

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452

NAAS Rating (2026): 4.49

Maths 2026; 11(3): 55-60

© 2026 Stats & Maths

<https://www.mathsjournal.com>

Received: 12-01-2026

Accepted: 21-02-2026

Consul

Department of Mathematics,
Faculty of Science, Niger Delta
University, Bayelsa State,
Nigeria

Juliana Iworikumo Yogo

Department of Mathematics,
Faculty of Science, Niger Delta
University, Bayelsa State,
Nigeria

Ifiemi Ilahn

Department of Mathematics,
Faculty of Science, Niger Delta
University, Bayelsa State,
Nigeria

Corresponding Author:

Juliana Iworikumo Yogo

Department of Mathematics,
Faculty of Science, Niger Delta
University, Bayelsa State,
Nigeria

Principal component analysis as a preprocessing technique for multiple linear regression models; dealing with superfluous variables

Consul, Juliana Iworikumo Yogo and Ifiemi Ilahn

DOI: <https://www.doi.org/10.22271/math.2026.v11.i3a.2276>

Abstract

Most data analysts often face the challenge of dealing with high dimensional data set with many variables. Often, several of these variables are highly correlated or convey overlapping information. Such variables or data are often termed superfluous data as they contribute little to the overall explanatory power of the dataset. This research uses both principal component analysis (PCA) as a reduction technique and multiple linear regression as predictive method and dealing with superfluous variables. This study demonstrates how PCA simplifies complex data structures, improves computational efficiency, and enhances the reliability of statistical models. A dataset collected using the personal interview method of 115 respondents' and fourteen variables was considered for data analysis. The R Software was used to implement the analysis. The multiple linear regression model of the dependent on the independent variables indicated that a very low (17.89%) proportion of the variance in the dependent variable was explained by the independent variables. A correlation analysis was performed to understand the relationship between the variables in the dataset which yielded some highly correlated variables. The PCA was performed on the set of independent variables and it showed that a subset of the top principal components (seven components) had a cumulative explained variance of 90%. Multiple linear regression model was performed using the principal components as independent variables and the dependent variable from the dataset. The result showed that 95.59% of the variance in the dependent variable was explained by the independent variables and thus, this model is a better model for the predicting the medical expenses.

Keywords: Superfluous variables, correlation, principal component analysis, predictive models, multicollinearity

1. Introduction

Most data analyst often deal with large-scale multidimensional data set with many variables which makes the calculations inconvenient. Often, several of these variables are highly correlated or convey overlapping information. They would want to reduce the number of variables but are not sure which variables to remove without losing information. Some data may include redundant or irrelevant features which will require reducing the dimensionality of the data while preserving its essential structure. Such data are called superfluous data as they contribute little to the overall explanatory power of the dataset. These redundant or superfluous variables do not contribute meaningful insight but instead increase data complexity, computational cost, and the likelihood of multicollinearity in predictive modeling. As a result, the accuracy, interpretability, and efficiency of data analysis or building predictive models may be compromised. There is the need for a statistical technique that can identify and reduce these redundant variables while preserving the key information within the data.

Superfluous variables do not enhance the analytical outcome but may distort statistical results or obscure meaningful patterns within the data. The inclusion of highly correlated variables can increase the dimensionality of data unnecessarily, complicate model interpretation, and introduce multicollinearity in regression analyses (Mundfrom & Kay, 2018) [12]. In regression or classification models, multicollinearity caused by these redundant variables can make it

difficult to determine the true effect of each predictor. Therefore, identifying and removing such variables is essential for maintaining the accuracy, reliability, and efficiency of data-driven decision-making. To address this problem, Principal Component Analysis (PCA) is employed as a dimensionality-reduction technique that transforms a set of correlated variables into a smaller number of uncorrelated components known as principal components (Syad, 2024). These components capture most of the variance present in the original variables while eliminating redundancy. Consequently, PCA simplifies complex datasets, enhances interpretability, and provides a clearer understanding of the underlying structure of the data. In this regard, the technique plays a crucial role in identifying and eliminating superfluous variables, thereby improving the quality and efficiency of modern statistical and data analyses (Fan *et al.*, 2021)^[3].

