Zanco Journal of Human Sciences (ZJHS)
ISSN (Print):2218-0222
ISSN (Online): 2412-396X

Vol.28 Issue 5 2024
ID No.2278 (PP 264 - 277)
https://doi.org/10.21271/zjhs.28.5.17

Research Article

Estimating Models and Evaluating their Efficiency under Multicollinearity in Multiple Linear Regression: A Comparative Study

Saman Hussein Mahmood/ Department of Statistics and Informatics, College of Administration and Economics, Salahaddin University-Erbil

CORRESPONDENCE
Saman Hussein Mahmood
saman.mahmood@su.edu.krd

Reiceved	16/05/2024
Accepted	11/07/2024
Published	15/10/2024

Keywords:
Multicollinearity,
Multiple linear regressions,
artificial neural network,
Ridge regression,
Principal components analysis

Abstract

Multicollinearity between independent variables occurs in multiple linear regression analysis characterized by high correlations, which complicates discerning individual variable effect, impacting model accuracy, stability, and interpretation of relationships. The research aims to diagnose the multicollinearity problem between explanatory variables in the linear regression model and identifying the variables causing this problem based on the variance inflation factor (VIF), then estimation and evaluate the performance of three alternative methods, which are Ridge regression, principal components analysis, and Feedforward Neural Networks (FFNN) models with one and two hidden layers and application of the models to compressive strength data for high-performance concrete. The results showed that Ridge regression and PCA effectively addressed multicollinearity problem, but the single hidden layer model FFNN showed superior predictive accuracy in estimating the compressive strength of high-performance concrete when comparing RMSE, MAE, and R^2 values.

1. Introduction

Multiple regression analysis is statistical methods used to study the linear association between a dependent variable and several independent variables. Its purpose is to identify and measure the direction and strength of the relationship between the variables under study. The effective use of the equation and estimated coefficients requires the availability of some assumptions, and the accuracy of these estimators depends on the validity of these assumptions. One of the assumptions is that there is no exact collinearity among the independent variables, also known as multicollinearity.

Multicollinearity occurs between independent variables when there is a high correlation between the independent variables in the multiple regression models, which leads to difficulty distinguishing the effect of each variable on the model, affects the accuracy and stability of the model. In addition, multicollinearity leads to inflation of variance and difficulty interpreting the actual relationships between the variables.

Ridge regression and principal component regression are two commonly used biased regression methods. The biased regression methods attack the collinearity problem by computationally suppressing the effects of the collinearity. Ridge regression does this by reducing the apparent magnitude of the correlations. Principal component regression attacks the problem by regressing Y on the important principal components and then parceling out the effect of the principal component variables to the original variables (Rawlings et al., 1998, p67). Artificial Neural Networks (ANN) has been widely used in prediction, modeling, and classification problems. It has the ability to fit any complex function through training. As a result, the effect of multicollinearity is no longer a problem because the flat area due to multicollinearity in the multiple regression line cannot be seen in Neural Network (Li & Wang, 2019,p6) (Chan et al., 2022,p8). However, each method has its own set of strengths and limitations, and the decision which to use depends on the specific analysis objectives and criteria.

2. Method and materials

2.1Multiple Linear Regression

Multiple linear regression analysis is performed by constructing a prediction model consisting of a dependent variable or a response variable y based on an assumed linear relationship with several k independent predictors (x_1, x_2,.., x_k). (Rencher & Schaalje, 2008,p2). The model for multiple linear regressions with linear terms is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \quad \ldots\ldots\ldots.(1)$$

A random error ε accounts for any factors not included in the model. The regression coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ represent the parameters, where β_0 is the intercept and coefficient β_j (j=1,2,..k) signifies the change in the expected value of y for a unit change in x_j while keeping all other variables constant . The errors are assumed to be unobservable, independent random disturbances following a normal distribution with zero mean and constant variance, denoted as $\varepsilon \sim N(0, \sigma^2)$. Under these conditions, the ordinary least squares (OLS) estimation method provides unbiased, consistent, and efficient estimates of the unknown parameters. (Doane & Seward, 2021,p475)

Multiple linear regression analysis involves several basic assumptions: (1) linearity, (2) homoscedasticity ($Var(\varepsilon_i)=\sigma^2$ for all i) (3) independence of errors (($Cov(\varepsilon_i \varepsilon_j)=0$ for all i≠j) (4) normally residuals($\varepsilon_i \sim N(0, \sigma^2)$ and (5) No multicollinearity (($Corr(x_i, x_j)\approx 0$ for all i≠j). (Rawlings et al., 1998,p29)

2.2 Detecting the problem of multicollinearity:

Multiple linear regression models that include two or more explanatory variables may encounter multicollinearity problems. Detecting the problem of multicollinearity means revealing the degree of multicollinearity, and its purpose is not the presence or absence of multicollinearity. The following points must be taken into consideration.

