Application of Principal Component Analysis and Hierarchical Regression Model on Kenya Macroeconomic Indicators

R SIVA GOPAL, B VENKATA RAMANA, D MADHUSUDANA REDDY
PROFESSOR[1,3] ,ASSISTANT PROFESSOR [2]
sivagopal222@gmail.com, bvramana.bv@gmail.com, madhuskd@gmail.com
Department of Mathematics, Sri Venkateswara Institute of Technology,
N.H 44, Hampapuram, Rapthadu, Anantapuramu, Andhra Pradesh 515722

Abstract

This research set out to examine Kenyan macroeconomic factors using a hierarchical regression model and Principal Component Analysis (PCA). A combination of descriptive and correlational research methods was used in the study. From 1970 to 2019, data for the 18 macroeconomic indicators were culled from the Kenya National Bureau of Statistics and the World Bank.For all the data analysis, the R programme was used.In order to lower the data's dimensionality, we used Principal Component Analysis, which reduced the original data set matrix to Eigenvectors and Eigenvalues. To find out whether the extracted components were excellent at forecasting economic development, we fitted them to a hierarchical regression model and used R2 as our measure of success. The study's findings showed that the first component was strongly connected with fifteen original variables and explained 73.605% of the total variance. As an added bonus, the two variables exhibited a larger positive loading into the second main component, which explained around 10.03% of the overall Variance. The third component, which had a strong correlation with only one of the initial factors, explained about 6.22 percent of the total variation. A p-value of 0.0001<5% indicated that the models were significant, while the first, second, and third models had F statistics of 2385.689, 1208.99, and 920.737, respectively. The third model was deemed the most effective predictor due to its 17.296 mean square error. The reliability of Principal Component 1 in describing economic growth was higher than that of other models since it had the largest Variance explained and the lowest p-value. As a result, we may anticipate economic development in Kenya based on the following macroeconomic variables: monetary economy, trade and openness to government operations, consumption, and investment. When working with more than 15 variables, the research suggests using principal component analysis (PCA). To find the partial variance change among the independent variables in regression modelling, the authors suggest using the hierarchical regression model construction approach.

Keywords: Economic growth, principal component analysis, eigenvalues, eigenvectors,, and macroeconomic indicators.

I. Introduction

A. Analysis of the Primary Components
To do Principal Component Analysis (PCA), one must first develop and then analyse a fully working method. This approach entails creating a single set of orthogonal axes, which, when ordered from most to least, identify the main directions of the sample variables [1]. To reduce the number of potential variables, researchers rely on principal component analysis (PCA) to organise variables into categories called principle components. It is crucial to use principal component analysis (PCA) to eliminate duplicate variables [2]. Consequently, principal component analysis (PCA) groups variables that measure the same constructs together after analysing the variables and determining the constructs that each of them measures. Through dimensionality reduction without information loss, the principal component analysis methodology streamlines massive data sets. By generating independent variables, principal component analysis (PCA) achieves its goal of maximising variance. When working with principal component analysis (PCA), eigenvalues and eigenvectors are used to ascertain the variation that variables explain. More than fifteen variables makes it challenging to design a regression model, as stated in [2]. So, principal component analysis (PCA) is the initial step in determining the influence of multiple factors, followed by fitting a regression or other model to the PCs [3]. It is important to be cautious not to over- or under-extract while using PCA. All inferences are derived from the components and related to the original data set, hence it is important that the components kept are accurate representations of the original data set matrix.

Various disciplines of research have published literature on the precise use of principal component analysis (PCA). To evaluate the impact of oil prices on food prices globally, for instance, a research by [4] used PCA. To find out how the macroeconomic index affected food costs, the researchers in the study employed principal component analysis. From 1961 to 2005, global macroeconomic variables including GDP, food production index, consumer price index, and crude oil prices were examined. The research used Scree plots and a fraction of variance (the Kaiser Criterion) to determine the optimal number of common components. The macroeconomic index found a connection coefficient between the consumer price index (0.36) and global gross domestic product (0.87) that varied across different economic variables. To sum up, the element that had the greatest impact on the macroeconomic index was the food production index. However, a correlation between the oil price index and the food production index was found by the researchers. Conversely, there was no discernible effect of oil prices on food prices. Parallel analysis, which helps determine how many components should be kept, was not used in this research, despite its great effectiveness. It is recommended to employ both parallel analysis and the Kaiser criteria when utilising PCA for component extraction and retention. Over and under extraction might be reduced in the long term using this.

Different methods are required to extract principal components in order to reduce the issue of under- or over-extraction. In their analysis of secondary school test data, the researchers in [5] used PCA. Finding out what factors, in terms of specific topics, are most important for the students' performance. The findings showed that there was a lot of agreement across all of the participants, with the first component having the most variation.The English subject was determined to be the most important factor in each student's overall test result. Statistical

tools used in the investigation included Catelli and Kaiser Scree plots. However, because the number of components that should be preserved in this research was never determined using the parallel analysis, the issue of excess and under extraction persists. The small number of variables (less than fifteen) also meant that principal component analysis was not the best tool to use in this investigation. Ultimately, you would get components with a high degree of variance explained if you increase the number of variables since PCA would work better with more variables.
Researchers must exercise caution and use appropriate methodologies in order to extract and preserve the PCs, taking into consideration the data being studied. According to [6], in order to keep all components with an eigenvalue greater than 1, researchers need a sample size of more than 250 observations and an average commonality of more than 0.6 to successfully use Keiser's criteria. A scree plot is recommended as the most successful factor extraction approach in a sample size of more than 300 observations, according to another proposal by [7]. Nevertheless, studies using the Keiser criteria, scree plots, and parallel analysis were contrasted in [8].The most effective strategy was determined by researchers to use parallel analysis to determine how many components to maintain. Following their collection, the major components might be renamed into new variables for use in inferences. The relationship between variables assessed by PCA cannot be well described without employing an alternative model. Thus, a hierarchical regression model comes into being.
Model B: Hierarchical Regression
Behavioural and social statisticians often use multiple linear regression when analysing data. Finding the best predictor is the ultimate goal of any multiple regression study. All the factors that provide credence to a research are located using a regression model. By evaluating the impact of variables beyond the previously input predictors, the hierarchical regression model allows for statistical control and the exploration of incremental validity. Each predictor variable is added to the analysis sequentially in hierarchical regression. The researcher is able to obtain control using the modified coefficient of determination in the hierarchical regression model, which effectively shows the influence of the predictor variables. The sequence in which the predictor variables are entered is determined by a particular theory in hierarchical regression. The researcher also has the option to choose the order in which the variables are entered. Instead than allowing the computer to choose the order of the variables as is done in stepwise regression modelling, researchers should typically choose the order since they know more than the computer (as stated in [9]). When looking at closely related or unrelated predictor variables, the hierarchical regression model is the way to go. It is often used to investigate how a predictor variable affects the result. Coefficients of determination are computed at each step of the study to guarantee quality control [9]. Once all of the primary components have been included into the model, the coefficient of

