projections to Latent (OPLS). Assuming a linear correlation between one output and multiple in a dataset, it is always possible to find a PLS (Partial Least Squares) model with a single component that captures the linear relationship after appropriately processing the data. Researchers Viroon et al. (2004) demonstrated that the number of PLS components in a PLS model processed with the OPLS algorithm can be reduced to a single component without affecting predictive performance.
The data matrix $X \in R^{mxn}$ consists of (m) rows of observations and (n) columns of explanatory variables, along with the response vector (y). Each variable is standardized to have a mean of zero and a standard deviation of one(J. Liu & Wong, 2011).
Orthogonal compounds (T-ortho) are not associated with the response variable y as: -
$t'_{ortho}\, y = w'_{ortho}\, X' y = [p' - (w'p)w']X'y$ ... (20)
Since this means that the above equation can be written as follows: -$X'y = W\,(y'y)$
$t'_{ortho}\, y = [p'w - (w'p)w'w]\, y'y = 0$
so that w'w=1 and w' p=p'w
Thus, the removed compounds are perpendicular to the response vector y. When new entries are available, each orthogonal compound is extracted in a repeated manner as follows: -
$t_{new_{ortho}} = X'_{new}\, w_{ortho}\,,\ X'_{new} = X'_{new} - t_{new_{ortho}}\, p'_{ortho}$ ... (21)
After removing all orthogonal compounds, the response value can be predicted using $\hat{y} = X'_{new}\, b$
The OPLS algorithm is illustrated by the following steps: -
1- $w' = y' X/(y'y)\ ,\ w = w/\|w\|$
2- $t = Xw\,,\ c = t' y/(t't),\ u = y\,c/c^2$
3- $p' = t'X/(t't)$
4- $w_{ortho} = P - (w'P)w\,,\ w_{ortho} = w_{ortho}/\|w_{ortho}\|$
5- $t_{ortho} = Xw_{ortho}\ , p'_{ortho} = t_{ortho}\, X/(t'_{ortho}\, t_{ortho})\ ,$
$E_{OPLS} = X - t_{ortho}\, p'_{ortho}$
6- Save the information obtained $T_{ortho} = [T_{ortho} \quad t_{ortho}]$
$,\ \ P_{ortho} = [P_{ortho} \quad p_{ortho}]\ ,\ W_{ortho} = [W_{ortho} \quad w_{ortho}]$
For additional orthogonal compounds we can refer to step two.
7- Perform steps (1-3) after extracting orthogonal compounds.
8. $q = y'\, u/u'\, u\ \ ,\ b = qw$

3.3 Principal Components Regression Method:

It is a statistical technique that can be expressed as a linear set of variables that retain the complete information of real data, and we can obtain eigenvalues and latent variables of the correlation matrix $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_m$ R through The following equation (Shang et al., 2017):-
$|R - \lambda I| = 0$ ... (22)

Principal component analysis works to reduce the information represented by variance and store it in components without losing information. The main idea of this method lies in converting linearly related variables that have a large amount of data into orthogonal and independent components that are arranged based on the amount of variance (Al- Rawi, Asmaa Ghaleb and Issa, 2019). This method is a traditional multiple method that depends on the covariance matrix between the explanatory variables and explains the largest part of the variance by finding linear structures independent of each other and each main component is called a linear combination for all variables, the first Principle component gives most of the variance and so on for the rest of the components and there are several algorithms to extract the main components are Jacobi, Mathematics of PCA Singular Value Decomposition (SVD) , NIPALS, Rotation (Saleh, n.d.). The principal component method transforms the original correlated explanatory variables, without removing any of them, into new orthogonal variables known as principal components (Hassan, Mahmoud Mahdi, 2020). In this research, we have addressed the Singular Value Decomposition (SVD) algorithm as well as the NIPALS algorithm.

3.3.1 Principal Components Selection Method:

It is possible to choose the number of significant principal components for summarizing data practically based on the following points:

1. The number of selected principal components should equal the number of eigenvalues greater than one.
2. Retain the number of components that explain 80% of the total variance.
3. Morrison (1976) indicated that explaining 75% of the total variance is sufficient, and the higher the proportion of explained variance with a smaller number of selected principal components, the better it is in terms of ease of discussing and interpreting the results (Mohammed, Haidar Yahya and Mohammed, 2020).

3.3.2 Singular value decomposition algorithm (SVD):

This algorithm is used to find the principal components as it gives the characteristic vector and the distinctive values that we need in the analysis of the principal components, and by analyzing the principal components we get the first principal component (PC) by dividing the nxp array X into three matrices (Ramzan Shahla, 2010).

$$\mathrm{X} = \mathrm{T_0\ S\ P'} \qquad \dots (23)$$

Suppose that $t = min\,\{n, p\}$

Whereas,

T₀: nxt orthogonal matrix which is the component matrix and is found from the characteristic vector of XX'

S: a diagonal matrix of $t \times t$ order which is equal to square roots to the characteristic values of X'X or XX' where $\mathrm{S} = \mathrm{daig}\{\lambda_1 > \lambda_2 \dots..\lambda_\mathrm{p} > 0\}$

P: Perpendicular matrix of Pxt order which is a load matrix and is found from the characteristic vector to XX' where the main components in the analysis are T can be (Saleh, n.d.):-

$$\mathrm{T} = \mathrm{T_0\ S} \qquad \dots (24)$$

$$\mathrm{X} = \mathrm{T_0\ S\ P'} + \varepsilon \qquad \dots (25)$$

As for the regression coefficients, they are: -

$$\beta_{\mathrm{PCR}} = \mathrm{P(T'T)^{-1}\ T'Y} \qquad \dots (26)$$

$$\beta_{\mathrm{PCR}} = \mathrm{P\ S^{-1}\ T_0'\ Y} \qquad \dots (27)$$

3.3.3 NIPALS PCA algorithm:

This algorithm is one of the many existing methods of finding latent variables that were originally developed for PCA but have been used in other ways as well, and this is an overview of the algorithm (Andrecut, 2009):

X=TP' ... (28)

Where T columns are called grade columns and P columns (P rows) are called loads, the algorithm begins to initialize h=1 and Xh=X, and this algorithm is done through the following steps:

1. Choose th as any column of Xh
2. Calculation of loads $\mathrm{P_h} = \mathrm{X_h' t_h / t_h' t_h}$
3. We impose $\mathrm{P_h} = \mathrm{P_h}/\sqrt{\mathrm{P_h' P_h}}$
4. Calculation of grades $\mathrm{t_h} = \mathrm{X_h P_h / P_h' P_h}$

Repeat step (3) and (4) until the convergence of the HTH of the principal component.
Let's assume and (Distinctive Value). $X_{h+`1} = X_h - t_h P'_h \lambda_h = t'_h t_h$. Increase h = h + 1 and repeat to the last main component.
Grouping T columns from ith and P columns from vectors Ph and the result PCs may be scaled in different ways. One way to scale PCA The solution is to specify the loads P=V and T=U'S.