Detecting multicollinearity depends on some basic rules, some informal and some formal, but the basic rules of thumb are all the same. Now we will look at some of these rules. We now consider some of these rules. (Gujarati & Porter, 2009,p320)

1. High R^2 value and non-significant estimated parameters of explanatory variables

2.The high correlation coefficients between explanatory variables and the fact that the correlation coefficients between pairs of variables in diagnosing the problem of multicollinearity because the mutual relationship between three or more variables may lead to a high degree of multicollinearity, even though the correlations between pairs of variables are low. Therefore, the best procedure to measure the degree of multicollinearity is to calculate the Eigenvalues of the correlation matrix and its corresponding condition index, in addition to variance inflation factors (VIF) and variance decomposition rates.

3.Tolerance and variance inflation factor: The variance inflation factor is the dominant approach to detect the presence of the multicollinearity problem. It measures the extent to which the variances of estimated regression coefficients are inflated when there is a linear relationship between explanatory variables. Can be found by relying on the coefficient of determination as in the following formula:-

$$VIF_J = \frac{1}{1-R_j^2} \quad j = 1,2,3,\dots,\mathrm{p} \quad \dots\dots.(2)$$

R_j^2 : It represents the coefficient of determination of the explanatory variable X_j in multiple linear regression.
P: Number of explanatory variables.
Tolerance is calculated as the reciprocal of VIF

$$TOL_J = \frac{1}{VIF_j} \quad \mathrm{j} = 1,2,3,\dots,\mathrm{p} \quad \dots\dots(3)$$

A low tolerance or high VIF indicates a high degree of multicollinearity between predictor variables. When the VIF value equals 1, it indicates no correlation among independent variables. A VIF value greater than one indicates a departure from orthogonally and typically suggests correlations among variables. If VIF falls between 1 and 5, it indicates moderate correlation among variables. The challenging value of VIF is between 5 to 10 as it specifies the highly correlated variables. If the VIF exceeds 5 to 10, multicollinearity among predictors in the regression model is present, and a VIF greater than 10 indicates weak estimation of regression coefficients due to multicollinearity. (Shrestha, 2020,p41)

2.3 Ridge regression analysis

When employing the least squares method with non-orthogonal data, it may lead to highly inaccurate estimates of regression coefficient as inflated variances and causing instability. This inflation is represented by the diagonal elements of the standard matrix. (Pati, 2020,p11), (Montgomery et al., 2012,p542).

Several procedures have been developed for obtaining biased estimators of regression coefficients. Among them is ridge regression, first proposed by Hoerl and Kennard [1970a, b], consists of adding a constant quantity to the diagonal elements of the matrix before taking its inverse, This leads to a biased estimator β_R of β , known as the ridge estimator, according to the following formula: (Kutner et al., 2005), (Montgomery et al., 2012,p543,p184)

$$\hat{\beta}_R = (X'X + kI)^{-1}X'Y \text{ …… (4)}$$

k takes the value of (0<k<1), if k = 0, the ridge estimator is the least-squares estimator.

Where ($\acute{X}X$ is a diagonal matrix representing the eigenvalues values of the matrix

The basic properties of the ridge solution include: (Duzan, 2020,p184)

i. The sum of squared residuals is a monotone which increases as a function of k
ii. The ridge estimator is a linear transformation of the OLS method since

$$\hat{\beta}_R = (X'X + kI)^{-1}(X'Y) = (X'X + kI)^{-1}(X'X)\hat{\beta} \text{ …………(5)}$$

iii. $E(\hat{\beta}_R)$ is a biased estimator of β**.**

$$E(\hat{\beta}_R) = (X'X + kI)^{-1}(X'X)\hat{\beta} \neq \beta \text{ …………(6)}$$

iv. The constant k is denoted as the biasing parameter. The covariance of $\hat{\beta}_R$ is

$$cov(\hat{\beta}_R) = \sigma^2(X'X + kI)^{-1}(X'X)(X'X + kI)^{-1} \text{ …..(7)}$$

v. The mean square error (MSE) of $\hat{\beta}_R$is given by

$$MSE(\hat{\beta}_R) = Var(\hat{\beta}_R) + (bais\ in\ \hat{\beta}_R)^2 \text{ …….(8)}$$

$$= \sigma^2 Tr[(X'X + kI)^{-1}(X'X)(X'X + kI)^{-1}] + k^2\beta'(X'X + kI)^{-2}\beta \text{……(9)}$$

$$= \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{\lambda_i + k} + k^2\beta'(X'X + kI)^{-2}\beta \text{ …..(10)}$$

In Equation (10), λ1, λ2,…, λp are the eigenvalues of X′X, the first term on the right-hand side is the sum of the variance of the parameters in $\hat{\beta}_R$ and the second term is the sum of the squared biases. It is clear that the sum of variance decreases as k increases, while the squared bias increases with k. (Montgomery et al., 2012,p542)