Principal component analysis (PCA) plays a vital role especially in information extraction and data simplification. It helps to eliminate noisy and redundant features, focusing the model on the most important patterns and improving performance. Principal component analysis is a widely used technique in predictive modelling like multiple linear regression. It is designed to simplify complex datasets, enhance interpretability and improve the performance of various predictive models. As datasets grow in size and complexity, especially in fields such as finance, healthcare, image processing, and natural language processing, the need for dimensionality reduction becomes increasingly critical. Principal component analysis addresses this need by reducing the number of variables (features) in a dataset while retaining the most meaningful information. When dealing with a lot of variables, they tend to move together in groups as more than one variable might be measuring the same driving principle of the dependent variable. Hence, we often have some sort of redundant information which we can remove and. Principal component analysis identifies this information.

Recent studies have continued to demonstrate the usefulness of PCA as a robust method for dimensionality reduction and data interpretation across a wide range of fields. Yin and Zhou (2024)^[16] revisited PCA's theoretical foundations in high-dimensional spaces, explaining how data normalization and covariance structures affect the interpretation of eigenvalues and eigenvectors. Ustuner (2024)^[15] applied a randomized version of PCA (R-PCA) for hyperspectral image classification. By reducing high-dimensional image data into a smaller set of features while maintaining classification

accuracy, the research by Ustuner (2024)^[15] highlighted PCA's importance in remote sensing and machine learning. Li *et al.* (2024)^[9] explored PCA in big data environments, demonstrating that weighted PCA methods can efficiently compress large multidimensional datasets without losing essential information.

The principal component analysis cannot only help us to balance the trade-off between the number of variables and information loss but also to get a better sense for high-dimensional data and develop a deeper understanding for the driving forces in a data set and thus, it is an important technique. Hence, this research is aimed at demonstrating how PCA simplifies complex data structures, improves computational efficiency, and enhances the reliability of statistical models. The primary goal is to use PCA is to identify the axes in the dataset that contain the most information, allowing for more efficient data analysis and improved performance in multiple linear regression model. The study will identify superfluous variables using correlation matrix, apply Principal Component Analysis (PCA) to the dataset and obtain the standard deviations of the principal components which are used to calculate the variance explained by each component, demonstrate how PCA transforms correlated variables into a smaller number of uncorrelated principal components that retain most of the original information and illustrate how PCA improves the accuracy, interpretability and efficiency of building a multiple linear regression model.

The R Software (R Development Core Team) was used to implement the analysis using a dataset collected through personal interview of 115 people.

2. Materials and Methodology

2.1 Data Collection

The study population in this research are peoples from different areas of the Bayelsa State in Nigeria. The personal interview method of data collection (Hamed Taherdoost, 2021)^[7] was used in this research. The researcher interviewed 115 persons from whom the information is being collected and considered for analysis. The respondents were required to fill their information such as gender, age, weight, height, number of children, status of smoking, status of drinking, region of residence, level of education, income earned, expenses on medicals, type of health facility, nature of job and body mass index (BMI). The attributes of the data are shown on Table 1.

Table 1: Attributes of data

S/No	Variables	Description
1	Age	Age of a person in healthcare treatment.
2	Gender	Gender of the person (<i>e.g.</i> , male or female).
3	Weight	Weight of a person in healthcare treatment.
4	Height	Height of a person in healthcare treatment.
5	Number of children	The number of children or dependents the person has.
6	Status of smoking	Smoking status of the person (<i>e.g.</i> , whether the person is a smoker or non-smoker).
7	Status of drinking	Drinking status of the person (<i>e.g.</i> , whether the person drinks or not).
8	Region of residence	The residence of the person (<i>e.g.</i> urban or rural).
9	Level of education	The level of education of the person (<i>e.g.</i> secondary, Bachelor's degree or Postgraduate degree).
10	Income earned	The estimated amount of income earned in a month in Naira.
11	Expenses on medicals	The estimated amount spent on medicals in a month in Naira.
12	Type of health facility	The type of health facility used by the person (<i>e.g.</i> Government or Private hospital).
13	Nature of job	The nature of job of the person (<i>e.g.</i> Civil servant, Private company or business person).
14	BMI	The body mass index based on the weight and height of the person.