3.4 Genetic algorithm:

Living organisms interact with their surrounding environment and attempt to adapt to changes. If they are unable to adapt, they face extinction. The organisms that possess strong traits survive, while those with weak traits weaken and die. Additionally, mutations, which occur at very low rates, are important factors that contribute to the development of hereditary traits passed down through genes. From this concept, an idea was adapted and applied in the field of computing known as Genetic Algorithm (GA). The genetic algorithm has significantly reduced the effort and time required for software and system designers by providing a general algorithm that can be relied upon to solve various types of problems, rather than developing a specific algorithm for each problem while considering the necessary adjustments to fit the particularities of each issue regarding the size and type of data used, the nature of the objective function, and the constraints of each problem(Abdul Hadi, Anwar Taher and Reda, 2020)***.***

This algorithm is considered as a random search technique based on natural selection and genetics to find the solution to the problem at hand. It was invented and developed by John Holland in 1975. The algorithm begins with a randomly selected set of chromosomes and ends with the best solutions. Applying the algorithm to a problem requires determining the correct encoding for the chromosomes and defining the efficiency of the fitness function(Al-Douri et al., 2020).Determining the correct encoding of chromosomes and determining the efficiency of the fitness function is also required. It is used to solve optimization problems that depend on phenomena in evolutionary biology such as genetics, mutations, natural selection, and hybridization. This algorithm can process many individuals in a population at the same time and evaluate multiple solutions in the search space (Jiang et al., 2023). New chromosomes are produced through crossover or mutation. The most suitable chromosomes are selected for the next generation, while the others are eliminated to maintain a constant population size across successive generations. This algorithm tries to converge towards the fittest chromosome which represents the optimal solution (Kale et al., 2022).

This algorithm operates through several stages to reach the optimal solution. Initially, there is the initialization of chromosomes, which is the process of generating many preliminary results (chromosomes) randomly. Following that, there is the fitness function, which evaluates the population of chromosomes by calculating the value of each chromosome, with the resulting value indicating the chromosome's efficiency. Afterwards, the process of selecting the best chromosomes takes place to form the parents for the next generations based on the fitness function evaluation.

Then comes the crossover or mating stage, which occurs between the chromosomes of the selected parents to produce offspring. The mutation stage follows, which is a natural change in the genes of the chromosome because of one or more gene mutations in the offspring's chromosome. The final stage is the stopping criterion, which continues to search through the generations sequentially to find the best solution, and it tests whether a stopping condition exists or not, according to the nature of the studied problem(Elvira-Ortiz et al., 2020) .

3.4.1 Problems addressed by the genetic algorithm:

We review below the problems that the bio-in-time algorithm can address (Hussain & Nassir, 2015):

1. Some problems require data analysis in large electronic databases, such as financial, astronomical, and other types of data.
2. The ineffectiveness of traditional research methods in solving the problem.

3. The problem of variables being related to a non-linear relationship or a relationship that is not well understood.
4. Problems for which an approximate solution is convincing, for example, image processing.

3.4.2 GA-PLS:

The traditional method of representing a solution to a feature selection problem using the GA-PLS is by a Boolean vector of the same length as the total number of variables in the dataset. Such vectors are generally known as chromosomes. Each Boolean value of vector corresponds to a variable that is either excluded from feature selection (0) or included (1). Many different chromosomes, when combined in a Boolean matrix, form a population. The Boolean values of each solution are typically initialized randomly at the beginning of the algorithm. The set of features defined by each chromosome in the community is used to measure PLS regression models on the dataset. The resulting regression coefficients are used to model the target response of the dataset, and the error associated with model predictions is reduced to a statistical measure of performance of the feature set, such as the Mean Squared Error (MSE). The cross-validation performance measure for each chromosome is converted into a fitness value, which is a numerical value representing the quality of the solution encoded for genetic evolution and determines what is meant by optimization(Eiben, A.E & Smaith, 2015).

The efficiency of a chromosome to the fitness value, i.e., its ability to solve the optimization task compared to all other candidate solutions in the community, determines how it is treated by the GA-PLS in the subsequent steps, where fitter individuals are favored compared to those with lower fitness. The bias within the genetic evolution cycle towards favoring high-fitness solutions is typically achieved using two stages of selection: parent selection and survivor selection. In the parent selection stage, it is decided which individuals will be used to generate new solutions that contribute to driving the evolutionary process forward. In the survivor selection stage, all individuals in the community compete for survival and transition to the next generation. In addition to favoring individuals known to have high fitness through selection factors, genetic algorithms have the capability to generate improved potential new solutions for the optimization task by combining and enhancing chromosomes that are already successful to form new chromosomes. This is achieved using a set of random variation factors: crossover and mutation.

Crossover takes a pair of parent chromosomes as input and produces one or several output chromosomes, similar to how genes are combined from parents to generate offspring. A frequently used crossover example in GA-PLS is the two-point crossover, which is performed by selecting two points along two original chromosomes and the resulting chromosomes to inherit sections from their parents alternately between the crossover points. The second variation factor, mutation, is a unary process that takes a single chromosome as input and outputs a slightly modified version the same chromosome, which replaces the original chromosome in the community (Leardi, 2003). The following figure shows a flowchart of GA-PLS modeling.