2.4 Principal Components Analysis

Principal components analysis (PCA) is a fundamental method in multivariate statistics that finds application in various fields. It explores and understands the interrelationship between many variables and transforms them into a set of new, unrelated variables (orthogonally), called principal components. Thus, it is possible to reduce the size of a data set that contains a large number of different variables that are related to each other, and these components correspond numerically to the variables under study. This facilitates the analysis process despite the possibility of complicating the objectives of the study, as is the case when dealing with a large volume of information. This approach allows for the selection and analysis of a concise set of key components, simplifying the investigation process while preserving the essence of the original data. The PCA can be written as follows: (Cohen et al., 2002,p428), (Rashid & Tofiq, 2022,p87)

$$Pc_i = a_{1i}x_1 + a_{2i}x_2 + \cdots\cdots\cdots + a_{mn}x_m \quad \ldots\ldots\ldots(11)$$

$$Pc_i = \sum_{j=1}^{m} a_{ji}x_j \qquad (i,j = 1,2,\ldots,m)$$

Whereas
PCi: represents the PCA
a_{ji}: represents the coefficient j in the principal component i
It can be written in matrix form as follows: $P_c = AX$

$$\begin{bmatrix} PC1 \\ PC2 \\ \vdots \\ PCp \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} \ldots\ldots\ldots\ldots(12)$$

The values of the principal component variables Pc1, Pc2, ..., Pcp correspond to the vector Pc, while the values of the random variables x1, x2, x3, ..., xp correspond to the vector x. The eigenvalues are represented by $\lambda_1 > \lambda_2 > \cdots . > \lambda_p$. The constants a_{ij} correspond to the elements of the ith eigenvector associated with the eigenvalueλ_i. (Al-Rawi, 1987,p487)

We calculate the correlation matrix or covariance matrix of the explanatory variables as shown in Equation (13).

$$\mathbf{R}= \begin{bmatrix} \mathbf{1} & \boldsymbol{r_{12}} & \cdots & \boldsymbol{r_{1p}} \\ \boldsymbol{r_{21}} & \mathbf{1} & \cdots & \boldsymbol{r_{2p}} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{r_{p1}} & \boldsymbol{r_{p2}} & \cdots & \mathbf{1} \end{bmatrix} \quad \mathbf{\ldots\ldots..(13)}$$

r_{ij} is simple correlation between two variables, when i,j=1,2,3,…., p

To find the characteristic roots (λ_i), we subtract from the diagonal values of the matrix R and then make its term equal to zero, so we obtain the characteristic equation of the matrix.

$$|\boldsymbol{R} - \boldsymbol{\lambda I}|= \begin{bmatrix} \mathbf{1-\lambda} & \boldsymbol{r_{12}} & \cdots & \boldsymbol{r_{1p}} \\ \boldsymbol{r_{21}} & \mathbf{1-\lambda} & \cdots & \boldsymbol{r_{2p}} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{r_{p1}} & \boldsymbol{r_{p2}} & \cdots & \mathbf{1-\lambda} \end{bmatrix} \mathbf{=0} \quad \mathbf{\ldots\ldots..(14)}$$

The form of the characteristic equation is a polynomial of degree p

$$\boldsymbol{\lambda^p + C_{m-1}\lambda^{p-1} + \cdots \ldots\ldots.. + C_1\lambda + C_0 = 0} \quad \mathbf{\ldots\ldots(15)}$$

When solving this equation we will get p roots and these roots ($\lambda_1 > \lambda_2 > \cdots . > \lambda_p$) are arranged from largest to smallest.

Explanation of variance and variance of the principal components is as follows:

$$var(x_i) = \sigma_{ii} \quad i = 1,2, \dots, p \qquad \dots\dots\dots\dots\dots(16)$$

$$Var(Pc_i) = \lambda_i \quad i = 1,2, \dots, p \qquad \dots\dots\dots\dots\dots(17)$$

$$Cov\left(Pc_i, Pc_j\right) = 0 \qquad \dots\dots\dots\dots\dots(18)$$

The following ratio provides the percentage of the original data's variability explained by the ith principal component. This can be calculated using the formula: (Johnson & Wichern, 2014,p430)

$$\frac{\boldsymbol{Var(Pc_i)}}{\boldsymbol{\sum Var(Pc_i)}} = \frac{\boldsymbol{\lambda_i}}{\boldsymbol{\sum \lambda_i}} \qquad \dots\dots\dots\dots\dots(19)$$

The percentage of variance that is explained by a few of the principal components by calculating the sum of the eigenvalues of those components and comparing this total to the sum of the eigenvalues

Significant principal components are chosen by evaluating the cumulative percentage of variance explained for each component. The number of selected principal components corresponds to the count of characteristic roots ($\lambda > 1$). Therefore, the first Principal component, derived from the first eigenvectors, explains the largest amount of variation in the original data, Subsequent principal components then explain the remaining variance in descending order. The amount of variation captured by each PC is given by their corresponding eigenvalues. (Samarasinghe, 2006,p287)

The first principal component (Pc1) explains the largest proportion of the total variance of the explanatory variables, followed by the second principal component (Pc2), and so forth for the remaining components. (Blbas et al., 2017,p47)

2.5 Artificial Neural Networks

The fundamental concept of Artificial Neural Networks (ANN) is to simulate the structure and function of the biological neural networks of the human brain. ANNs are a set of models and algorithms that have demonstrated an increasingly noteworthy role in the practical solution of difficult and diverse problems. An artificial neural network (ANN) consists of interconnected artificial neurons that use a mathematical or computational model to process information, derived from the connectionist approach to computation. An artificial neural network essentially consists of a network of basic processing units (neurons) capable of exhibiting complex processes and overall behavior, dictated by the connections between these processing units and their respective parameters. (Fausett, 1994,p25) (Samarasinghe, 2006,p11)

The various advantages to ANN including provide highly accurateresults when compared with regression model, easily updated, suitable for dynamic environment, generallyrobust to missing or inaccurate data (Sharma & Chopra, 2011, P34).