2.2 Multiple Linear Regression Model

Multiple Linear Regression (MLR) is a statistical technique used to model the linear relationship between a dependent variable and two or more independent (or predictor) variables (Kaya Uyanık & Güler, 2013) [8]. The dependent variable represents the outcome of interest, while the independent variables represent the predictors or explanatory factors. Regression models are used to describe how changes in the independent variables influence the dependent variable, and they also serve as tools for prediction, forecasting, and decision-making. The multiple linear regression model specification is given as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_K X_K + \varepsilon$$

Where:

$X_1, X_2, \dots, X_K$ = Independent variables

$\beta_1, \beta_2, \dots, \beta_K$ = Regression coefficients

ε = Error term

Each coefficient β_i represents the expected change in Y due to a one unit change in X_i , while holding other variables constant. The R-squared and Adjusted R-squared measures of how well the model fits the data. R-squared indicates the proportion of the variance in the dependent variable explained by the independent variables. Multicollinearity occurs when two or more independent variables in a regression model are highly correlated.

2.3 Principal Component Analysis

The main use of principal component analysis is to balance the number of variables and the information loss from removing variables (Mishra *et al.*, 2017). To reduce the information loss, we need to determine the importance or amount of information of a variable in our data set. For this, we can use the variance of a variable as a higher variability which indicates a signal while a lower variability indicates noise.

The principal component analysis works by linearly combining the variables in our data set and then transforming the newly derived variables to maximise their variance (that is the importance of representing the dependent variable). The linear combination of the variables is represented in a covariance matrix while the principal components are determined by the Eigenvectors and Eigenvalues of the covariance matrix. The principal component analysis thereby gives a measure of how variables relate to each other (by covariance matrix), the directions in which the data is dispersed (by Eigenvectors) and the relative importance of each direction (by Eigenvalues).

3. Analysis, Results and Findings

3.1 Multiple Linear Regression of the medical expenses on all independent variables

The Multiple Linear Regression method was used to explain the relationship between several independent variables and the dependent variable, namely "medical expenses". The coefficients of parameters of multiple linear regression are shown on Table 2.

Table 2: The coefficients of parameters of the multiple linear regression

Variables	Coefficients
Expenses on medicals (Y)	
Intercept (β_0)	-0.00006
Age (β_{age})	-0.0275
Gender (β_{gender})	0.0402
Weight (β_{weight})	-0.0040
Height (β_{height})	0.00004
Number of children ($\beta_{no\ child}$)	-0.00120
Status of smoking (β_{smoke})	-0.00390
Status of drinking (β_{drink})	0.00160
Region of residence (β_{region})	0.00420
Level of education ($\beta_{education}$)	-0.0039
Income earned (β_{income})	0.03290
Type of health facility (β_{health})	-0.00120
Nature of job (β_{job})	-0.00210
BMI (β_{bmi})	0.00011

The Multiple R-squared value is given as 0.1789. It can be seen that the R-squared indicates the model's ability to explain variations in the data is low. This indicates that only 17.89% of the variance in the dependent variable was explained by the independent variables. The intercept of $\beta_0 = -0.00006$ indicates a negative relationship for the medical expenses when the independent variables are centered around its mean values.

Identifying Superfluous variables using Correlation matrix

A correlation analysis is performed to understand the relationship between the variables in the dataset. Superfluous data often manifested where multiple predictor variables are highly correlated. The lowest correlation (0.01) which is between the height and the level of education. This may be due to the fact that there is no relationship between the height

of a person and the level of education attained. The highest correlation (0.94) is between the BMI and the weight of the respondent.

Also, another highly correlated variables are status of drinking and status of smoking (0.91), region and level of education (0.70) and height and BMI (0.77). Some of these highly correlated variables or irrelevant variables are considered as superfluous or data with little predictive power. In some cases of the highly correlated variables, one of the two variables can be eliminated from the model without losing much information. One of the variables may contain information that is already captured by other variables in the dataset. For instance, the variable BMI and weight or BMI and height are highly correlated with each other, one might be considered superfluous as it doesn't add unique information. Again, the BMI was calculated using the height and weight.