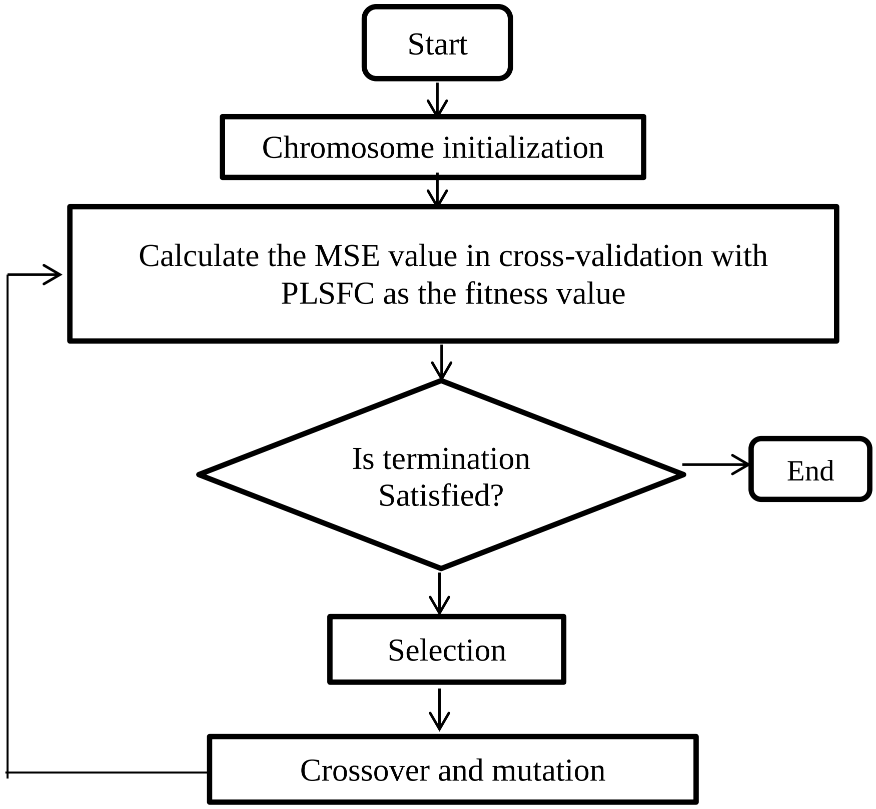
Start
Chromosome initialization
Calculate the MSE value in cross-validation with PLSFC as the fitness value
Is termination Satisfied?
End
Selection
Crossover and mutation

Figure 1: The flowchart of GA-PLSFC modeling.
Source: (Kaneko, 2022)

3.4.3 GA-PCA

This process can be divided into two steps. The first step involves PCA (Principal Component Analysis) transforming the original variables into principal components, and the second step involves the feature selection strategy based on GA (Genetic Algorithm) selecting the appropriate principal components and using them as new variables for the predictive model. The goal is to achieve better predictive capability in terms of performance on the original data. Genetic algorithms are optimization algorithms inspired by the processes of natural selection and genetics. They include creating a population of potential solutions (chromosomes), evaluating their fitness based on a specified objective function, selecting the fittest individuals for reproduction, and applying genetic operations such as crossover and mutation to create new offspring.

In the context of using GA-PCA, the objective is to utilize GA to search for a distinctive set of features that enhances the model's performance. This is achieved by encoding the feature selection problem as a chromosomal representation in GA. Initially, a population of chromosomes is created, with each chromosome being evaluated using a fitness function that combines the performance of the selected features with the data transformed by PCA. GA improves the population through selection, crossover, and mutation operations to produce new generations.

The selection process favors chromosomes with higher fitness values, ensuring the retention of better feature sets. PCA is applied to the selected feature sets to reduce the dimensionality of the data while retaining the most important information.

This helps in enhancing the model's performance. The GA-PCA process is iterated over several generations until stopping criteria are met, such as reaching a maximum number of iterations or achieving an acceptable level of performance (Ding et al., 2014). The following figure shows a flowchart of GA-PCR modeling.

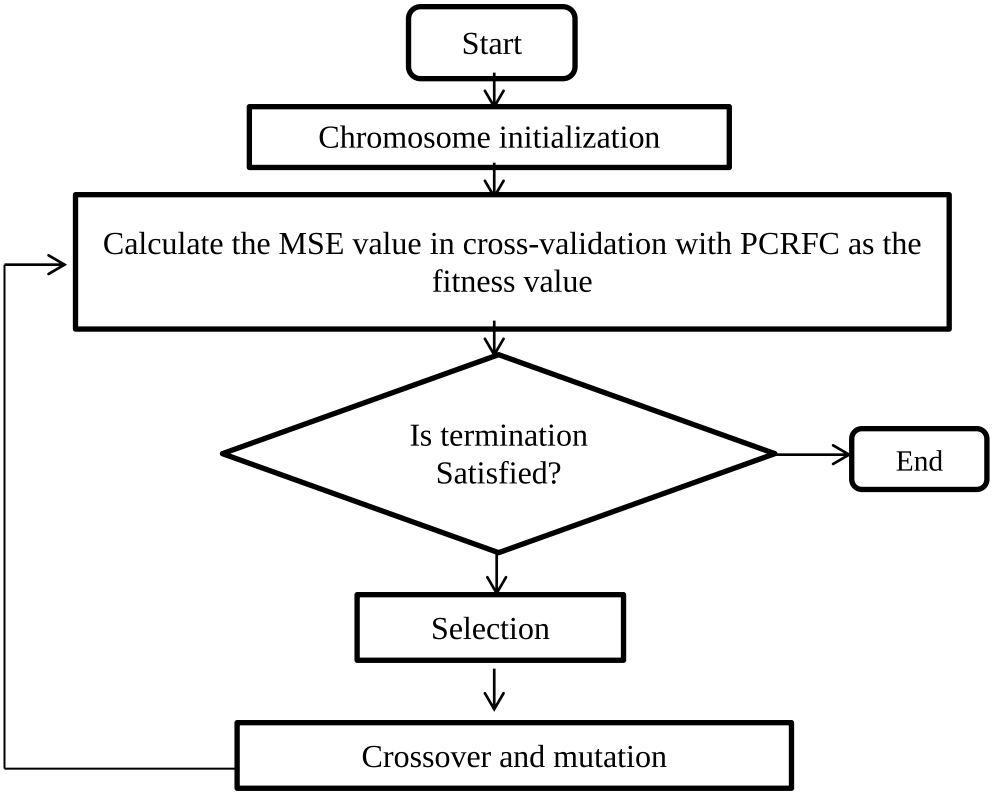
Start
Chromosome initialization
Calculate the MSE value in cross-validation with PCRFC as the fitness value
Is termination Satisfied?
End
Selection
Crossover and mutation

Figure 2: The flowchart of GA-PCRFC modeling.