perseverance at each stage helps to explain the rise in variability.
When compared to the hierarchical regression model, the multiple regression model fails to adequately describe how the predictor variable influences the dependent variable. One example is the study that looked at how several macroeconomic factors affected Pakistan's GDP (refer to reference 10). Principal comment analysis and a multiple regression model were used to get the results of the study. Out of seventeen macroeconomic variables, three were selected for further analysis. A

multiple regression model was used to fit all of the retrieved components, and it was found that each of them affected GDP. Unfortunately, the research could not provide convincing evidence of the influence of any predictor variable. In order to get a better understanding of the connection between the predictor variables, a hierarchical regression model might be used. According to some research, a hierarchical regression model outperforms more traditional forms of analysis in terms of estimate accuracy. As an example, a hierarchical regression model was used to examine the effects of neuroblastoma and numerous paternal occupational exposures on children in a research conducted by [11]. Neuroblastoma risk in children and their fathers' employment histories was the primary focus of the research. In all, 405 sick and 302 healthy individuals participated in the research. Both hierarchical regression and traditional maximum likelihood were used to evaluate the effects of each experience. The total accuracy was much higher when using hierarchical regression as opposed to conventional analysis. Hierarchical regression allowed the researchers to improve estimate accuracy, account for linked exposures, and make certain estimates more accurate using previous information, mitigating some of the drawbacks of the standard technique. De Roos found that among the regression models tested, the hierarchical regression model yielded the best results.

On Kenyan macroeconomic variables, this research used principal component analysis (PCA) and a hierarchical regression model. There were three main goals that we tried to achieve: first, using principal component analysis (PCA) to make Kenya's macroeconomic data more manageable by reducing its dimensionality and classifying it into principal components; second, using these components to fit a hierarchical regression model to economic growth in Kenya; and third, finding the best predictors of economic growth.

II. LITERATURE REVIEW

Analysing the primary components Formed in 1901 [12], PCA was conceived by Karl Pearson. Right now, principal component analysis is used for both constructing prediction models and exploratory data analysis. To get to this conclusion, we break down the auto values of the covariance matrix. For this data analysis, we use a principal component analysis (PCA) on the factor scores [12]. The main goal of principal component analysis (PCA) is to construct a linear combination of the variables that are being studied so that the variance and covariance of a random vector made up of random variables can be explained. The main parts are combinations of linear expressions. Take into account a random vector of interest X' = (X1, X2,..., XN) with a covariance matrix $\sum$ and eigenvalues $\lambda!\geq \lambda"\geq \lambda\#\geq ... \geq \lambda\$\geq 0$. The following are the linear combinations that we have:

$$Y_1 = a_1'X = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p \quad (1)$$
$$Y_p = a_2'X = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p$$
$$\vdots$$
$$Y_p = a_p'X = a_{p11}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p$$

We have $Var(Y_i) = a'_i \sum a_i$ and $Cov(Y_i Y_k) = a'_i \sum a_k$ where $i, k = 1,2,3, \ldots, p$

TheprincipalcomponentsY_1,Y_2,...,Y_p)Should,therefore,captureasmuchinformationas possible.(1)Let ∑be the covariance matrix with the eigenvalue eigenvector pairs $(\lambda_1, \ell_1), (\lambda_2, \ell_2), \ldots, (\lambda_p, \ell_p)$,and$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_n \geq 0$, then the$i$th principal component is given by:

$$Y_1 = \ell_1'X = \ell_{11}X_1 + \ell_{12}X_2 + \ldots + \ell_{1p}X_p \quad (2)$$

for$i$=1,2,3, . . . ,p

It'sworthnotingthatthe varianceofthe$i$thprincipal componentequalsthe$i$th eigenvalue.

$Var(Y_i) = \ell'_i \sum \ell' = \lambda'$and $Cov(Y_i Y_k) = \ell'_i \sum \ell_k = 0$ where$i$=1,2,3, ... ,$p$and $i \neq k$

Linear combinations of random variables produce the primary components. They areuncorrelatedand havevariances equaltothe eigenvaluesofΣ (the covariancematrix);thus,there is no need to make any assumptions regarding multivariate normality distributional assumptions in their construction [13].

Thekth principalcomponent's shareof total variancecan beexpressedas follows:

$$K^{th} = \frac{\lambda_1}{\lambda_1 + \lambda_2 + \ldots + \lambda_p} \quad (3)$$

This is where λ(is the eigenvalue of the kbh PC. With just a little loss in accuracy, k variables may replace the original p variables as the first k PCs can explain the bulk of the variation in population covariance [13]. A Model for Hierarchical Regression When developing their models, advanced regression models often turn to hierarchical regression. As the number of predictors increases, a popular statistical strategy involves building successive linear regression models. We are curious as to whether the dependent variable is better explained by the following model than the one before it [14].If there is a statistically significant difference in R between the two models, we may conclude that the additional factors in the subsequent model provide a better explanation for the dependent variable than the variables in the earlier model. The hierarchical regression model has the benefit of being easily visualised in the degrees of freedom in the majority of statistical programmes. Thus, the results of the hierarchical regression and the statistical significance shown by the regression are accurate and significant. After the study, selecting the best predictor is less of a challenge since, unlike with arbitration, variable entry judgements were made manually based on research [14].

