International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064

Statistical Analysis of Factors that Influence Voter Response Using Factor Analysis and Principal Component Analysis

[1]Violet Omuchira, [2]John Kihoro, [3]Jeremiah Kiingati

Jomo Kenyatta University of Agriculture and Technology, P.O box 62000, Nairobi, Kenya

[2, 3]Department of Statistics, Jomo Kenyatta University of Agriculture and Technology, P.O box 62000, Nairobi, Kenya

Abstract: ***General elections in any country provides an avenue through which citizens exercise their democratic rights in electing leaders of their choice to lead them through a predefined constitutional term in office. Leaders are elected through campaigns in which they demonstrate their intention to lead the populace and they are selected or elected through various modalities that the electorate chooses. The election 2013, for instance provided an avenue through which the electorate chose six levels of leadership starting from the president, governor, Senator, Mp, Women Rep and County Reps. Before the electorate chooses who leads them at whichever level, various factors usually take center stage. These factors range from tribal influence, policies as stipulated in the manifestos, past development record of the contestants etc. This research determined through statistical methods the factors that affect voting patterns and choices that the electorates make. The project adopted random sampling of the JKUAT community via methods of structured questionnaires in which then respondents filled in each category the basis for their choosing a candidate. After field based questionaring, the research used the respondents' samples so collected and performed analysis on the data frame using factor analysis and PCA which is a technique that is used to reduce a large number of variables into fewer numbers of factors. It is also a mathematical tool which can be used to examine a wide range of data sets. This technique extracts maximum common variance from all variables and puts them into a common score. After analysis, the results are the grouping of major factors that affect voting patterns through statistical grouping of study variables. The area of study is JKUAT community located in Juja constituency, of Kiambu County.***

Keywords: Election, voters, community, factor analysis, information, variable

1. Introduction

1.1 Background to the Study

Overtime there has been lack of properly documented data and information on voting patterns and why specifically people vote the way they do and select candidates the way they have done in the past. This research seeks to find out the factors that inform voter decisions on who and who not to elect to a given positions as their representative. The research also wants to have clear guidelines and documented data on election decisions that may be used for future inferences on elections and furthermore give campaigners informed planning while laying their campaign strategies. The research will adopt factor analysis methods and principal component analysis methodologies in analyzing variables as collected in the field.

1.2 Principal Component Analysis (PCA)

PCA starts extracting the maximum variance and puts them into the first factor. After that, it removes that variance explained by the first factors and then starts extracting maximum variance for the second factor. This process goes to the last factor. Eigen values: Eigen values are also called characteristic roots. Eigen values shows variance explained by that particular factor out of the total variance. From the commonality column, we can know how much variance is explained by the first factor out of the total variance. For example, if our first factor explains 68% variance out of the total, this means that 32% variance will be explained by the other factor.

1.3 Factor Analysis

Factor analysis is a technique that is used to reduce a large number of variables into fewer numbers of factors. It is also a mathematical tool which can be used to examine a wide range of data sets. The motivation behind factor analysis is the notion that the data that we observe are a function of some smaller number of unobserved variables called factors. This technique extracts maximum common variance from all variables and puts them into a common score.

1.4 Statement of the Problem

Over time, there has been need to keenly study voter patterns and voting preferences of voters towards what their voting is informed by. To this end, there has been deficiency of information and statistical data collected and analyzed to make conclusions and inferences on why voters vote the way they do and why they choose one individual over the other. Since independence Kenya has held several elections but data on patterns and main factors that influence the voting has not been yet documented. This research aims at solving the problem of data availability by use of statistical analyses of patterns using PCA and factor analysis and have these documented for future inferences.

1.5 Objectives

1.5.1 General objective

The main objective of this research is to come up with a clear picture of factors that determine the choice of political leaders in all the positions by use of factor analysis and PCA.

1.5.2 Specific Objectives

1. To suggest a methodology that will be used in future to get clear reasons of why people vote for the political leaders they chose in Kenya.
2. To establish the factors that affect voting patterns in Kenyan elections.