In order to achieve the best network architecture that accurately understands input and output data, two basic factors are taken into consideration:

i. Choose the most accurate training algorithm.

ii. Determine the appropriate number of hidden neurons.

Accordingly, different training algorithm and hidden nodes were evaluated to determine to determine the best training algorithm and the optimal number of hidden nodes that would produce the most accurate network structure. (Ekiugbo et al., 2021,p235)
Neural network regression uses feedforward ANNs according to the type of supervised training to process the regression function. Where the input data is received by the input layer, each neuron is connected to neurons in the subsequent layer, known as the hidden layer (Figure 1). (Sharma & Chopra, 2011,P35), (Haykin, 1999,p132)

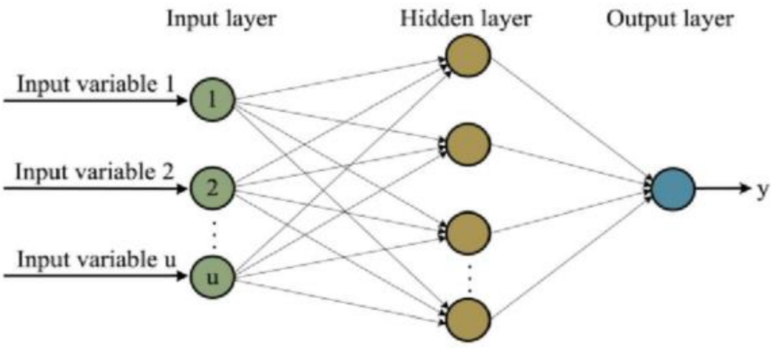

Fig. 1 Feed Forward Neural Network with one hidden layer and one output layer

Mathematically, this procedure can be expressed as follows: (Haykin, 1999,p132)

$$H_j = v_j + \sum_{i=1}^{n} v_{ij} X_i \quad \ldots\ldots\ldots\ldots.(20)$$

Where: v_j=the bais term for hidden unit j
v_{ij}=the weight from input note i to hidden node j
X_i=i^{th} input variables
At each hidden node j, an activation function is applied to estimate the output of the hidden layer unit:

$$h_j = f(H_j) \quad \ldots\ldots\ldots\ldots.(21)$$

The activation values from each node in the hidden layer are sent to the output layer, and then an output unit collects (Y_k; k=1,...,m) its weighted input signals as follows:

$$Y_k = W_k + \sum_j^P h_j W_{jk} \quad \ldots\ldots.(22)$$

Where: w_{jk}=the weight from hidden node j to output node k
Finally, the output layer generates the corresponding outputs based on the provided inputs and then applies an activation function to estimate the outputs of the output layer unit: (Li & Wang, 2019,p5)

$$y_k = f(Y_k) \quad \ldots\ldots\ldots\ldots.(23)$$

Backpropagation ANN is a supervised learning algorithm widely used to train feedforward neural networks. It adjusts the network weights and biases based on the least square error between the predicted (NN output) (y_k) and the actual (model data) output (t_k) until the optimal weights are reached, and propagates this error across the network inversely. (Haykin, 1999,p133) (King, 1999,p159)

$$Min \sum_{k=1}^{K} (y_k - t_k)^2 \quad \ldots\ldots\ldots\ldots\ldots.(24)$$

It includes forward and backward propagation stages, calculates outputs and adjusts parameters to reduce errors. Using gradient optimization, it updates weights and biases to reduce errors.
The objective function of NN in Equation (24) is similar to that used in multiple linear regressions. Both techniques aim to minimize the sum of squared differences between observed and expected values. (King, 1999,p161)

3. Results and Discussion

The practical aspect of this research includes estimating of the multiple linear regression model and then diagnosing the multicollinearity problem between the explanatory variables for the compressive strength of high-performance concrete (HPC) and identifying the variables causing this problem based on the variance inflation factor (VIF). Then (Ridge regression, PCA, ANN) were used to estimate parameters and factors as methods for treating multicollinearity to reach estimates that are more expressive of the effect of the explanatory variables on the compressive resistance function of concrete.
The experimental data used in this study were obtained from the from a machine learning repository managed by the University of California, Irvine (UCI) and curated by (Yeh, 1998,p1800). Concrete samples assessed by different university research facilities to evaluate the prediction capabilities of each AI technique. The data consists of 8 independent variables in addition to the dependent variable, and the number of samples used in this research 400 samples.
We will apply the three statistical measures; Root Mean Squared Error (RMSE), Mean absolute error (MAE) and coefficient of determination(R^2) to determine the best model among the estimated models.