Each of the J-level 2 units in a two-level hierarchical model is given its own level 1 model [14]. Imagine a scenario where the dependent variable is continuous and there is

one continuous predictor, X, at the level 1 level of the model.
Level 1 models are structured as follows:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \bar{X}) + \varepsilon_{ij} \quad (4)$$

In this context, ωij is the predictor variable found on the ith level 1 unit within the jth level 2 unit, **ν**- is the intercept for the jth level 2 unit, Xij is the level 1 predictor or covariate, XA is the grand mean of X', **ν**! is the regression coefficient associated with level 1 predictor X for the jth level 2 unit, and Ɛit is the random error associated with the ith level 1 unit within the jth level 2 unit.
In the level 2 models, we associate all of these regression coefficients (**ν**- and **ν**!), which are included as variables at level 2. The following is the form of a level 2 model that includes a continuous predictor or covariate:

$$B_{0j} = \gamma_{00} + \gamma_{01}W_j + V_{oj}; B_{1j} = \gamma_{10} + \gamma_{11}W_j + V_{1j} \quad (5)$$

where (**ν**-, and **ν**!,) represent the slope and intercept for the jth level 2 unit, γ--and γ-! stand for the overall mean slope and intercept after adjusting for W, W for the level 2 predictor or covariate, γ!- and γ!! for the remaining variables. (X., and X!, respectively) represent the random effects of the jth level 2 unit on the intercept and slope, corrected for W

$$X = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{p1} & \cdots & a_{pn} \end{bmatrix} (P \times N) \quad (7)$$

[14], and are associated with the level 2 predictor W as regression coefficients relative to the level 2 slopes and intercept, respectively. Similar to how level 1 predictors are modelled, level W predictors may be modelled either in their original metric or in relation to their grand mean. The level 2 model is substituted into the level 1 model to create the combined model.

$$B_{0j} = \gamma_{00} + \gamma_{01}W_j + V_{10}(X_{ij} - \bar{X}) + \gamma_{11}W_j(X_{ij} - \bar{X}\ldots) + V_{0j} + V_{1j}(X_{ij} - \bar{X}\ldots) + \varepsilon_{ij} \quad (6)$$

III. METHODOLOGY

A. Study Plan

Several research methods, such as descriptive and correlative designs, were used in the study.

The research variables were described using a descriptive approach.To determine the sort of inherent link between the study's variables, researchers used a correlational research technique [15].

B. Gathering Information

The purpose of this research was to analyse 18 macroeconomic factors from 1970 to 2019. Since it spans more than 30 years, it was determined to be an effective timeframe. A m × m matrix including the macroeconomic variables made up the data set. The World Bank and the Kenya National Bureau of Statistics (KNBS) websites provided the secondary data used in this research. We uploaded the data from Excel sheets into statistical software for analysis.

C. Analysis of the Primary Components

Below is an example of a data matrix X with n columns and p rows that summarises the high-dimensional data set with macroeconomic variables; the rows represent the observations, while the columns include the variables.

The macroeconomic variables are represented by a', where i=1,2,..., n and b=1,2,3,..., k. If the original multivariate dataset is unavailable, PCA may still utilise the covariance matrix to derive PC values. To determine PC, we used the correlation matrix rather than the covariance matrix since the data

set's variables were measured with various units and, hence, had different variances. In order to get the main part, we break down the covariance matrix of the random vector. It is possible to determine the transformed random vector's components by referring to the covariance matrix in relation to the transformed vectors. Covariance matrices of the first standardised variables are used to construct the main components in this case. The main components obtained from the initial variables via the use of the correlation matrix are identical to these. Here is one way to depict the PCA model:

$$\mu_{m\times 1} = W_{m\times d}\, X_{d\times 1} \qquad (8)$$

where the original time series data is projected onto an m-dimensional vector. W is the transpose of E, X is the original matrix, and d is the dimensional datavector ($m<<d$). The major axes, which are m projection vectors that maximise the Variance of u, are derived from the eigenvectors $e1$, $e2$, …, em of the data set covariance matrix S. These eigenvalues correspond to the m biggest non-zero values. $\theta 1$, $\lambda 2$,..., λm. scans of the dataset's coefficients of variation Where μ is the mean vector of x. Solving this system of equations yields the eigenvectors ei:

λi are the eigenvalues of the set S. A primary component's eigenvalues provide the entire variance explained by that component. The total squared component loadings for all items in a given component give rise to this. Eigenvalues that are not zero are desirable, however eigenvalues that are negative are not acceptable since variance cannot be

the eigenvalues are negative and negative. When the eigenvalues are approaching 0, it means that multicollinearity is present since the initial component might take up all of the variance. A component with an eigenvalue greater than one is often considered a factor or main component in a principal component analysis (PCA), as the communality for a single item is 1. A Scree Plot was used to select the components for this study. This plot shows the eigenvalue, or total variance explained, for each component. With the use of a Bi-plot visualisation approach, the dimensionality of these vectors was also visibly decreased. Eigenvectors provide the weight for each eigenvalue. The decision of how many main components to retain was helped by the parallel analysis.
Following their calculation, the eigenvectors are ordered according to the magnitude of their

$$L = \left[\sqrt{\widehat{\lambda_1}\widehat{\varrho_1}}, \sqrt{\widehat{\lambda_2}\widehat{\varrho_2}}, \dots, \sqrt{\widehat{\lambda_m}\widehat{\varrho_m}}\right] \quad (12)$$

respective eigenvalues, as seen in the following. The next step is to choose the top m eigenvectors. The following is the procedure for calculating the PCA projection matrix:

$$W = E^T \quad (11)$$

This is a Wisan $m \times d$matrix, where E has the same eigenvectors as its columns. When principle component analysis (PCA) is used, dimensionality is reduced since the maximum Disparity between the components of the input feature vector devoid of input space transformation. Explanations with a probability greater than 75% were examined in this study. Assumption of uncorrelatedness among the elements in the study is key to orthogonal rotation techniques. The four orthogonal approaches described by Marczyk [16] are "equinox, orthomax, quartimax, and varimax." Alternatively, oblique rotation approaches presuppose a correlation between the components. This research used orthogonal rotation because factors are thought to be uncorrelated. The orthogonal or oblique rotating tools suggested by Marczyk [16] are varimax and Promax, respectively. By minimising variance in loadings within factors and maximising disparity between high and low loadings, varimax rotation optimises a set of factors. Consequently, this study used the varimax approach to create the rotated component matrix. Multiplying the eigenvector by the square root of the eigenvalue yields the component loadings. To find the correlation between each item and its appropriate main component, we utilise the factor loadings that were acquired. To get the communality, add up all the squared loadings of the components. There should be "salient" or "significant" loadings (i.e., 0.30 or higher) [17]. This means that variables with loadings greater than 0.30 are classified into several factors.
The estimated factor loadings matrix, L, is provided by

WhereListhematrixoffactorloadingsandλ'Isit h eigenvalue.

A. HierarchicalRegressionModel

Using a hierarchical regression model, you can see how changing one variable affects another.The general regression model of the study was

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_1 X_i + \varepsilon$$

Thiswasexpanded as;

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \quad (14)$$

Where;

Y

βi=regression coefficientsi =1,2,3,…,nXi =the principal components

ε=errorterm

Thesuccessivehierarchicalmodelswereasfollows;

Model 1 $$Y = \beta_0 + \beta_1 X_1 \quad (15)$$

Model 2 $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \quad (16)$$

.

Model i $$Y = \beta_0 + \beta_1 X_A + \beta_2 X_2 + \cdots + \beta_n X_n \quad (17)$$

In a hierarchical regression model, the main emphasis is on the changes in predictability that are associated with the Principal Components (predictor variables) that are input later in the analysis, as opposed to the predictor variables that are input earlier. The researchers in this study calculated the change in R2 by introducing predictor variables into the analysis in stages. Each component's ability to explain variance dictated the relative importance of the predictor variables. We started by entering the component with the highest Variance, then all the others at the same time, and finally, the component with the lowest Variance. The most relevant statistics were those that showed changes in R2, F, and p-values. One way to find out how much variance each primary component explained was to look at the change in R2. The evaluation of the predictor variables based on their corresponding βs and the structure of the coefficients was given less attention when they were included in the analysis, since the emphasis was on R2 instead of the β or the structure of the coefficients. How much of the variation in the dependent variable can be explained by each model is shown by the coefficient of determination. The coefficient often falls somewhere between zero and one, with the range being 0 to one. An almost-one-to-one coefficient of determination indicates that this component accounts for the majority of the variance. The projected values often end up being rather similar to the actual data values, as pointed out by Trustrum [18] in his discussion of well-fitting regression models. If there are no independent variables that provide useful information, the mean model may be used, which takes the average of all the predicted values and uses it. When compared to the mean model, the suggested regression model ought to provide a more satisfactory match. Three often used statistics in regression analysis for model fit assessment are R-squared, Root Mean Square Error (RMSE), and overall F-test (ANOVA). All three are based on the sum of squares.

Because this research aims to provide explanations and predictions, the F-test was used. Analysis of Variance was used to calculate the F-test. The F-test contrasts two hypotheses: one states that all regression coefficients are equal to zero, and the other states that at least one of the regression coefficients is not equal to zero. "An equivalent null hypothesis" would be an R-squared value of zero [19]. The F-test is a statistical tool for checking the reliability of a hypothesised connection between a predicted variable and its predictors. The observed R-squared is trustworthy and the data does not

generate false findings, according to a significant F-test. A significance threshold, such as 0.05, is used to compare the F-test p-value. We may say that the regression model provides a better fit to the data than the mean model if the p-value is determined to be less than the significance threshold. These results show that the central components of the model, which are the independent variables, enhance the fit.

IV. RESULTSAND DISCUSSION

A. Statistics for Description

To better understand the distribution and behaviour of the predicted variable, descriptive analysis was performed on the predictor variables and trend analysis was applied to the data before any specific analysis was performed.

TABLEI:DESCRIPTIVESTATISTICS

Variables (Natural Logs)	Mean	Std. Dev.	Correlation of Variation	Skewness	Kurtosis
lnGDP	6.884	1.762	0.255	0.633	2.022
lnAQPROD	25.674	2.094	0.081	-0.093	1.754
lnBM	14.949	0.233	0.015	-0.007	2.648
lnCERPROD	3.144	0.234	0.074	0.613	2.536
lnDCPS	21.826	0.855	0.039	0.081	2.203
lnEXGS	17.936	1.791	0.099	0.108	3.119
lnFDINIC	23.254	1.048	0.045	0.325	2.225
lnGGFCE	21.435	0.956	0.044	0.211	2.249
lnGCFC	21.656	1.004	0.046	0.440	2.247
lnGDS	21.082	0.707	0.033	0.078	2.293
lnGNE	23.323	1.087	0.046	0.331	2.164
lnHHSNPISHS	22.896	1.147	0.050	0.312	2.107
lnIMGS	22.087	1.010	0.045	0.241	2.072
lnINF	2.2671	0.667	0.294	-0.429	3.626
lnLINTR	2.7396	0.358	0.131	0.566	2.809
lnECHR	3.4932	1.006	0.288	-0.417	1.503
lnREM	18.848	1.673	0.088	0.051	1.876
lnPOPPTT	17.093	0.448	0.026	-0.214	0.851
lnUNEMP	1.0222	0.024	0.024	-0.571	3.314