2. Material and Methods

2.1 Data sets

A structured questionnaire was used which was administered to registered voters, this will help in establishing the parameters to be used in the research. The data collection approach and sampling technique is arrived at after assessing the distribution of the study population. In this random sampling method was used. After data collection, the primary data was analyzed through statistical methods and packages. The package that was used in this research is mainly SPSS. To come up with comprehensive results, the study proposed to adopt factor analysis and PCA methods to analyze the data. The idea behind PCA and factor analysis is to reduce the dimensions of the observation, put the common variance together and come up with common factors and specific factors.

2.2 Method

2.2.1 The Principal Component Analysis (PCA) and Factor Analysis

Eigenvalue is the Criteria for determining the number of factors, According to the Kaiser Criterion; Eigenvalues is a good criterion for determining a factor. If Eigenvalues is greater than one, we should consider that a factor and if Eigenvalues is less than one, then we should not consider that a factor. According to the variance extraction rule, it should be more than 0.7. If variance is less than 0.7, then we should not consider that a factor.

Definition:

For a square matrix A of order n, the number lambda (λ)is an eigenvalue if and only if there exists a non-zero vector x such that

$Ax= \lambda x$

X is called an eigenvector corresponding to λ, and the pair (λ, x) is called the eigenpair for A. Using the matrix multiplication properties, we obtain

$(A- \lambda I_n)x =0.$

This is a linear system for which the matrix coefficient is $A-\lambda I_n$. also the system has one solution if and only if the matrix coefficient is invertible i.e

$Det (A- \lambda I_n) \neq 0$

Here the zero vector is a solution and x is not the zero vector, then we must have

$|A- \lambda I_n| =0$

Therefore PCA which uses eigenvalue to explain the total variance is given by;

Let $X= (X1,X2,\ldots Xj)'$ be a random vector with covariance matrix $\sum$ whose eigenvalues are $\lambda 1 \geq \lambda 2 \geq \ldots\ldots \geq \lambda j \geq 0$

The relationships between the variables are usually driven by underlying latent variables.

Therefore suppose we have two variables x_1 and x_2 and one common factor F. suppose our variables x_1 and x_2 are turned into standardized variables z_1 and z_2. Let F be a common underlying factor and a unique factor for each variable y_1 and y_2.

$Z_{1i} =b_1F_i +u_1y_{1i}$
$Z_{2i} = b_2F_i+u_2y_{2i}$

We assume that both F and y's are standardized and that they are all uncorrelated with one another that is E (F) = 0 and Var (F) =1. The assumption is that the factors are standardized this is because the factors are not observable.
Variance of Z_1 will then be;

$E (Z_1^2) =E (b_1F+u_1y_1)^2$
$=E (b_1^2 F^2+u_1^2y_1^2+2b_1^2u_1^2Fy_1)$
$=b_1^2 E (F^2) +u_1^2 E (y_1^2) +2b_1^2u_1^2 E (Fy_1)$
$Var (Z_1) = b_1^2 Var (F) +u_1^2 Var (y_1) +2b_1^2u_1^2 Covar (Fy_1)$

Assuming that variance is equal to one and covariance zero;
$Var (Z_1) = b_1^2 + u_1^2$
$1 = b_1^2 + u_1^2$

Therefore the variance in variable Z_1 is determined by the contribution of the common factor and the unique factor.

The covariance between Z_1 and the common factor F is;
$Covar (Z_1, F) =E (Z_1, F)$
$=E ((b_1F+ u_1y_1) F)$
$= b_1 E (F^2) + u_1 E (Fy_1)$
$= b_1 Var (F) + u_1 Covar (Fy_1)$
$= b_1$

Since both F and Z_1 are standardized, the covariance is the same as the correlation (r), so the correlation (r) between Z_1 and F is b_1.