3.1 Diagnosing multicollinearity in data

The multiple linear regression model estimation and analysis results in Table(1) indicate that the VIF value of some variables exceeded 10, which is an indication of the presence of multicollinearity in the model. Multicollinearity can affect the reliability of regression coefficients and predictors, so it is necessary to consider addressing multicollinearity problem in the model, especially for variables with high VIF values. Find some statistical indicators (RMSE, MAE and R^2) which are equal to (8.10, 6.23 and 78.8%)

Table (1) Results of multiple linear regression

Variables	Unstandardized Coefficients	VIF
Constant	-37.013	
Cement (x1)	0.120	**21.159**
Blast Furnace Slag (x2)	0.136	**11.983**
Fly Ash (x3)	0.064	**15.057**
Water (x4)	-0.154	**4.177**
Superplasticizer (x5)	-0.096	**2.458**
Coarse Aggregate (x6)	0.033	**11.211**
Fine Aggregate (x7)	0.020	**11.270**
Age x8	0.279	**1.015**

4.2 Ridge regression analysis

Ridge regression enhances the stability of parameter estimates, particularly for variables exhibiting high VIF values. Various methods were employed to determine the optimal value of the ridge parameter (k), typically set between 0 and 1. and by increasing the value of (k) by 0.05 for each iteration with finding (VIF) and statistical indicators, we found that the optimal value was found to Parameter = 0.02. The estimated coefficients for the variables in the ridge regression model are presented in Table 2, where VIF<5 indicate the absence of multicollinearity problem in the model. Find some statistical indicators (RMSE, MAE and R^2) which are equal to (8.24, 6.27 and 69.03%)

Table (2) Model Results for Ridge Parameter = 0.02

Parameter	Unstandardized Coefficients	standardized Coefficients	***Variance Inflation Factor***
(Constant)	102.78		
Cement	0.043	0.427	**0.590**
Blast Furnace Slag	0.061	0.369	**0.649**
Fly Ash	-0.037	0.017	**0.724**
Water	-0.199	-0.208	**0.773**
Superplasticizer	0.198	-0.011	**0.902**
Coarse Aggregate	-0.022	-0.022	**0.737**
Fine Aggregate	-0.035	-0.078	**0.630**
Age	0.231	0.499	**0.698**

3.3 Principal components method

The principal components regression method is applied using 8 input variables, and the variables that have different measurement units, were standardized. Then, eigenvalues greater than one were determined, from which the first four components were extracted from the explanatory variables after rotate them. These components collectively explained up to 83.659% of the total variance. The results are in Table (3).

Table(3) Total variance explaine

Component	**Total**	**% of Variance**	**Cumulative %**
1	**2.514**	31.42	31.420
2	**1.712**	21.39	52.814
3	**1.438**	17.98	70.794
4	**1.029**	12.87	**83.659**

Table (4) presents the rotated component matrix obtained from the principal component analysis. Each variable is associated with different components based on their loadings. Higher absolute values indicate stronger associations with the respective components.

Component 1, which explains (31.42%) of the variance, is strongly associated with fly ash, coarse aggregate and cement.

Component 2 which explains (21.39%) of the variance, is strongly with water and superplasticizer.

Component 3 which explains (17.98%), is strongly associated with Fine Aggregate and Blast Furnace Slag.

Component 4, which explains (12.87%) is strongly related to Age.

Table (4) Rotated Component Matrix

	Component			
	1	2	3	4
Fly Ash	**-0.86**	-0.09	0.28	-0.08
Coarse Aggregate	**-0.83**	-0.12	-0.38	-0.06
Cement	**0.82**	0.23	-0.14	-0.09
Water	0.08	**-0.94**	-0.09	-0.01
Superplasticizer	0.39	**0.82**	0.03	0.03
Fine Aggregate	0.09	0.19	**0.92**	0.08
Blast Furnace Slag	0.49	0.18	**-0.60**	0.19
Age	-0.02	0.01	0.011	**0.98**

The multiple regression method was used to estimate the concrete compressive strength equation through regression on the extracted factors (PC1, PC2, PC3, and PC4). To obtain the best input factor for the model (PCA-MLR), the stepwise algorithm is used, and the VIF values for all input variables indicate that there is no multicollinearity problem and significant values for all predictors ($P < 0.05$). The adjusted R square indicates that approximately 74.8% of the variance in concrete compressive strength is explained by the four factors. The result is in Table 5 and 6

We write the multiple regression equation as follows

Concrete compressive strength (Y) =42.612+10.743PC1+5.023PC2-3.877PC3+9.064PC4

Table(5) VIF values and cofficients for the principal component regression method

Model	B	t	Sig.	VIF
(Constant)	42.612	95.395	<0.001	
PC1	10.743	24.02	<0.001	1
PC2	5.023	11.231	<0.001	1
PC3	-3.877	-8.668	<0.001	1
PC4	9.064	20.267	<0.001	1

Table (6) Result of the stepwise algorithm for the (PCA-MLR) model.