4. Results:

This aspect aims to describe data related to hypertension by identifying the most significant factors (explanatory variables) that have a substantial impact on this disease. To achieve the research objective of finding the best method between PLS (Partial Least Squares) and PCR (Principal Component Regression) in building a regression model, one of the artificial intelligence algorithms, namely the genetic algorithm, will be employed in these methods.

4.1 Sample Description:

A sample of 60 patients diagnosed with hypertension was collected from the patients hospitalized at Al-Rifai General Hospital. The research included 8 variables; we denoted the response variable as y and the explanatory variables as X_i $(i = 1,2, \ldots ,7)$. These are the common variables collected and analyzed in hypertension studies, helping to understand the factors affecting this disease and to identify the best strategies for its prevention and treatment. The table below illustrates the variables in order:

- Search variables: -

The response variable y_i represents hypertension, which is known to be divided two categories: the syst component and the diastolic component. This research, data collection was based on the systolic.

Table 1: The code of each of the illustrative variables

Variable	Variable name
X_1	Sex
X_2	Lifetime
X_3	Weight
X_4	Heartbeat
X_5	Blood sugar rate
X_6	Hemoglobin in the blood
X_7	Cholesterol

Source: Prepared by the researchers

4.2 Detecting multicollinearity:

There are several tests through which we can detect the existence of the problem of linear multiplication between the explanatory variables of people with blood pressure disease using the SPSS program, and the test that was used to detect this problem is: -

4.2.1 Variance Inflation Factor (VIF):

This scale is one of the most important measures that are used to detect the problem of Multicollinearity as it can be measured through the value of VIF coefficients, if the value is greater than 10, this indicates the existence of the problem of Multicollinearity, as shown in the table (2) below: -

Table 2: The VIF values for each variable

Variables	X_1	X_2	X_3	X_4	X_5	X_6	X_7
VIF	12.09616	10.06426	5.47845	3.87880	1.03811	2.47612	12.19626

Source: Prepared by the researchers

We note through table (2) that the values of the VIF for the illustrative variables and for each of the first variable, which is gender, is equal to 12.09616, the second variable is age, with a value of 10.06426, and the seventh variable, the percentage of cholesterol, its value is equal to 12.19626, and these variables have values greater than 10, this means that they suffer from the problem of Multicollinearity multiplicity between illustrative variables.

4.3 Applied results:

This section presents the applied results and subsequently analyzes them to determine the suitability and accuracy of the actual data in addressing the multicollinearity issue, based on two methods: Partial Least Squares (PLS) and Principal Component Regression (PCR).

To compare the two methods to select the one according to the comparison criterion of Mean Squared Error (MSE), the genetic algorithm was also employed in the methods, and the comparison was based on the first component. The applied results obtained through the R program are presented as follows.

Table 3: The values of the MSE comparison standard between the PLS and PCA methods

Methods	PCR. SVD	PCR. NIPALS	PLS. SIMPLS	PLS. OPLS
MSE	203.40609	198.25367	189.09135	180.10590

Source: Prepared by the researchers

We note from table (3) that the values of the MSE comparison criterion for the OPLS algorithm are the lowest and therefore the PLS method is the best. Its advantage is due to its ability to deal with high-dimensional data and build models based on improving prediction.

Table 4: The MSE of PLS and PCA methods when employing the genetic algorithm in the PLS.GA and PCA.GA methods

Methods	PCR.SVD.GA	PCR.NIPALS.GA	PLS.SIMPLS.GA	PLS.OPLS.GA
MSE	178.14174	175.81513	175.26477	171.75359

Source: Prepared by the researchers

As shown in table (4), when employing the genetic algorithm in the methods, we also find that the MSE value for the OPLS algorithm is the, indicating that the PLS method is the best.

Based on the results presented in tables (3 and 4), it is evident that the comparison criterion, MSE for the Partial Least Squares (PLS) method is lower than that for the Principal Component Analysis (PCA). This means that the PLS method is the best and most optimal choice. It can be stated that this method is considered the most suitable option for data analysis, especially in cases where multicollinearity issues need to be addressed, as well as for predicting future values or dealing with complex data.

Additionally, it serves as an effective tool for use in various fields. Since this method is the best, this implies that the PLS model is more effective in building accurate predictive models.

4.3.1 Estimation of regression model parameters

The first component was relied upon to obtain the parameter estimates of the model using the PLS and PCA methods, as well as in the case of employing the artificial intelligence algorithm (genetic algorithm) in the methods. The tables below illustrate the estimation results for each method.

Table 5: The parameter values of the regression model for the first component

Methods	PCR. SVD	PCR. NIPALS	PLS. SIMPLS	PLS. OPLS
β0	94.92774	65.62464	79.72178	81.78017
β1	-0.51731	2.69115	2.50203	3.52516
β2	0.06962	-0.01959	0.07793	-0.01320
β3	0.05356	0.19680	0.06719	0.13874
β4	0.16268	-0.01043	0.17917	-0.03662
β5	0.05652	0.15476	0.08218	0.14638
β6	0.00580	1.09293	-0.54668	-0.57845
β7	0.10472	0.12510	0.13817	0.15712

Source: Prepared by the researchers

Table 6: The parameter values of the regression model for the first component when employing the genetic algorithm

Methods	PCR.SVD.GA	PCR.NIPALS.GA	PLS.SIMPLS.GA	PLS.OPLS.GA
β0	83.34480	68.69653	75.73453	76.78136
β1	4.00759	5.61416	5.57559	6.02722
β2	-0.05712	-0.10190	-0.05435	-0.09983
β3	0.17318	0.24741	0.18187	0.21746
β4	-0.14178	-0.23006	-0.13689	-0.24306
β5	0.17817	0.22818	0.19154	0.22348
β6	-0.58777	-0.05844	-0.86274	-0.88933
β7	0.16813	0.17407	0.18674	0.19349