Covariance between Z_1 and Z_2 is;
$Covar (Z_1, Z_2) = E (Z_1,Z_2)$
$= E ((b_1F+ u_1y_1) (b_2F+ u_2y_2))$
$=E (b_1b_2F^2+ b_1u_1F y_1+u_1u_2y_2 y_1)$
$= b_1b_2 Var (F) + b_1u_1Covar (Fy_1) + u_1u_2Covar (y_1y_2)$
$= b_1b_2$

Since both Z_1 and Z_2 are standardized, the covariance is the same as the correlation (r), so the correlation (r) between Z_1 and Z_2 is b_1b_2.

$Cor (r_{z1,z2})= b_1b_2$

Correlation matrix between a set of variables is completely determined by their common factors. From the decomposition of the variance in Z_1, we can define the **communality** h_j^2 of each variable as b_j^2, this is the proportion of variance explained by the common factor. The uniqueness of the variables is given by $1- h_j^2$.

Generally for a set of j variables and m factors;
$Z_{ji} =b_{j1}F_{1i} + b_{j2}F_{2i}+ b_{j3}F_{3i}+\ldots+ b_{jm}F_{mi}+ u_jy_{ji}$

Where b's are the **factor loadings** which tell us the correlation coefficient between each factor and the observed variables. Factor loading also shows the variance explained by the variable on that particular factor. In the structure equation modeling (SEM) approach, as a rule of thumb, 0.7 or higher factor loading represents that the factor extracts sufficient variance from that variable. F is the common underlying factor, u is the mean of the standardized variables y's.

In factor analysis model:
Consider a p-dimensional random vector X with mean μ and covariance matrix Var (X) =∑. Then a simple factor analysis model of X in matrix notation is given by
X=QF+μ…….. 1

Where F is the k-dimensional vector of the k-factors. It is often assumed that the factors F are centered, uncorrelated and standardized i.e. E (F) =0 and Var (F) =I_k

We then split the influence of the factors into common factors and specific ones. This leads to the generalized factor model which together with the assumptions constitute the orthogonal factor model given by

$X_{(p\times1)} = Q_{(p\times k)}F_{(k\times1)} + U_{(p\times1)} + \mu_{(p\times1)}$
Where
U_j=mean of j-th specific factor
μ_j=mean of variable j
F_l=l-the common factor
Q_{jl}=loading of the j-th variable on the l-th factor

The random vectors F and U are unobservable and uncorrelated. Using factor analysis I will come up with the set of correlated continuous variables. Factor analysis attempts to identify a small set of factors that represent the underlying relationship among a group of variables. There are three main steps in conducting factor analysis and principal component analysis; assessment of the suitability of the data for factor analysis, factor extraction and factor rotation.

Step 1: Assessment of the suitability of the data for factor analysis
There are two main issues to consider in determining whether a particular data set is suitable for factor analysis: **sample size**, and **the strength of the relationship among the variables** (or items).While there is little agreement among authors concerning how large a sample should be, the recommendation generally is: the larger, the better. In small samples the correlation coefficients among the variables are less reliable, tending to vary from sample to sample. There should also be a ratio of at least five cases for each of the variables. The correlation matrix should also show a correlation of r=0.3 or greater. The Bartlett's test of sphericity should be statistically significant at $p<0.05$, Bartlett's test of sphericity (Bartlett, 1954) and Kaiser-Meyer Olkin value should be 0.6 or above, Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy (Kaiser, 1970, 1974) .The KMO index ranges from 0 to 1, with .6 suggested as the minimum value for a good factor analysis (Tabachnick & Fidell, 2001).

Kaiser-Meyer-Olkin measure of sampling adequacy
KMO determines if the sampling is adequate for analysis (Kaiser 1974a). The KMO compares the observed correlation coefficients to the partial correlation coefficients. Small values for the KMO indicate problems with sampling. A KMO value of 0.9 is best, below 0.50 is unacceptable.

Ant-image correlation matrix shows if there is a low degree of correlation between the variables when the other variables are held constant. Ant- image means low correlation values will produce large numbers. A KMO value less than 0.50 means we should look at the individual measures that are located on the diagonal in the Ant-image matrix. Variables with small values should be considered for elimination.

Sphericity test: is a statistical test for the overall significance of all correlations within a correlation matrix.