The superfluous variable will have little or no relationship with the dependent variable (medical expenses) or do not influence the model's output significantly.

Again, including superfluous variables can cause overfitting of the data and thereby increasing the complexity of a model.

Table 3: Table of PCA and total variance

Principal Components	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	1.8374	1.8168	1.3394	1.0949	0.95199	0.79703	0.7354
Proportion of Variance	0.2597	0.2539	0.1380	0.0922	0.06971	0.04887	0.0416
Cumulative Proportion	0.2597	0.5136	0.6516	0.7439	0.81355	0.86241	0.9040
Principal Components	PC8	PC9	PC10	PC11	PC12	PC13	
Standard deviation	0.6306	0.6068	0.4696	0.4341	0.2679	0.03415	
Proportion of Variance	0.0306	0.0283	0.0170	0.0145	0.0055	0.00009	
Cumulative Proportion	0.9346	0.9629	0.9799	0.9944	0.9999	1.00000	

The standard deviations of the principal components are used to calculate the variance explained by each component using a cumulative variance threshold which involves setting a specific threshold for the total proportion of variance to be explained and selecting the minimum number of components required to reach that threshold. In this case, we choose 90%. The first seven components have a cumulative explained variance of 90%.

Figure 1 shows a scree plot is a line plot of the eigenvalues (which represent the amount of variance) against the corresponding principal component number. A scree plot is a line plot of the eigenvalues (which represent the amount of variance) against the corresponding principal component number.

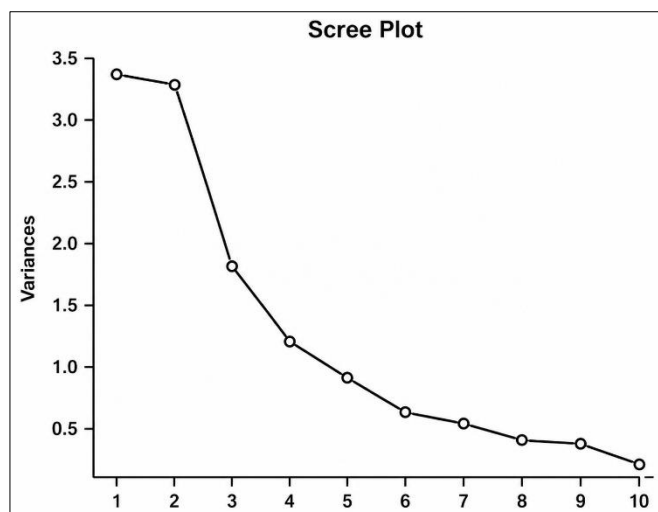


Fig 1: A scree plot is a line plot of the eigenvalues

The plot shows a steep drop, followed by a leveling-off. The elbow rule is determined by locating the elbow point where the slope of the graph distinctly changes or starts to level off. Components to the left of this point are considered significant and are retained. In this case, we choose the first seven components.

3.3 Obtaining the principal components and their scores.

The principal components (loadings or eigenvectors) show how each original variable contributes to each PC. The loadings range from -1 to +1. They indicate the weight of each variable on a Principal component. A high absolute value (close to 1 or -1) indicate that the variable has a strong influence on the component. Value near 0 indicate that the variable has little influence. The sign (+ or -) indicates the

3.2 Performing Principal Component Analysis (PCA)

We perform principal component analysis by applying it to the set of independent variables to create the principal components. The output in R is given in Table 3. Table 3 shows the PCA and the total variances.

direction of the relationship. Appendix A shows the loadings. For instance, PC1 is influenced by four variables, sex (0.4480), weight (0.4035), height (-0.4367) and BMI (0.470). PC2 is influenced by three variables smoking (-0.4117), region (0.4263) and drinking (-0.4178). This suggests a contrast between region and smoking and drinking. All other PC's are explained in similar way.

3.4 Performing Multiple Linear Regression using the principal components

Multiple Linear Regression model was performed using the principal components as independent variables and the dependent variable from the dataset. The output is given in Table 4.