Source: Prepared by the researchers

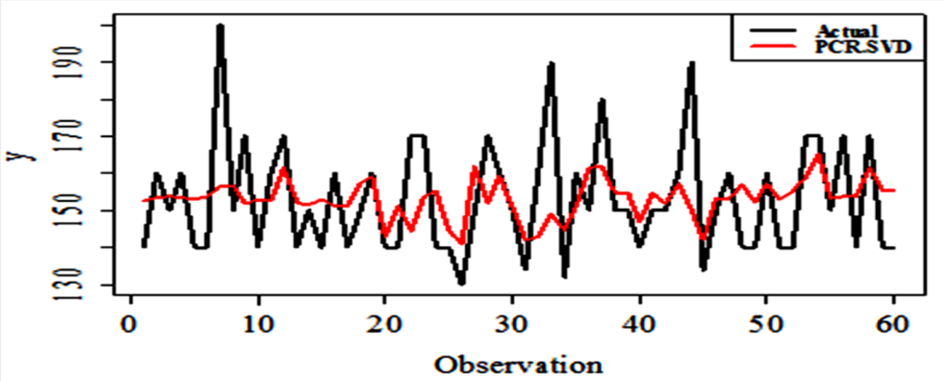

Figure 1: The SVD-PCA algorithm
Source: Prepared by the researchers

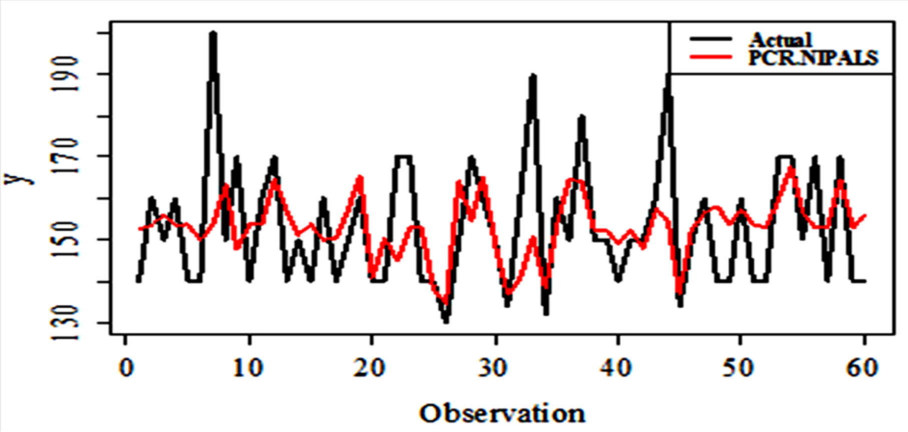

Figure 2: The NIPALS-PCA algorithm
Source: Prepared by the researchers

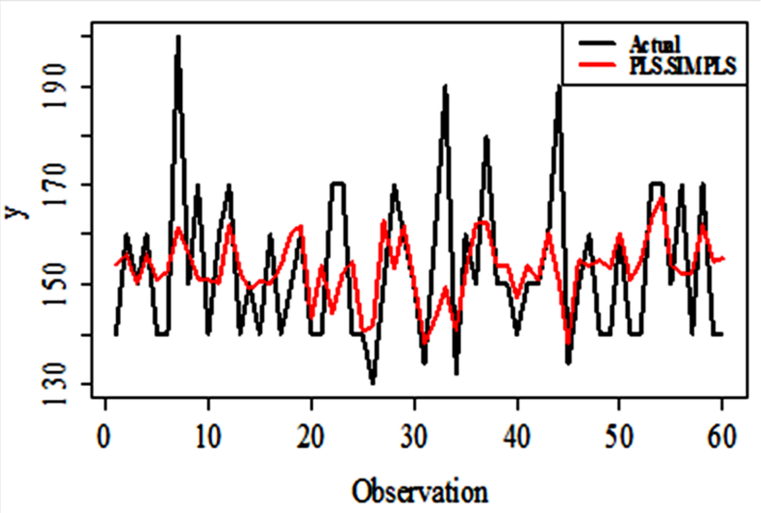

Figure 3: The SIMPLS-PLS algorithm.
Source: Prepared by the researchers

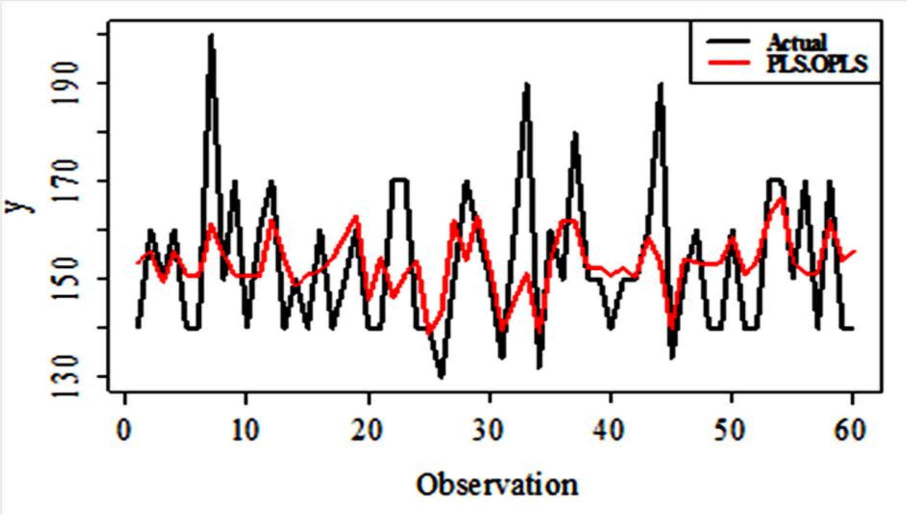

Figure 4: The OPLS-PLS algorithm.
Source: Prepared by the researchers

The shapes above illustrate the graph of original and predicted values, along with the variable for the model using Partial Least Squares (PLS) and Principal Component (PCA). We observe through these that the algorithms that come to the true values with the predicted values, based the lowest Mean Squared Error (MSE), are the (O-PLS) algorithm, followed by the (SIMPLS) algorithm, and then the (NIPALS) algorithm. The last place is held by the (SVD) algorithm. This confirms that the Partial Least Squares (PLS) method possesses an optimal regression model and high predictive capability.