Step 2: Factor extraction and Interpretation
Factor extraction involves determining the smallest number of factors that can be used to best represent the interrelations among the set of variables. There are a variety of approaches that can be used to identify (extract) the number of underlying factors or dimensions. Some of the most commonly available extraction techniques are: Principal components, Principal factors, Image factoring, maximum likelihood factoring, Alpha factoring, unweighted least squares, and generalized least squares.

The method I will use is the principal components analysis. It determines how well the factors explain the variation. The goal is to identify the linear combination of variables that account for the greatest amount of common variance. The first factor accounts for the greatest amount of common variance. The factors in the PCA shows individual relationship, much like the beta values in regression. In principal components analysis there are a number of techniques that can be used to assist in the decision concerning the number of factors to retain:

- Kaiser's criterion
- Scree test and
- Parallel analysis.

Kaiser's Criterion/ Eigenvalues Rule
This is one of the most commonly used techniques. Using this rule, only factors with an eigenvalues of 1.0 or more are retained for further investigation. The eigenvalues of a factor represents the amount of the total variance explained by that factor. Kaiser's criterion has been criticized, however, as resulting in the retention of too many factors in some situations.

Eigen value above one (1) represents the number of factors needed to describe the underlying dimensions of the data.

All the factors below eigen value one (1) does not contribute an adequate amount to the model to be included. Each of the factors at this point are not correlated with each other (they are orthogonal as described below).

The Scree test

This is another approach that can be used. It is also Catell's scree test (Catell, 1966) Catell, 1966 since it was first proposed by Catell. This involves plotting each of the eigenvalues of the factors and inspecting the plot to find a point at which the shape of the curve changes direction and becomes horizontal. Catell recommends retaining all factors above the elbow, or break in the plot, as these factors contribute the most to the explanation of the variance in the data set.

Step 3: Factor rotation and interpretation

Once the numbers of factors have been determined, the next step is to interpret them. To assist in this process the factors are 'rotated'. This does not change the underlying solution rather; it presents the pattern of loadings in a manner that is easier to interpret. SPSS does not label or interpret each of the factors but it just shows which variables 'clump together'. From the understanding of the content of the variables (and underlying theory and past research), it is up to the researcher to propose possible interpretations. There are two main approaches to rotation, resulting in either orthogonal (uncorrelated that is covariance is zero) or oblique (correlated) factor solutions. According to Tabachnick and Fidell (2001), orthogonal rotation results in solutions that are easier to interpret and to report; however, they do require the researcher to assume that the underlying constructs are independent (not correlated). Oblique approaches allow for the factors to be correlated, but they are more difficult to interpret, describe and report (Tabachnick & Fidell, 2001, p. 618). In practice, the two approaches (orthogonal and oblique) often result in very similar solutions, particularly when the pattern of correlations among the items is clear (Tabachnick & Fidell, 2001). Many researchers conduct both orthogonal and oblique rotations and then report the clearest and easiest to interpret. Within the two broad categories of rotational approaches there are a number of different rotational techniques provided by SPSS (orthogonal: Varimax, Quartimax, Equamax; oblique: Direct Oblimin, Promax). The most commonly used orthogonal approach is the Varimax method, which attempts to minimize the number of variables that have high loadings on each factor. The most commonly used oblique technique is Direct Oblimin. For a comparison of the characteristics of each of these approaches, see Tabachnick and Fidell (2001, p. 615). Rotation method makes it more reliable to understand the output. Eigen values do not affect the rotation method, but the rotation method affects the Eigenvalues or percentage of variance extracted. From the different rotational techniques I am going to use Varimax in the orthogonal category and Direct Oblimin in the Oblique categories. Each of these can be easily selected in SPSS, and we can compare our variance explained by those particular methods.