Where lnGDP is the natural logarithm of GDP and lnBM is the natural logarithm of broad money times the current LCU. Logarithm of domestic credit to the private sector as a percentage of GDP is equal to lnDCPS. natural log of exports of goods and services times the current US dollar Foreign Direct Investment Net Inflows (FDINIC) = Natural Log of FDI, net inflows (BoP, current US$) the natural logarithm of the general government's ultimate consumption expenditure in current US dollars lnGCFC = Natural Log of Gross Capital Formation(current US$) LnGDS is the natural logarithm of GDP in current US dollars.(in current US dollars) The natural logarithm of total national spending The natural log of household and NPISHs final consumption spending in current US dollars is equal to lnHHSNPISHS. lnIMGS = Natural Log of Goods and Services Imported and Current US Dollars consumer prices (per annum) as a function of the natural logarithm of inflation LnLINTR is equal to the natural logarithm of the lending interest rate expressed as a percentage. Natural log of personal remittances received times the current US dollar amount is equal to lnREM. Natural log of cereal output in metric tonnes is denoted as lnCERPROD. lnAQPROD is the natural logarithm of the metric tonne output of aquaculture.

Natural log of the official exchange rate (LCUperUS$, periodaverage) is equal to lnECHR.Natural logarithm of the total population, lnPOPTT Unemployment rate as a percentage of the work force, natural logarithm The average annual Gross Domestic Product (GDP) from 1970 to 2019 was 23.254 billion USD, according to Table I. The standard deviation was 1.048 USD, showing that GDP varied from one year to the next. During the same time period, FDI averaged 17.936 per year with a standard deviation of 1.791, while inflation averaged 2.267 per annum with a standard deviation of 0.667. The data showed that yearly goods and services imports averaged 22.087 with a standard deviation of 1.011. The distribution exhibited platykurtic behaviour since all the values in Table I were less than 3, when taking Kurtosis values—a measure of a distribution's tail behavior—into consideration. Furthermore, if the Skewness values fall between the range of -3 to 3, the data is deemed to be normal. All of the Skewness scores were within the typical range of -3 to 3, indicating that the data utilised for the study was normal, according to this research. Coefficient of variation (CV), a statistical measure of data points dispersion around the mean, was lastly explored in the research. Relatively significant variation (standard deviation > mean) is indicated by a coefficient of variation larger than 1, but a coefficient of variation less than 1 is considered excellent. The coefficient of variation decreases as the estimate becomes more exact. This allows us to compare the amount of variance across several data sets using the coefficient of variation. Table I shows that the data set showed minimal variance, with all coefficients of variation being less than 1.

Section A: Bartlett's Test of Sphericity and the Sampling Adequacy

TABLEII:KMOANDBARTLETT'STEST

Parameter	Measure	Statistic	Remark
Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.865>0.6	PCA recommended for analysis
Bartlett's Test of Sphericity	Approx. Chi-Square	2470.514	Significance
	Df	153	
	p-value	0.0001<0.05	

(1) Findings from the Examination

The statistical output is as follows: $\chi 2=2470.514$, df =153, $p<0.0001$. Statistical Evaluation The p-value of 0.0001 was lower than the 0.05 threshold of significance, therefore rejecting the null hypothesis that there is no meaningful difference between the variables' correlation matrices and an identity matrix. This meant that the sample correlation matrix did not originate from a population where the correlation matrix is an identity matrix, and that the variables' correlation matrices are substantially different from an identity matrix. It is common practice to propose principal component analysis (PCA) for analysis when the KMO statistics are more than 0.6 and Bartlett's test of sphericity is statistically significant (p-value less than 5%).According to the rule of thumb, principal component Networks

Table III displays the outcomes of an analysis that calculated the iit commonality by squaring the loadings of the iit variables on the 'n' common components.

analysis (PCA) is effective for the variables under investigation when the Kaiser-Meyer-Olkin measure of sample adequacy is larger than 0.7. Table II shows that the correlations matrix met the criteria for component analysis with a KMO score of 0.865, which is higher than the suggested 0.7. Table II's Bartlett's Test was adequate for the investigated data. This is due to the fact that a value of 2470.514 was calculated using Bartlett's Test of Sphericity to compare the variables' correlation matrices with the identity matrices.It was clear from this that a difference existed. Consequently, the identity matrix and the measured variables' correlation matrices were quite different, which was in line with the factorable premise of the matrix.
Section B:

TABLEIII: COMMUNALITIES

Variables (Natural Logs)	Initial	Extraction
lnBM	1.000	0.982
lnDCPS	1.000	0.886
lnEXGS	1.000	0.979
lnFDINIC	1.000	0.664
lnGGFCE	1.000	0.975
lnGCFC	1.000	0.983
lnGDS	1.000	0.860
lnGNE	1.000	0.987
lnHHSNPISHS	1.000	0.985
lnIMGS	1.000	0.982
lnINF	1.000	0.888
lnLINTR	1.000	0.853
lnREM	1.000	0.918
lnCERPROD	1.000	0.685
lnAQPROD	1.000	0.848
lnECHR	1.000	0.932
lnPOPPTT	1.000	0.973
lnUNEMP	1.000	0.793

ExtractionMethod:PrincipalComponentAnalysis.

As all of the macroeconomic variables, from total unemployment to broad money, had communalities higher than 0.65, they followed a similar trend and were strongly connected. Additionally, the strong correlation shows that all of the factors had a significant impact on economic development throughout the research period.

Section A: Charting and Parallel Analysis

Results are shown in Fig. 1 from the screen plot and parallel analysis that helped determine the primary components to be kept.