Input	**Adj. R^2**	**Sig**	**VIF**
PC1	31.9%	0.00	No problem
PC1,PC4	58.7%	0.00	No problem
PC1,PC4,PC3	68.9%	0.00	No problem
PC1,PC4,PC3, PC2	74.8%	0.00	No problem

3.4 Artificial Neural Networks ANN

Application of the Feedforward Neural Network (FFNN) model to analyze the compressive strength of concrete with changing hidden layer (one and two layers):-

3.4.1 Feedforward Neural Networks one hidden layer

The model consists an input layer containing eight variables as inputs, one hidden layer, with the final layer being the output layer responsible for predicting the concrete's compressive strength. The input data will be randomly partitioned into three sets and 70% of the data will be assigned to the training set, while 15% will be allocated to both the validation and testing sets.

In order to determine the best FFNN model, we varied the number of nodes in the hidden layer from 1 to 10 and repeated each experiment (500) times for each node and then calculated the average of the statistical indicators (RMSE, MAE and R^2). The results in table (7) indicate that the FFNN model with the (8:10:1) architecture performs optimally and has minimum values (RMSE and MAE), as well as a high R^2 value.

Table (7) Comparison of FFNN models for compressive strength of concrete

No. of Nodes in hidden layer	1	2	3	4	5	6	7	**8**	9	10
RMSE	**8.04**	**7.08**	**6.50**	**6.07**	**5.80**	**5.63**	**5.58**	**5.34**	**5.19**	**5.17**
MAE	**6.20**	**5.43**	**4.91**	**4.56**	**4.33**	**4.16**	**4.10**	**3.92**	**3.78**	**3.75**
R^2	**79.6**	**84.2**	**86.7**	**88.5**	**89.5**	**90.1**	**90.3**	**91.1**	**91.61**	**91.67**

3.4.2 Feedforward Neural Networks two hidden layer

Similarly, to our previous approach, to determine the number of nodes in two hidden layers from 1 to 10 and repeated each experiment (500) times for each node, followed by calculating the average statistical indicators. From the results, the FFNN model (8:10:6:1) is the best model with the lowest value (RME = 5.543 and MAE = 3.941) and the highest R^2 value is 90.45%.

The analysis detects that adding two hidden layers fails to improve the values of the statistical indicators. Therefore, we will choose the first model. We'll now proceed to estimate the importance of each input variable within this model FFNN(8:10:1), as illustrated in Figure 2 and Table (8) which represents all the basic variables. The x-axis shows the normalized importance or percentage impact on concrete strength. According to the analysis "Cement" has been identified as the most important variable followed by Age, Water and other variables with less effect.

Table 8 The Importance of predictors as illustrated by FFNN1

Variables	Importance	Normalized Importance
Cement	0.248	(0.248/0.248) = **100%**
Blast Furnace Slag	0.114	(0.114/0.248) = **46.0%**
Fly Ash	0.035	(0.035/0.248) = **14.1%**
Water	0.126	(0.126/0.248) = **50.7%**
Superplasticizer	0.093	(0.093/0.248) = **37.5%**
Coarse Aggregate	0.068	(0.068/0.248)= **27.3%**
Fine Aggregate	0.085	(0.085/0.248) = **34.4%**
Age	0.232	(0.232/0.248)= **93.9%**

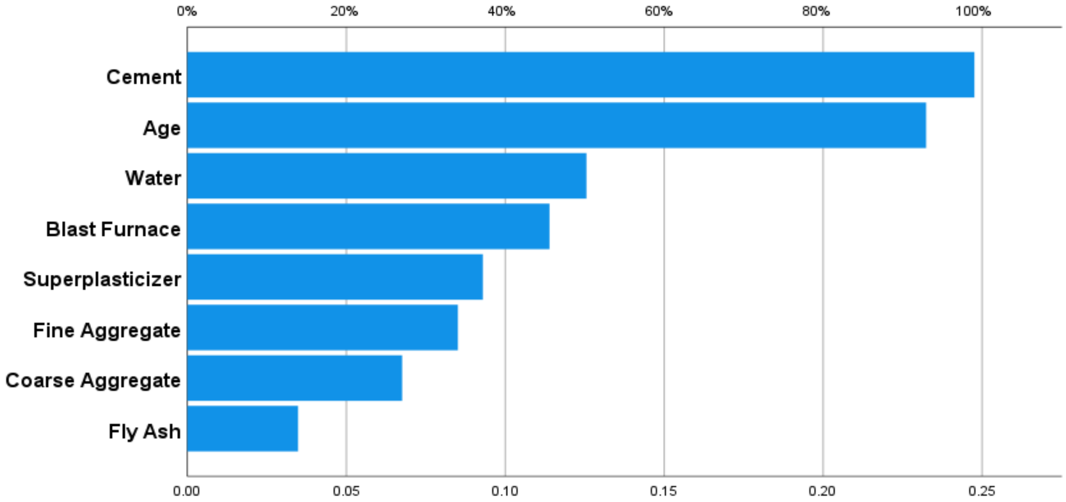

Fig.2.The importance of independent variables

3.5 Compare model results

1. Comparing the standardized coefficients from Ridge regression (Table 2), the parameter values across four factors in PCA (Table 4), and the importance of the predictors in FFNN1 (Table 5) . From the results, it can be concluded that the basic variables have a greater effect on the strength of concrete: cement, age, water, blast furnace slag and other variables have less effect.