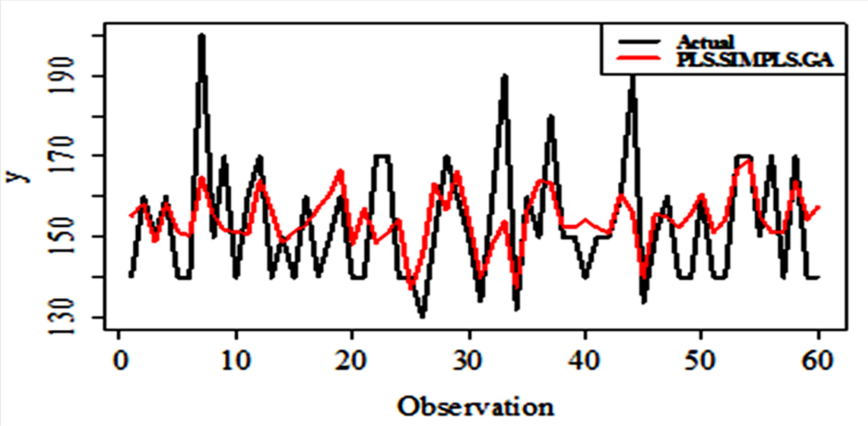

Figure 5: The SVD algorithm. GA-PCA
Source: Prepared by the researchers

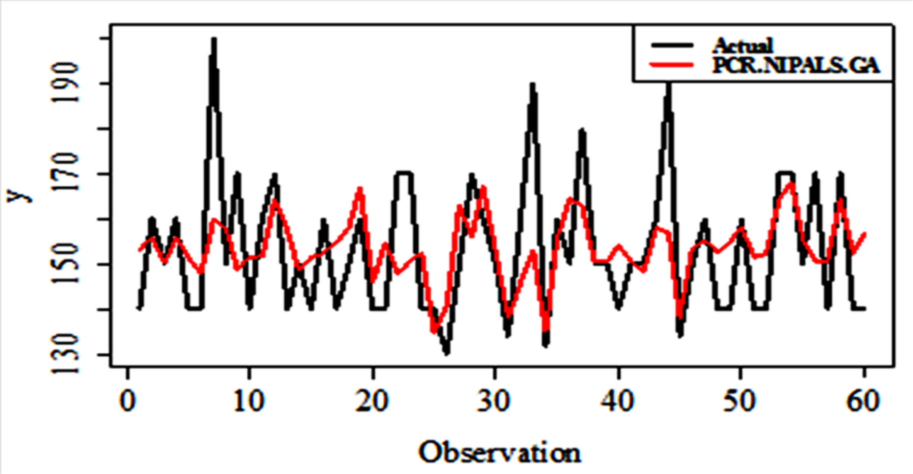

Figure 6: The NIPALS algorithm. GA-PCA.
Source: prepared by the researcher.

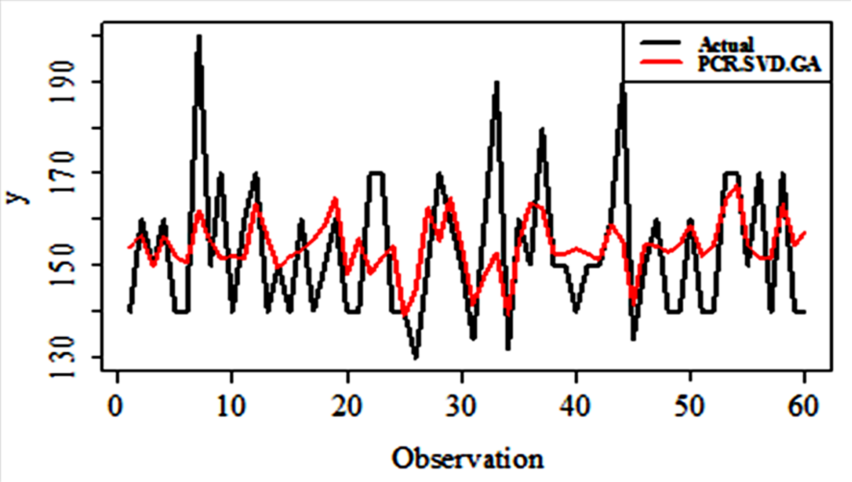

Figure 7: represents the SIMPLS algorithm GA-PLS.
Source: prepared by the researcher.

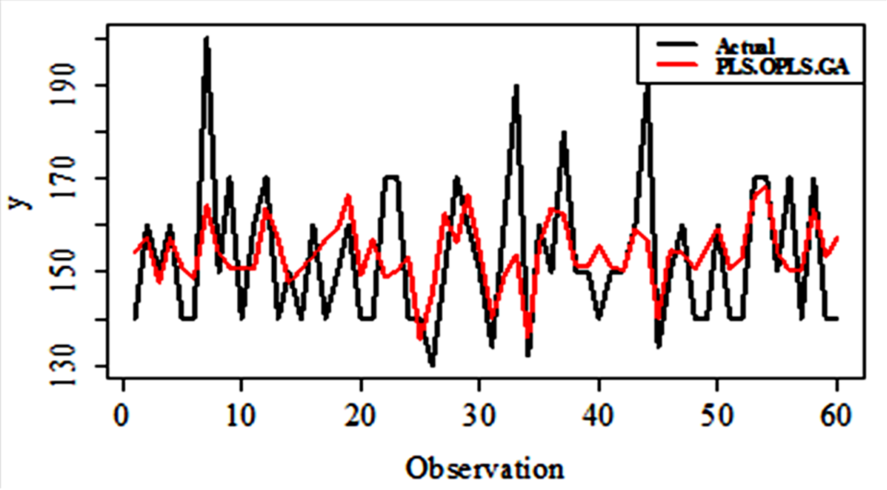

Figure 8: represents the OPLS algorithm GA-PLS.
Source: prepared by the researcher.