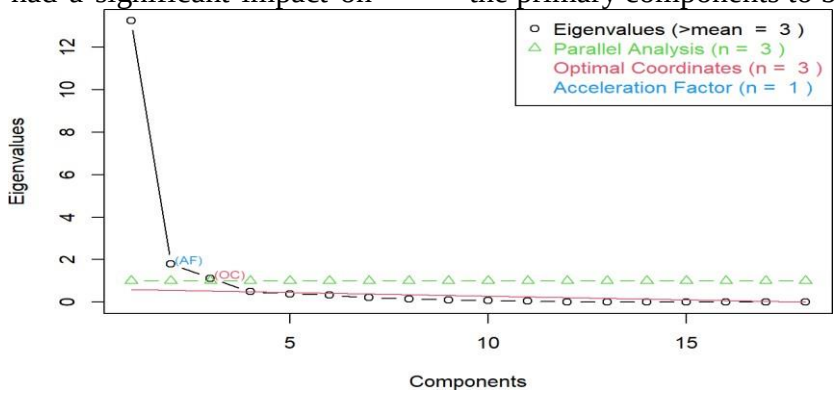

Fig.1.Scree plot.

We used the Scree test to visually inspect the Eigenvalues for inflection spots. After the third principal component (PC), the graph seems to dip, suggesting that the three PCs before it may be accurately summarised to represent the variables in their entirety. Using Eigenvalues, we were able to extract three main components from the data, as shown in the screen picture.Furthermore, in order to prevent over- and under-extraction, we employed parallel analysis.Following the completion of parallel analysis,

Three parts were selected for further analysis. The three variables that were retrieved and kept are representations of the original ones.There are a lot of unique factors in each of the three parts. But the amount of initial

variables varies among components. Using the rotated component matrix, we were able to determine the total number of variables in each component.

Table IV summarises the findings of an analysis of component variation performed using the eigenvalues.

A. The Explanation of Total Variance

TABLEIV:TOTALVARIANCE EXPLAINED

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% Variance	Cumulative %	Total	% of Variance	Cumulative %
1	13.249	73.605	73.605	13.249	73.605	73.605	12.994	72.190	72.190
2	1.806	10.034	83.639	1.806	10.034	83.639	1.957	10.870	83.060
3	1.119	6.217	89.856	1.119	6.217	89.856	1.223	6.796	89.856
4	0.493	2.738	92.594						
5	0.390	2.169	94.763						
6	0.334	1.855	96.618						
7	0.208	1.158	97.776						
8	0.151	0.840	98.616						
9	0.106	0.587	99.203						
10	0.073	0.403	99.606						
11	0.043	0.238	99.843						
12	0.016	0.089	99.932						
13	0.007	0.037	99.969						
14	0.003	0.019	99.988						
15	0.001	0.007	99.995						
16	0.001	0.004	99.999						
17	0.000	0.001	100.000						
18	6.172E-06	3.429E-05	100.000						

The first component accounted for 73.61% of the total variance, according to Table IV, when using the eigenvalue-one criteria. The second component accounted for 10.03% of the total variance. As a percentage of the overall variance, the third, fourth, fifth, and seventh PCs accounted for about 6.217%, 2.74%, 2.17%, 1.86%, and 1.16 percent, correspondingly. Each of the other components accounted for less than 1% of the overall Variance. A declining pattern becomes apparent as we go from one component to the next, with the first component obviously explaining the most variance. On top of that, the first and second components together accounted for around 83.64% of the total variation.Through the utilisation of the orthogonal variation max technique, the

The analysis discovered that the original set of data variables had a summarised total variance of roughly 89.86% per the three preserved components, indicating uncorrelated factor structures.

A. Matrix of Rotated Components

Making decisions is made easier with the help

of the rotating component matrix. It contains the estimated major components and the computed correlations between all of the variables. It includes both the variables' estimated correlations and their computed major components. Each item's loading on each spinning component is indicated in the Rotated Component Matrix, as illustrated in Table V.

TABLEV:ROTATED COMPONENTMATRIX

Variables (Natural Logs)	Component		
	1	2	3
lnBM	0.944		
lnDCPS	0.888		
lnEXGS	0.983		
lnFDINIC	0.797		
lnGGFCE	0.983		
lnGCFC	0.989		
lnGDS	0.889		
lnGNE	0.988		
lnHHSNPISHS	0.985		
lnIMGS	0.985		
lnINF			-0.927
lnLINTR		0.871	
lnREM	0.939		
lnCERPROD	0.818		
lnAQPROD	0.919		
lnECHR	0.831	0.489	
lnPOPPTT	0.940		
lnUNEMP		0.804	0.357

According to Table V, component 1 had a strong relationship with fifteen initial variables. These variables included: total population, broad money, domestic credit to the private sector, gross capital formation, exports of goods and services, general government final consumption expenditure, net inflows of foreign direct investment, personal remittances received, gross domestic savings, gross national expenditure, aquaculture production (metric tonnes), households, and NPISHs. Cereal output (metric tonnes), official exchange rate, final consumption expenditure, and imports of goods and services. The monetary economy, government-sponsored trade and investment, consumption, and investment all share characteristics with this part of the economy. Loan interest rate and total unemployment (as a proportion of the labour force) both loaded positively into the second major component to a greater extent. There is a tight relationship between this part and the monetary and labour economies. Lastly, just one of the original variables—inflation—was substantially associated with the third main component.This part is similar to the economic openness and commerce. A biplot, factor analysis, and variable graph may all be used to visually display this.

Model A: Hierarchical Regression

We ran the predictor variables through the hierarchical regression model to see how they impacted the predicted variable. Table VI summarises the results of fitting three models.