2. Table (9) indicates that the Feedforward Neural Network with one hidden layer have better performance than other methods, as it achieving the lowest RMSE and MAE values, in addition to the highest R^2 value.

Table 9 Comparison of Models Performances

Methods	RMSE	MAE	R^2
MLR	8.10	6.23	78.8%
Ridge MR	8.24	6.27	69.03%
MLR-PCA	8.88	6.93	74.80%
FFNN(one Hidden)	5.17	3.75	91.67%
FFNN(two Hidden)	5.543	3.941	90.45%

4. Conclusion

1. Results of applying and studying different methods (ridge regression, principal component analysis, and feedforward neural network) to estimate compressive strength concrete models more effectively than traditional multiple linear regression, especially in addressing multicollinearity, were compared. The comparison of results revealed that the FFNN(8:10:1) model outperformed other methods, as it achieved the lowest RMSE and MAE values, as well as the highest R^2 value.
2. From the results different methods, it can be concluded that the basic variables have a greater effect on the strength of concrete: cement, age, water, blast furnace slag and other variables have a lesser effect
3. Determining the optimal number of layers and nodes in the hidden layer is the basic and difficult aspect of neural networks, so each experiment was repeated (500) times for each node, and then the average statistical indicators were calculated. Increasing the second hidden layer of the model (FFNN) did not improve the model, and therefore it is the best model in neural networks (FFNN (8:10:1)).

5. References

- Al-Rawi, K.M., 1987. Introduction to Regression Analysis. Dar Al-Kutub for Printing and Publishing, Mosul University.
- Blbas, T.A., Mahmood, H. & Omer, A., 2017. A Comparison results of factor analysis and cluster analysis to the migration of young people from the Kurdistan Region to Europe. ZANCO Journal of Pure and Applied Sciences, 29(4), pp.44-55.
- Chan, Y.-L. et al., 2022. Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. Mathematics, 10(1283), pp.1-17.
- Cohen, , Cohen, , West, S.G. & Aiken, L.S., 2002. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, Third edition. Routledge.
- Doane, D.P. & Seward, L., 2021. Applied Statistics in Business and Economics. 7th edition.McGraw-Hill Education.
- Duzan, , 2020. A comparison between the method of least squares and ridge regression in the Presence of Multicollinearity in regression analysis. International Science and Technology Journal, 23, pp.182-201.
- Ekiugbo, A., Amiolemhen, P. & Ariavie, G.O., 2021. Performance of Multiple Linear Regression and Artificial Neural Network in Predicting Risk Index. Journal of Science and Technology Research, 3(4), pp.233-44.
- Fausett, L., 1994. Fundamentals of Neural Networks: Architectures, Algorithms, and Applications. Prentice-Hall.
- Gujarati, D.N. & Porter, D.C., 2009. Basic Econometrics. 5th ed. McGraw-Hill Irwin.
- Haykin, S.S., 1999. Neural Networks: A Comprehensive Foundation. 2nd ed. Prentice Hall.

- Hoerl, A. & Kennard, R., 1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 12, pp.55-67.
- Johnson, & Wichern, , 2014. Applied Multivariate Statistical Analysis. 6th ed. Pearson Education Limited.
- King, S.L., 1999. Neural Networks vs. Multiple Linear Regression for Estimating previous diameter. In Stringer, & Loftis, D.L., eds. 12th Central Hardwood Forest Conference. Kentucky, 1999,pp159-166
- Kutner, M.H., Nachtsheim, C.J., Neter, J. & Li , , 2005. Applied Linear Statistical Models, 5 Edition. McGraw-Hill Irwin.
- Li , M. & Wang, , 2019. An Empirical Comparison of Multiple Linear Regression and Artificial Neural Network for Concrete Dam Deformation Modelling. Mathematical Problems in Engineering, 1(1), pp.1-13.
- Montgomery, D.C., Peck, E.A. & Vining, G., 2012. Introduction to Linear Regression Analysis. 5th ed. John Wiley & Sons.
- Pati, K.D., 2020. Estimate The Parameters In Presence Of Multicollinearity And Outliers Using Bisquare Weighted Ridge Least Median Squares Regression (WRLMS). Journal of University of Duhok, 32(2), pp.9-24.
- Rashid, M.J. & Tofiq, O., 2022. Application of Principal Component Analysis for Steel Material Components. Kurdistan Journal of Applied Research, 7(2), pp.85-94.
- Rawlings, O., Pantula, G. & Dickey, D.A., 1998. Applied Regression Analysis: A Research Tool, 2nd Edition. Springer.
- Rencher , A.C. & Schaalje, G., 2008. *Linear models in statistics*. 2nd ed. John Wiley & Sons.
- Samarasinghe, , 2006. *Neural Networks for Applied Sciences and Engineering: From Fundamentals to Complex* Pattern Recognition. Auerbach Publications.
- Sharma, A., & Chopra, A., 2011, *Artificial Neural Networks: Applications In Management*, IOSR Journal of Business and Management (IOSR-JBM), Volume 12, Issue 5, PP 32-40
- Shrestha, N., 2020. Detecting Multicollinearity in Regression Analysis. American Journal of Applied Mathematics and Statistics, 8(2), pp.39-42.
- Yeh, I.C., 1998. Modeling of strength of high-performance concrete using artificial neural networks. Cement and Concrete Research. Elsevier, 28(12), pp.1797-1808.