Table 4: The coefficients of parameters of the multiple linear regression of the principal components

Variables	Coefficients
Expenses on medicals (Y)	
Intercept (β_0)	44.58261
PC1	2.11477
PC2	-1.71092
PC3	0.95248
PC4	-0.08371
PC5	-0.58055
PC6	-1.43292
PC7	4.59181

The Multiple R-squared value is given as 0.9559 which indicates the model's ability to explain variations in the data is high. This implies that only 95.59% of the variance in the dependent variable was explained by the independent variables. The intercept of $\beta_0 = 44.5861$ indicates a positive relationship for the medical expenses when the principal component scores variables are centered on its mean values. From the evaluation table above, it can be seen that the Multiple Linear Regression model has a high R-squared (R^2), indicating the model's ability to explain variations in good data. Thus, this model is better at predicting medical costs by combining several independent variables, such as age, gender, body mass index, number of children, smoking habits, and region.

4. Conclusion

In conclusion, Principal Component Analysis (PCA) is a powerful and versatile tool for dealing with superfluous data in predictive models like multiple linear regression. PCA has the ability to identify the most important features in high-

dimensional datasets makes it indispensable in many fields. As data complexity continues to increase, PCA will undoubtedly remain a critical tool in simplifying data, improving model performance, and uncovering valuable insights. This PCA approach is very useful in addressing multicollinearity (high correlation) among original predictor variables and reducing model complexity.

Principal component analysis transforms the original variables that are correlated into new variables that are not correlated with the reduction of the number of variables that have smaller dimensions but can explain most of the diversity of the original variables. It transforms the original data into a new coordinate system, where the axes (called principal components) represent the directions of maximum variance in the data. PCA ensures that only the most significant features are retained for further analysis.

This research recommends that researchers should use PCA as a technique for dealing with variables are highly correlated or convey overlapping information in predictive models like multiple linear regression model which reduces computational complexity and improves the efficiency of the model. Some limitations of this study include the reliance on linear relationships for modelling and hence, an area of further research will be the mathematical modelling of non-linear models using PCA.

References

- Ahmed SS, Yisa EM, Jibrin M, Yahaya G. Principal component analysis for investigating the relationship between semester results and academic performance of students in a polytechnic in Niger State, Nigeria. *J Appl Sci Environ Manage*. 2023;27(10):2355-2359.
- Akbar TA, Javed A, Ullah S, Ullah W, Pervez A, Akbar RA, *et al*. Principal component analysis (PCA)-GIS modeling for groundwater and associated health risks in Abbottabad, Pakistan. *Sustainability*. 2022;14(21):14572.
- Fan C, Chen M, Wang X, Wang J, Huang B. A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data. *Front Energy Res*. 2021;9:652801.
- Gao J, Hu W, Chen Y. Revisiting PCA for time series reduction in temporal dimension. *arXiv*. 2024.
- Greenacre M, Blasius J, Nenadić O. Correspondence analysis and related methods. Boca Raton: Chapman & Hall/CRC; 2022.
- Guo L, Yi Q. An exploration of the application of principal component analysis in big data processing. *Appl Math Nonlinear Sci*. 2024;9(1):1-24.
- Taherdoost H. Data collection methods and tools for research: a step-by-step guide to choose data collection technique for academic and business research projects. *Int J Acad Res Manag*. 2021;10(1):10-38.
- Kaya Uyanık G, Güler N. A study on multiple linear regression analysis. *Procedia Soc Behav Sci*. 2013;106:234-240.
- Li G, Wang X, Liu H, Zhang Q. An exploration of the application of principal component analysis in big data processing. *Appl Math Nonlinear Sci*. 2024;9(4):664-675.
- Li H, Tan Y, Zeng D, Su D, Qiao S. Attitude-predictive control of large-diameter shield tunneling: PCA-SVR machine learning algorithm application in a case study of the Zhuhai Xingye Express Tunnel. *Appl Sci*. 2025;15(4):1880.
- Mishra S, Sarkar U, Taraphder S, Datta S, Swain D, Saikhom R, *et al*. Principal component analysis. *Int. J Livest Res*. 2017;1.
- Mundfrom D, Kay L. The effect of multicollinearity on prediction in regression models. *Gen Linear Model J*. 2018;44:24-28.
- R Development Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2008.
- Hamina S. Data reduction using principal component analysis: theoretical underpinnings and practical applications in public health. *J Contemp Med Educ*. 2024;14(1):1-14.
- Ustuner M. Randomized principal component analysis for hyperspectral image classification. *arXiv*. 2024.
- Yin Y, Zhou W. High-dimensional PCA revisited: insights from general spiked models and data normalization effects. *arXiv*. 2024.