TABLEVI:ESTIMATESOFPARAMETERINMODELS

	Model	Coefficients	Std. Error	T value	Sig.
1	(Constant)	5.086	0.373	13.650	0.000
	PC1	1.039	0.021	48.844	0.000
2	(Constant)	5.281	0.401	13.168	0.000
	PC1	1.033	0.022	47.752	0.000
	PC2	-0.041	0.032	-1.267	0.212
3	(Constant)	5.549	0.388	14.286	0.000
	PC1	1.049	0.021	49.639	0.000
	PC2	-0.035	0.030	-1.135	0.262
	PC3	-0.302	0.111	-2.722	0.009

In order to determine the association between the first component (with 15 original variables),the second component with two variables, the third component with only one original variable, and the GDP (dependent variable), the researcher conducted a hierarchical multiple regression analysisaspresentedinTableVI.AspertheRgeneratedoutput,theequation(Y= $f(PC1,PC2,PC3)$+e)successivelybecomes;

Mode 1 $$Y = 5.086 + 1.039PC_1 \quad (18)$$

Model 2 $$Y = 5.281 + 1.033PC_1 - 0.041PC_2 \quad (19)$$

Model 3 $$Y = 5.549 + 1.049PC_1 - 0.035PC_2 - 0.302PC_3 \quad (20)$$

In (19), the coefficient of the second component was negative and insignificant. Similarly, in (20), the coefficients of the second component were negative and insignificant, while the third component was negative and significant. After the insignificant coefficients were dropped, the retained coefficients formed (21), (22), and (23).

Model 1 $$Y = 5.086 + 1.039PC_1 \quad (21)$$

Model 2 $$Y = 5.281 + 1.033PC_1 \quad (22)$$

Model 3 $$Y = 5.549 + 1.049PC_2 - 0.302PC_3 \quad (23)$$

The three models' intercepts—constant terms—were 5.086, 5.281, and 5.549, correspondingly. According to Table VIII, they exhibited a rising tendency when more components were included into the corresponding models. All three constants were considered significant since the p-values of the derived t-values were 0.000, which was lower than the crucial threshold of 0.05. The GDP figures that are independent of the macroeconomic factors are 5.086, 5.281, and 5.549. Due to a t-value (48.844) and p-value (coefficient of X1 = 1.039, p = 0.000) that were lower than the 0.05 critical value, there was a positive and statistically significant association between the first component with 15 original variables and GDP.With all other factors held constant, this indicates that a one-unit rise in the first major component would result in a 1.039-unit increase in GDP, a substantial effect. The results also showed that after including the second principal component (with two original variables), there was a negative but non-significant relationship between the second component and economic growth, but a positive and statistically significant relationship between the first principal component and GDP.With p-values of 0.00 and 0.212, respectively, and t-values of 47.752 and -1.267, this was confirmed. The first major component's coefficient was 1.033 while the second's was -0.041. That is to say, if the first major component increases by one unit, Kenya's GDP rises by 1.033 units, while if the second component increases by one unit, GDP falls by 0.041 units.
In the end, all three main components were included in the third model according to the original variables.A positive correlation was discovered between the first component and GDP, whereas a negative correlation was seen between the second component and economic growth; nevertheless, the second component's p-value (0.262) was higher than the 5% significance threshold, so the correlation was not deemed statistically significant. The correlation between the GDP and the third component was shown to be statistically significant. Furthermore, the first component was the most important as a one-unit increase in that variable caused Kenya's GDP to grow by 1.049 units. A p-value of 0.262 and a coefficient of -0.035 for the second main component suggest that GDP falls by 0.035 units for every unit rise in the second component, but this effect is not statistically significant.The final component's coefficient was -0.302, and the matching p-value was 0.009, indicating that it was inversely related to GDP. It can be shown that a one-unit increase to the third component would result in a 0.054-unit decrease to GDP.

TABLEVII:GOODNESSOFFITOFMODEL

	Model	Sum of Squares	df	Mean Square	F	Sig.
1	Regression	51.716	1	51.716	2385.689	0.000
	Residual	1.019	47	0.022		
	Total	52.734	48			
2	Regression	51.750	2	25.875	1208.991	0.000
	Residual	0.984	46	0.021		
	Total	52.734	48			
3	Regression	51.889	3	17.296	920.737	0.000
	Residual	0.845	45	0.019		
	Total	52.734	48			

With only the first principal component as an independent variable, the first model in Table VII achieved an R-squared value of 0.98. Principal component 1 (PC1), which includes all fifteen of the original variables, accounts for 98% of the variance in economic growth. The model is statistically significant since the p-value was determined to be 0.0001, which is lower than the significance level of 0.05. It follows that PC 1 is a solid predictor of GDP growth. Model 2 was fitted after the inclusion of one variable, main component 2. Model 2's adjusted R-squared value of 0.981 indicates that PC 1 and PC 2 account for 98.1% of the variance in GDP growth. The adjusted R-squared values were higher in this model compared to model 1, suggesting that the additional variable improves the model. The predictor variables can't be trusted to forecast growth, however, since the model 2 P-value was 0.212, which is more than 0.05 and indicates that the model is insignificant. Model 3, which had all three PCs, was installed after the addition of the third PC. The three PCs accounted for 98.3% of the variance in economic growth, according to the Model 3 adjusted R-squared value of 0.983. Model 3's adjusted R-squared value was much higher than that of models 1 and 2, suggesting that component 3 substantially improves the model. We may trust Model 3 to forecast economic growth since its P-value of 0.009 is less than the significance level of 0.05.

TABLEVIII:ANOVA

With an F-statistic of 2385.689 and a p-value of 0.001, the findings demonstrate that model 1 is significant, since it is less than the crucial

SE	F Change	Sig. F Change
0.147	2385.689	0.000
0.146	1.605	0.212
0.137	7.408	0.009

value of 0.05. The second model is likewise statistically significant; its F-statistic was 1208.991 and its P-value was 0.001, both of which are less than the significance level of 0.05. Another noteworthy model, the third one had a P-value of 0.001 and an F-statistic of 920.737, both below the significance level of 0.05. There is a statistically significant difference between the means of when the p-value is smaller than the crucial value at the 0.05 significance level.

the factors under consideration. Since the p-values for all the models are below the critical limit, the ANOVA results generally show that there are statistically significant differences between some of the means. With an MSE of 51.716 for Model 1, 25.875, and 17.296 for Model 3, the Mean Square Error (MSE) was trending downwards. A better match is indicated by lower MSE values. According to the author, "MSE is a good indicator of how accurately the model predicts the response and it is the most relevant criterion for fit" when the primary goal of the model is prediction. Model three had the lowest mean square error of the three fitted models, indicating that it was the best predictor, while model one had the greatest, as shown in Table VIII.