The shapes above illustrate the graph of original and predicted values, along with the response variable for the model estimated using Partial Least Squares (PLS) and Principal Component Analysis (PCA) when employing the genetic algorithm for both methods. We observe from these shapes that the algorithms that come closest to the true values with the predicted values, based on the lowest Mean Squared Error (MSE), are the (O-PLS) algorithm, followed by the (SIMPLS) algorithm, then the (NIPALS) algorithm, The last place is held by the (SVD.GA) algorithm. This confirms that the Partial Least Squares (PLS) method possesses an optimal regression model and high predictive capability, distinguishing it as the most efficient method for predicting future values of the response variable.

4.3.2 Analysis of Variance

The analysis of the variance table can be used in regression analysis to measure the impact of explanatory variables on the response variable and to determine whether there are statistically significant effects. In this research, the results of the analysis of the variance table were derived using the best method, which is the PLS method. Below are the results of the analysis of variance.

Table 7: The analysis of variance for the OPLS-PLS algorithm

S.O.V	D.F	S.S	M.S	Fcal	P-Value
Regression	7	3327.191	475.313	2.287	0.04142
Error	52	10806.354	207.814		
Total	59	14133.545	-		

Source: Prepared by the researchers

Table 8: The analysis of variance when employing the genetic algorithm in (OPLS. GA-PLS)

S.O.V	D.F	S.S	M.S	Fcal	P-Value
Regression	7	4226.113	603.730	3.046	0.00921
Error	52	10305.215	198.177		
Total	59	14531.328	-		

Source: Prepared by the researchers

To assess the significance of the linear relationship and test the extent of the impact of the explanatory variables on the response variable, the hypothesis concerning the regression model was tested as follows.

$$H_o: B_1 = B_2 = \ldots\ldots\ldots = B_7$$
$$H_1: B_1 \neq B_2 \neq \ldots\ldots\ldots \neq B_7$$

It was found from the results of table (7) that the null hypothesis is rejected, and the alternative hypothesis is accepted at a significance level of 0.05 after comparing it with the P-Value. This indicates that there are significant differences between the explanatory variables and the response variable similarly, with regard to table (8), the results indicate the rejection of the null hypothesis and the acceptance of the alternative hypothesis, which shows that there are significant differences between the explanatory variables and the response variable.

5. Discussion of results:

The results indicate the superiority of the partial least squares (PLS) method over the principal components (PCA) method in terms of having the lowest value of the comparison criterion (MSE) before and after employing the genetic algorithm. As the graphs showed, the model estimated by the partial least squares method gives predictive values closer to the original values. This indicates that the partial least squares method has an ideal regression model and high predictive ability, which makes it the most efficient method for predicting future values of the response variable.

6. Conclusions:

The data were tested for multicollinearity, by testing the variance inflation factor (VIF), which indicated that the data suffer from the problem of multicollinearity. The problem of multicollinearity was addressed using the partial least squares method with algorithms (SIMPLS, O-PLS) and the principal components method using algorithms (NIPALS, SVD), and a comparison was made between the two methods based on the first component and using the mean square error (MSE) as a standard comparison. The statistical analysis revealed the superiority of PLS with the algorithm (O-PLS), as it achieved the lowest MSE value. In order to improve the results, the genetic algorithm was used for the two estimation methods PLS with the algorithm (SIMPLS, O-PLS) and the principal components analysis with the algorithm (NIPALS, SVD), and a comparison was made between the two methods and the results of the comparison showed that the PLS method with the algorithm (O-PLS) was the best due to its lowest value of the mean square error. The regression model parameters were estimated based on the first component using the algorithms for both methods, and the regression model parameters were estimated using the improved methods. After concluding that the PLS method is better than the PCA method before and after using the genetic algorithm, we built an analysis of variance table based on the superior PLS method. The results indicated that there were significant differences between the explanatory variables and the response variable, which indicates the quality of the estimated model.

Authors Declaration:

Conflicts of Interest: None

-We Hereby Confirm That All The Figures and Tables In The Manuscript Are Mine and Ours. Besides, The Figures and Images, which are Not Mine, Have Been Permitted Republication and Attached to The Manuscript.

- Ethical Clearance: The Research Was Approved by The Local Ethical Committee in The University.

References:

Abdi, H. (2010), Partial least squares regression and projection on latent structure regression (PLS Regression). WIREs Comp Stat, 2: 97-106. https://doi.org/10.1002/wics.51

Abdulhadi, A. T., & Reda, S. M. (2020). Estimation of survival function using genetic algorithm. *Journal of Economics and Administrative Sciences*, *26*(122), 440–454. https://doi.org/10.33095/jeas.v26i122.2018

Abdullah, A. N., & Abbas, B. K. (2020). Comparison between Partial Least Squares Regression and Dendritic Regression Using Simulation. *Journal of Economics and Administrative Sciences*, *26*(120), 411–425. https://doi.org/10.33095/jeas.v26i120.1924

Albadrani, D. R. M., & Al-mawla, T. A. T. (2016). Comparison between Principal Component Regression and Partial Least Squares Methods with Application to Kirkuk Cement Plant. *Tikrit Journal of Pure Science*, *21*(7), 185–203. https://doi.org/10.25130/tjps.v21i7.1126

Albayati, M. (2012). A practical application of analyzing statistical data using the program (SPSS). *Al-Jazeera Press and Publishing.*

Albayati, M. M. & Shaker, H. H. (2018). Comparison between Partial Least Squares and SVPD for Estimating Logistic Regression Model Parameters in Case of Multicollinearity Problem Using Simulation. *Journal of Economic and Administrative Sciences, 24*(109), 458–471. https://doi.org/10.33095/jeas.v24i109.1559

Aldouri, Y. K., AlChalabi, H., & Lundberg, J. (2020). Risk-based life cycle cost analysis using a two-level multi-objective genetic algorithm. *International Journal of Computer Integrated Manufacturing*, *33*(10–11), 1076–1088. https://doi.org/10.1080/0951192X.2020.1757157.