Appendix A

% Results for the loadings of the PCA

```
> Loadings <- my_pca$rotation
> Print (loadings)
```

Variable	PC1	PC2	PC3	PC4	PC5
Age	0.32935807	-0.26646155	0.148341332	-0.01303792	-0.09041638
Sex	0.48015810	-0.19491138	-0.010854265	0.03018828	0.06283531
Weight	0.40353732	0.27318862	-0.002574464	-0.17811725	0.13959057
No child	0.19833573	0.15364257	0.328919461	0.04079870	0.01924632
Smoke	0.06241380	-0.41168624	-0.390739130	0.18521883	-0.21763691
Region	-0.08880636	0.42633499	-0.212524319	0.04613333	-0.18871635
Levedu	-0.15391213	0.37934112	-0.316815801	0.15222940	-0.23418201
Income	-0.04811870	-0.02601745	0.01407097	-0.73608281	-0.51250367
Type of health	0.18855856	0.17148944	0.177116562	0.37728166	-0.68541984
Nature job	0.05637062	0.19635563	-0.592082642	0.05651236	0.23664191
Height	0.43666521	-0.04040364	0.222801374	0.08614976	0.07238330
BMI	0.47073226	0.22591817	-0.081018274	-0.16460972	0.06898188
drink	0.03865440	-0.41781119	-0.364995392	0.15849816	-0.17807069

Variable	PC1	PC2	PC3	PC4	PC5
Age	-0.22316597	0.17513587	0.05264400	-0.09935817	-0.14750199
Sex	0.03357035	0.15710225	-0.27816499	-0.09923096	0.11588817
Weight	0.09990941	0.01180401	-0.35017833	0.17399725	0.20293281
No child	0.76263056	0.17335804	-0.23492548	-0.01066378	-0.03946998
Smoke	0.17460149	-0.01905893	0.06918191	0.13068034	0.20824620
Region	-0.12048692	0.29462254	-0.14243276	0.06491708	-0.40855131
Levedu	-0.05989154	0.31695383	-0.15997379	-0.18874739	0.69511475
Income	0.35544254	0.10340985	-0.07321519	-0.15720996	-0.06346416
Type of health	-0.21351878	-0.26169934	0.26927253	-0.25128662	-0.18481743
Net wealth	0.14832264	0.10531103	0.12785580	-0.53969545	-0.44355899
Height	0.06674351	0.37403777	0.58970201	-0.04231082	0.04002568
BMI	0.06134784	-0.12290308	0.16386076	0.13629720	0.12991696
Drink	0.23107448	-0.02361094	0.20591025	0.29690136	0.13546466

Variable	PC1	PC2	PC3
Age	-0.42354703	0.12720975	-0.00042059
Sex	0.75050072	-0.11989489	-0.00888480
Weight	0.02311312	-0.10532642	-0.56193308
No child	0.02929543	-0.00432393	-0.01537629
Smoke	-0.15645023	-0.71724901	0.00815536
Region	-0.15686476	-0.04748583	-0.00181798
Levedu	0.07312543	0.01493564	-0.00283052
Income	0.14462145	-0.01507766	-0.00978220
Type of health	-0.06552877	0.04016720	-0.00154970
Net wealth	0.07261398	0.07607952	-0.00350630
Height	0.37055689	-0.15113430	0.29807532
BMI	-0.11317906	-0.01172951	0.77128514
Drink	0.14279105	0.64083377	-0.00030158