Varimax method: is the most common of the rotation methods that are available. This first involves scaling the loadings which maximizes the sum of variances of the squared loadings (squared correlations between variables and factors). This is achieved if any given variable has a high loading on a single factor but near zero loading on the remaining factors also if any given factor is constituted by only a few variables with very high loading on this factor, while the remaining variables have near zero loadings on this factor. If this condition holds, the Varimax rotation brings the loading matrix closer to a simple structure. Scaling the loading, we divide the loading by the corresponding communality as shown;

$b_{jm} = b_{jm}/h_j$,

Here the loading of the j^{th} variable on the m^{th} factor rotation, where h is the communality for the variable j. To find the rotation which maximizes this quantity; the Varimax procedure, as defined below selects the rotation to find this maximum quantity.

V= $\frac{1}{p}\sum_{m-1}^{i}\left\{\sum_{j-i}^{p}(b_{jm})\text{^}4 - \frac{1}{p}((b_{jm})\text{^}2)\text{^}2\right\}$ which is the sample variance of the standardized loadings for each factor, summed over the i factors. Here we find a factor rotation that maximizes this function.

Factor Score

The factor score is also called the component score. Factor score is a composite measure created for each observation on each factor extracted in the factor analysis. The factor weights are used in conjunction with the original variables values to calculate each observation's score. The factor score are standardized according to a z-score. This score is of all row and columns, which can be used as an index of all variables and can be used for further analysis. We can standardize this score by multiplying a common term. With this factor score, whatever analysis we will do, we will assume that all variables will behave as factor scores and will move.

3. Data Analysis

Data was analyzed using Statistical Package for Social Sciences (SPSS) which has inbuilt PCA and factor analysis properties. In factor analysis and PCA, the variable will be analyzed in SPSS to generate factors. Factors with eigen value one and above were considered for more analysis. Using the questionnaire, data analysis will be in three parts; the first part is to check whether the previous Government performed well in intangible services like infrastructures and if the incoming Government will improve the services. The second part is to check the factors that influenced the voters. In step one; we first check whether the data is fit for factor analysis and PCA. Here we check the correlation matrix and most of the values are greater than 0.3. Therefore we proceed to check the KMO and Bartlett's test.

Table 3.1: KMO and Bartlett's test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.84
Bartlett's Test of Sphericity	*Approx. Chi-Square*	390.402
	df	45
	Sig.	0

From the table above KMO is 0.840 and Bartlett's test of sphericity is statistically significant with the value p=0.000. We then concluded that the data is appropriate for PCA and Factor analysis.

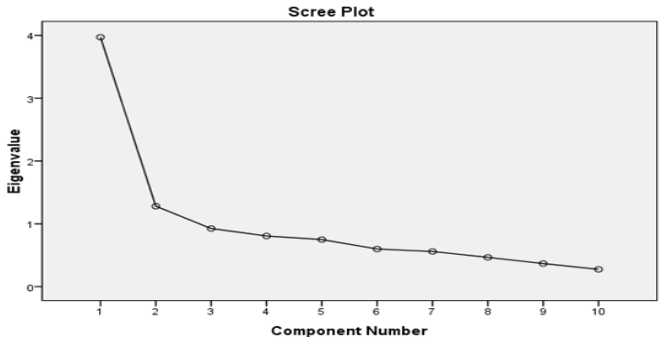
Table 3.2: Total variance explained by the components

Comp--onent	Initial Eigenvalues			Rotation Sums of Squared Loadings		
	Total	*% of Variance*	*Cumulative %*	*Total*	*% of Variance*	*Cumulative %*
1	3.971	39.714	39.714	3.321	33.213	33.213
2	1.28	12.8	52.514	1.93	19.301	52.514
3	0.925	9.252	61.766			
4	0.806	8.064	69.83			
5	0.749	7.491	77.322			
6	0.599	5.992	83.314			
7	0.56	5.598	88.912			
8	0.466	4.659	93.571			
9	0.368	3.678	97.249			
10	0.275	2.751	100			

Extraction Method: Principal Component Analysis.

In factor analysis we only use the components that have an Eigen value of one or more. From the total variance explained table, only the first two components recorded Eigen values above one, which is 3.971 and 1.280. The components explain a total of 52.514 per cent of the variance from the cumulative variance column.