Alkhafaji, M. A., & Saleh, R. A. (2021). A Statistical Study on the Parameters of the Skew Normal Distribution Depending on the Use of the Genetic Algorithm Using the Simulation Method. *Journal of Physics: Conference Series*, *1879*(3). https://doi.org/10.1088/1742-6596/1879/3/032017

Alrawi, A. G., & Issa, A. M. (2019). Use Principal Component Analysis Technique to Dimensionality Reduction to Multi Source. *Journal of Economics and Administrative Sciences*, *25*(115), 464–473. https://doi.org/10.33095/jeas.v25i115.1778

Alsabaah, S. A., & AlQuraishi, Z. K. M. (2018). Use Principle Component Regression Method In Addressing Linear Multiplicity Problem. *Karbala University Scientific Journal*, *16*(2), 248–261. https://www.iasj.net/iasj/article/153425

Alsafawi, S. Y., AlDin, S. D., & Shaker, S. M. (2010). Using the Partial Least Squares Method to Eliminate Multicollinearity. *Iraqi Journal of Statistical Sciences*, *10*(1), 115–128.

Andrecut, M. (2009). Parallel GPU implementation of iterative PCA algorithms. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, *16*(11), 1593–1599. https://doi.org/10.1089/cmb.2008.0221

Ding, J., Zhao, L., Liu, C., & Chai, T. (2014). GA-based principal component selection for production performance estimation in mineral processing. *Computers and Electrical Engineering*, *40*(5), 1447–1459. https://doi.org/10.1016/j.compeleceng.2013.12.014

Eiben , A.E & Smaith, J. (2015). Introduction to evolutionary computing. In *Natural Computing Series* (Vol. 28). https://doi.org/10.1007/978-3-662-43631-8_2

Elvira-Ortiz, D. A., Jaen-Cuellar, A. Y., Morinigo-Sotelo, D., Morales-Velazquez, L., Osornio-Rios, R. A., & Romero-Troncoso, R. de J. (2020). Genetic algorithm methodology for the estimation of generated power and harmonic content in photovoltaic generation. *Applied Sciences (Switzerland)*, *10*(2). https://doi.org/10.3390/app10020542

Hassan, M. M., Shaker. H. H., & Mohammed. N. J. (2020). Comparison between the two methods of regression of the letter and regression of the principal components using Monte Carlo simulation through the mean square error (MSE). *Journal of Al-Rafidain University College of Science*, *46*(1), 335–352. https://doi.org/10.55562/jrucs.v46i1.86.

Hubert, M., & Branden, K. V. (2003). Robust methods for partial least squares regression. *Journal of Chemometrics*, *17*(10), 537–549. https://doi.org/10.1002/cem.822

Hussain, J., & Nassir, A. (2015). Cluster Analysis as a Strategy of Grouping to Construct Goodness-of-Fit Tests when the Continuous Covariates Present in the Logistic Regression Model. *British Journal of Mathematics & Computer Science*, *10*(1), 1–16. https://doi.org/10.9734/bjmcs/2015/18616

Hussein, S. M., & Saleh, R. A. (2014). Comparison of some robust methods for estimating partial least squares regression parameters. *Journal of Economics and Administrative Sciences*, *20*(75), 413–431. https://doi.org/10.33095/jeas.v20i75.587

Jiang, S., Tian, H., Wang, Y., Jin, L., Rong, J., Kang, S., ... & Liu, Z. (2023). Optimization of source pencils loading plan with genetic algorithm for gamma irradiation facility. *Radiation Physics and Chemistry*, *207*, 110839. *Https://Doi.Org/10.1016/J RADPHYSCHEM.2023.110839*.

Kale, I. R., Pachpande, M. A., Naikwadi, S. P., & Narkhede, M. N. (2022). Optimization of advanced manufacturing processes using socio inspired cohort intelligence algorithm. *International Journal for Simulation and Multidisciplinary Design Optimization*, *13*. https://doi.org/10.1051/smdo/2021033

Kaneko, H. (2022). Genetic Algorithm-Based Partial Least-Squares with only the First Component for Model Interpretation. *ACS Omega*, *7*(10), 8968–8979. https://doi.org/10.1021/acsomega.1c07379

Leardi, R. (2003). Genetic algorithm-PLS as a tool for wavelength selection in spectral data sets. *Data Handling in Science and Technology*, *23*(C), 169–196. https://doi.org/10.1016/S0922-3487(03)23006-9

Liu, C., Zhang, X., Nguyen, T. T., Liu, J., Wu, T., Lee, E., & Tu, X. M. (2021). Partial least squares regression and principal component analysis: Similarity and differences between two popular variable reduction approaches. *General Psychiatry*, *35*(1), 1–5. https://doi.org/10.1136/gpsych-2021-100662

Liu, J., & Wong, D. S. H. (2011). Developing soft sensors based on orthogonal projections to latent structures with kernel algorithm. *IFAC Proceedings Volumes (IFAC-PapersOnline)*, *44*(1 PART 1), 14342–14347. https://doi.org/10.3182/20110828-6-IT-1002.00100

Mohammed, H. Y. and Mohammed, L. A. (2020). In the analysis of the Principle components of kernel. *Journal of Management and Economics*, *123*, 376–394. https://doi.org/10.4324/9781315755533-9

Ramzan , S., & Zahid, F.M. (2010). prediction method for Time- series regression models with multicollinearity. *World Applied Sciences Journal*, *11*(4), 443–450.

Saleh, R. A. (2016). Comparison between partial least squares and principal component methods using simulation. *Journal of Economics and Administrative Sciences*, *22*(87), 50–71. https://doi.org/10.33095/jeas.v22i87.725

Samosir, R. D., Salaki, D. T., & Langi, Y. (2022). Comparison of Partial Least Squares Regression and Principal Component Regression for Overcoming Multicollinearity in Human Development Index Model. *Operations Research: International Conference Series*, *3*(1), 1–7. https://doi.org/10.47194/orics.v3i1.126

Shang, X., Li, X., Morales-Esteban, A., & Chen, G. (2017). Improving microseismic event and quarry blast classification using Artificial Neural Networks based on Principal Component Analysis. *Soil Dynamics and Earthquake Engineering*, *99*(May), 142–149. https://doi.org/10.1016/j.soildyn.2017.05.008

Van Roon, P., Zakizadeh, J., & Chartier, S. (2014). Partial Least Squares tutorial for analyzing neuroimaging data. *The Quantitative Methods for Psychology*, *10*(2), 200–215. https://doi.org/10.20982/tqmp.10.2.p200