خەمڵاندنی مۆدێلەکان و هەڵسەنگاندنی کاراییەکانیان لە ژێر فرە پەیوەیستی هێڵەکی لە لاریبوونی هێڵی فرەیی: لێکۆڵینەوەیەکی بەراوردکاری

سامان حسین محمود

بەشی ئامار و زانیارییەکان، کۆلێژی بەریوەبردن و ئابووری، زانکۆی سەلاحەدین-
هەولێر،هەولێر،هەرێمی کوردستان،عیراق
saman.mahmood@su.edu.krd

پوختە

فرە پەیوەیستی هێڵەکی نێوان گۆڕاوە سەربەخۆکان لە شیکاری لاریبوونی هێڵی فرەیی ڕوودەدات کە بە پەیوەندییە بەرزەوە ، کە تێگەیشتن لە کاریگەری گۆڕاوە تاکەکان ئاڵۆز دەکات، کاریگەری لەسەر وردی مۆدێل، سەقامگیری و لێکدانەوەی پەیوەندییەکان هەیە. ئامانجی توێژینەوەکە دەستنیشانکردنی کێشەی فرە پەیوەیستی هێڵەکی نێوان گۆڕاوە سەربەخۆکان لە مۆدێلی لاریبوونی هێڵی فرەیی و دەستنیشانکردنی ئەو گۆڕاوانەی کە هۆکاری ئەم کێشەیەن لەسەر بنەمای فاکتەری هەڵاوسانی جیاوازی (VIF)، پاشان خەملاندن و هەڵسەنگاندنی سێ شێوازی بەدیل، کە بریتین لە لاریبوونی ڕیج، سەرەکی شیکاری پێکهاتەکان، و مۆدێلی تۆڕی دەماری دەستکرد (FFNN) بە یەک و دوو چینە شاراوەکان و بەکارهێنانی مۆدێلەکان بۆ داتاکانی هێزی پەستان بۆ کۆنکرێتی کارایی بەرز. ئەنجامەکان دەریانخست کە پاشەکشەی ڕیج و PCA بە شێوەیەکی کاریگەر کێشەی فرە پەیوەیستی هێڵەکی چارەسەر کرد، بەڵام مۆدێلی FFNN وردبینی پێشبینیکردنی بەرزتری نیشان دا لە خەملاندنی هێزی پەستانی کۆنکرێتی کارایی بەرز کاتێک بەهاکانی RMSE، MAE و R^2 بەراورد دەکرێت.

ووشە سەرەکییەکان : فرە پەیوەیستی هێڵەکی ، لاریبوونی هێڵی فرەیی ، تۆڕی دەماری دەستکرد، پاشەکشەی ڕیجی، شیکاری پێکهاتە سەرەکییەکان

تقدير النماذج وتقييم كفاءتها في ظل العلاقة الخطية المتعددة في الانحدار الخطي المتعدد: دراسة مقارنة

سامان حسين محمود
قسم الإحصاء والمعلوماتية، كلية الإدارة والاقتصاد، جامعة صلاح الدين-أربيل،أربيل.اقليم
كوردستان،العراق
saman.mahmood@su.edu.krd

ملخص

تحدث العلاقة الخطية المتعددة بين المتغيرات المستقلة في تحليل الانحدار الخطي المتعدد الذي يتميز بارتباطات عالية، مما يعقد تمييز تأثير المتغير الفردي، مما يؤثر على دقة النموذج والاستقرار وتفسير العلاقات. يهدف البحث إلى تشخيص مشكلة التعدد الخطي بين المتغيرات التفسيرية في نموذج الانحدار الخطي وتحديد المتغيرات المسببة لهذه المشكلة بالاعتماد على عامل تضخم التباين(VIF) ، ثم تقدير وتقييم أداء ثلاث طرق بديلة وهي انحدار ريدج، تحليل المكونات الرئيسية، ونماذج الشبكات العصبية المغذية (FFNN) ذات الطبقة المخفية الواحدة والطبقتين، وتطبيق النماذج على بيانات قوة الضغط للخرسانة عالية الأداء. أظهرت النتائج أن انحدار ريدج و PCA قد عالجا بشكل فعال مشكلة الخطية المتعددة، لكن نموذج الطبقة المخفية الفردية FFNN أظهر دقة تنبؤية فائقة في تقدير قوة الضغط للخرسانة عالية الأداء عند مقارنة قيم كل من RMSE ، MAE و R^2.

الكلمات المفتاحية: التعدد الخطي، الانحدارات الخطية المتعددة، الشبكة العصبية الاصطناعية، انحدار ريدج، تحليل المكونات الرئيسية