A. Top Economic Growth Predictors

Table IV shows that the first component, which included 15 of the original variables,

may explain as much as 73.605 percent of the overall variance. The significant level of variance suggests that the 15 initial variables in component one were strongly associated

with each other. Also shown is that the components may be used to provide more accurate predictions when contrasted with the other two components with smaller variance explanation. Model 1, with its independent variable component 1, is able to accurately forecast economic development, as shown in Table VII, as it had the lowest P-value of 0.001 compared to models 2 and 3. Everybody agrees that the first principal component is the strongest predictor of economic growth when using the Total Variance explained and the P-value. Component 1's initial 15 variables are, hence, the most reliable indicators of GDP growth. The monetary elements, trade and openness with government operations, consumption, and investment are the most important economic variables, according to these fifteen variables, which show how the economy grows.

V. CONCLUSIONAND RECOMMENDATION

A. Last Thoughts

We were able to extract and keep three components after using the PCA algorithm. A total of 89.856% of the variance in the original data set was explained by the extracted components. This research effectively used the principal component analysis (PCA) approach, as seen by the high percentage of variance. Based on the largest Variance explained (73.605%), PC 1 was shown to have a substantial influence on economic development among the 15 original variables connected in principle component 1.Households, NPISHs, and wide money were the fifteen microeconomic factors. The official exchange rate, aquaculture production in metric tonnes, imports of goods and services, private sector domestic credit, population total, exports of goods and services, general government final consumption expenditure, net inflows of foreign direct investment, personal remittances received, gross domestic savings, gross national expenditure, and final consumption expenditure. The monetary economy, commerce, and openness to foreign investment are hence macroeconomic factors.

Kenya's economic growth tendency may be mostly explained by the government's operations, consumer spending, and investment.

Section B: Suggestion

It was suggested that when studying more than 15 variables, principal component analysis should be used to group the variables into principal components and reduce their dimensionality. The hierarchical regression model provides better model building techniques by calculating the R-squared change from one model to another, which captures the partial variance change among the independent variables.

References

[1] PearsonK.Principalcomponentsanalysis.TheLondon,Edinburgh,andDublin Philosophical Magazine and Journal of Science. 1901; 6(2): 559.
[2] ZouH,HastieT,TibshiraniR.SparsePrincipalComponentAnalysis.Journalof Computational and Graphical Statistics. 2006; 15(2): 265-286.
[3] CornerS.Choosingtherighttypeofrotationin PCAandEFA.JALTTesting&Evaluation SIG newsletter. 2009;13(5): 38-45.
[4] EsmaeiliA,ShokoohiZ.Assessingtheeffectofoilpriceonworldfoodprices:Application of principal component analysis.
EnergyPolicy.2011;39(2):1022-1025.
[5] NjorogeE,NjorogeG,MuriithiD.EvaluatingSecondarySchoolExaminationResults: Application of Principal Component Analysis. Journal of Statistical and Econometric Methods. 2014; 3(2): 31-46.
[6] BoligonA,VicenteI,VazR,CamposG,SouzaF,CarvalheiroRetal.Principalcomponent analysis of breeding values for growth and reproductive traits and genetic association with adult size in beef cattle1. Journal of Animal Science. 2016; 94(12): 5014-5022.
[7] Field,A.P.DiscoveringstatisticsusingSPSS(2ndedition).London:SAGEPublication, 2005.
[8] GranatoD,SantosJ,EscherG,FerreiraB,MaggioR.Useofprincipalcomponentanalysis (PCA) and hierarchical cluster analysis (HCA) for the multivariate association between bioactivecompoundsandfunctionalpropertiesinfoods:Acriticalperspective.Trendsin Food Science & Technology. 2018; 72: 83-90.
[9] Richardson D, Hamra G, MacLehose R, Cole S, Chu H. Hierarchical Regression for AnalysesofMultipleOutcomes.AmericanJournalofEpidemiology.2015;182(5):459-467.
[10] HussainA,SabirH,KashifM.Impactof macroeconomicvariablesonGDP:Evidencefrom Pakistan. European Journal of Business and Innovation Research. 2016; 4(3): 38-52.
[11] DeRoosA,PooleC,TeschkeK,OlshanA.Anapplicationofhierarchicalregressioninthe investigation of multiple paternal occupational exposures and neuroblastoma in offspring. American Journal of Industrial Medicine. 2001; 39(5): 477-486.
[12] BrownJ.Choosingtherightnumberofcomponentsorfactorsin PCAandEFA.JALT Testing & Evaluation SIG Newsletter. 2009; 13(3): 20-25.
[13] LeverJ,KrzywinskiM,AltmanN.Principalcomponentanalysis.NatureMethods.2017; 14(7): 641-642.
[14] LiuQ,CookN,BergströmA,HsiehC.Atwo-stagehierarchicalregressionmodelfor meta- analysis of epidemiologic nonlinear dose-response data. Computational Statistics & Data Analysis. 2009; 53(12): 4157-4167.
[15] FieldA.AnAdventureinStatistics:TheRealityEnigmaEd.1.London:SAGEPublications, 2016.
[16] MarczykG,DeMatteoD,FestingerD.Essentialsofresearchdesignandmethodology.John Wiley & Sons Inc., 2005.
[17] FanJ,LiaoY,WangW.Projectedprincipalcomponentanalysisinfact

ormodels.The Annals of Statistics. 2016; 44(1): 219.
[18] TrustrumK,FoxJ.RegressionDiagnostics:AnIntroduction.TheStatistician.1993;42(2): 201.
[19] GelmanA,GoodrichB,GabryJ,VehtariA.R-squaredforBayesianRegressionModels. The American Statistician. 2019; 73(3): 307-309.