Figure 3.1: Scree Plot



From the scree plot, we look for a change (elbow) in the shape of the plot. The only components above this point are retained. Component one and two explain much more of the variance than the remaining components; we therefore extract two components only.

Table 3.3: Rotated Component Matrix using Varimax Rotation with Kaiser Normalization

Rotated Component Matrix[a]

	Component	
	1	2
G1	0.779	
G2		0.637
G3	0.719	
G4		0.509
G5	0.783	
G6	0.783	
G7		0.708
G8		0.602
G9	-0.459	0.521
G10	0.688	

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 3 iterations.

From the rotation of the data using Varimax rotation, the main loadings on component one are variables G1, G3, G5, G6 and G10. From the questionnaire this items are; the Government will improve the national economy, will improve the country's security, better health services, and better education and will develop a global partnership for development in the country. All these items are positive showing that these are the variables which influenced voters towards choosing leaders. The main loadings of component two are G2, G4, G7, G8 and G9. From the questionnaire these variables are negative in nature, that is; will not improve food security, will be unable to solve negative ethnicity, will not promote gender equality and will not improve infrastructure.

4. Conclusions

From factor analysis and principal component analysis, we first came up with two main components from Kaiser Criterion and scree plot. Using Varimax rotation we have seen that voters usually choose leaders because of the positive intangible services which will improve every human beings life. This method can also bring out the tangible factors that voters usually consider when voting. Such approaches as adopted in this research are of great significance as can be applied across county and national governments in Kenya and Africa as a whole. The scope of this research can be expanded to cover other factors and parameters that determine voter patterns and further be expanded to cover monitoring for voter expectations and whether or not they are fulfilled after the elections.

References

[1] Bartlett, M. S. (1954). A note on the multiplying factors for various chi square approximations. Journal of the Royal Statistical Society, 16 (Series B), 296–298.
[2] Catell, R. B. (1966). The scree test for number of factors. Multivariate Behavioral Research, 1, 245–276
[3] Tabachnick, B. G., & Fidell, L. S. (2001). Using multivariate statistics (4th edn). New York: HarperCollins.
[4] Kaiser, H. (1970). A second generation Little Jiffy. Psychometrika, 35, 401–415.
[5] Kaiser, H. (1974). An index of factorial simplicity. Psychometrika, 39, 31–36.
[6] Bartholomew, D.J, Steele, F, Galbraith, J, Moustaki, I. (2008). Analysis of Multivariate Social Science Data. Statistics in the Social and Behavioral Sciences Series (2nd ed.). Taylor & Francis.
[7] Meng, J. (2011). "Uncover cooperative gene regulations by microRNAs and transcription factors in glioblastoma using a nonnegative hybrid factor model". International Conference on Acoustics, Speech and Signal Processing.
[8] Brown, J. D (16 April 2012.) "Principal components analysis and exploratory factor analysis – Definitions, differences and choices. Shiken: JALT Testing & Evaluation SIG Newsletter.
[9] MacCallum, Robert (June 1983). "A comparison of factor analysis programs in SPSS, BMDP, and SAS". Psychometrika **48** (48)
[10] Bryant, F. B and Yarnold, P. R (1995) Principal components analysis and exploratory and confirmatory factor analysis. Washington, DC: American Psychological Association

[11] L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding multivariate analysis*. Washington, DC: American Psychological Association.
[12] Fabrigar, L. R, Wegener, D. T, MacCallum, R. C. & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272-299.
[13] Hair, J. F., Jr., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate data analysis with readings* (4th ed.). Upper Saddle River, NJ: Prentice-Hall.
[14] Hutcheson, G. and Sofroniou, N. (1999). *The multivariate social scientist: Introductory statistics using generalized linear models*. Thousand Oaks, CA: Sage Publications.
[15] Kim, J. -O., & Mueller, C. W. (1978a). *Introduction to factor analysis: What it is and how to do it*. Newbury Park, CA: Sage Publications.
[16] Pett, M. A, Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. Thousand Oaks, CA: Sage Publications