STATISTICAL ANALYSIS OF STROKE PATIENTS

PROJECT SUBMITTED TO THE UNIVERSITY OF CALICUT IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF THE DEGREEOF MASTER OF SCIENCE IN STATISTICS

HANEENA V P

(REGISTER No**: DVAVMST010)**

DEPARTMENT OF STATISTICS

ST.JOSEPH'S COLLEGE (AUTONOMOUS)

DEVAGIRI, CALICUT-8

JUNE 2023

DEPARTMENT OF STATISTICS

ST.JOSEPHS COLLEGE (AUTONOMOUS)

DEVAGIRI,CALICUT

(AUTONOMOUS) DEVAGIRI, CALICUT

CERTIFICATE

Certified that the project entitled **“Statistical Analysis of Stroke patients”** is a bonafide work done by **HANEENA V P** in partial fulfillment of the requirement of the award of Master of Science in Statistics Degree from University of Calicut.

Mrs.Ashwany Muralidaran M
Lecture
Dept. of Statistics
St. Joseph’s College (Autonomous)
Devagiri, Calicut

Calicut

June 2023

DECLARATION

I, HANEENA V P hereby declare that this project entitled “STATISTICAL ANALYSIS OF STROKE PATIENTS” submitted to St Joseph’s College (Autonomous) Devagiri, affiliated to the University of Calicut in partial fulfillment for the award of Master of Science is a bonafide record of the work carried out by me during 2021-2023, under the supervision and guidance of Mrs. Aswany Muralidaran.

Devagiri,Calicut

HANEENA V P

Register no:DVAVMST010

Department of Statistics

St.Joseph’s College(Autonomous)

ACKNOWLEDGEMENT

First of all, I express my heartfelt gratitude to God the Almighty, the compassionate and merciful, for his grace due to which this project has been made.

It is a matter of great pleasure for me to express a deep sense of gratitude to my supervisor Mrs.ASHWANY MURALIDARAN who enlightens my thoughts through her valuable guidance throughout the period of this dissertation work.

I express my sincere thanks to the Principal, DR. BOBY JOSE, for providingfacilities to me.

I wish to express my special thanks to Dr.ANIL KUMAR, Mr.JOMON JOSE and Mrs.JISHA T for their constant encouragement and suggestions related to the work.

I remember my parents who have always been with me in providing moral and mental support for the completion of this work.

I extend my sincere thanks to all their encouragement and also to the library staff for their timely assistance. I also thank my teachers and friends for their assistance, support and words of goodwill during the period of my dissertati

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

Stroke is a medical condition in which poor blood flow to the brain causes cell death. There are two main type of stroke: ishemic, due to lack of blood fow, and hemorrhagic, due to bleeding. Both cause parts of the brain to stop functioning properly. Signs and symptoms of a stroke may include an inability to move or feel on one side of the body, problems understanding orspeaking, dizziness, or loss of vision to one side. Signs and symptoms often appear soon after the stroke has occured. if symptoms last less than one or two hours, the stroke is a transient ischemic attack (TIA), also called a mini- stoke. A hemorrhagic stroke may also be associated with a severe headache. The symptoms of a stroke can be permanent. Longterm complications may include pneumonia and loss of bladder control.

The main risk factor for stroke is high blood pressure. Other risk factors include tobacco smoking, obesity, high blood cholestrol, diabetes mellitus, a previous TIA, end-stage kidney disease, and atrial fibrillation. An ischemic stroke is typically caused by blockage of a blood vessel, though there are also less common causes. A hemorrhagic stroke is caused by either bleed- ing directly into the brain or into the space between the brain's membranes.

Bleeding may occur due to a ruptured brain aneurysm. Diagnosis is typically based on a physical exam and supported by medical imaging such as a CT scan or MRI scan. A CT scan can rule out bleeding, but may not necessarily rule out ischemia, which early on typically does not show up on a CT scan. Other tests such as an electrocardiogram (ECG) and blood tests are done to determine risk factors and rule out other possible causes. Low blood sugar may cause similar symptoms.

Prevention includes decreasing risk factors, surgery to open up the arteries to the brain in those with problematic carotid narrowing, and warfarin in people with atrial fibrillation. Aspirin or statins may be recommended by physicians for prevention. A stroke or TIA often requires emergency care. An ischemic stroke, if detected within three to four and half hours, may be treatable with a medication that can break down the clot. Some hemorrhagic strokes benefit from surgery. Treatment to attempt recovery of lost function is called stroke rehabilitation, and ideally takes place in a stroke unit; however, these are not available in much of the world.

According to World Health Organisation (WHO), stroke are the second leading cause of death and the third leading cause of disability globally. Stroke is the sudden death of some brain cells due to lack of oxygen when the blood flow to the brain is lost by blockage or rupture of an artery to the brain, it is also a leading cause of dementia and depression.

Stroke is disease that affects the arteries leading to and within the brain. Accordingto the American Stroke Association, itis the No.5 casue of death and a leading cause of disability in the United States. Studies have also shown that there is a higher occurence in blacks than people of biracial origins.

The impact of stroke on people’s lives represents an important challenge for society. In addition to being a sudden event, stroke affects both the individual and family members who are unprepared to deal with the process of rehabilitation or the disabilities that results from this condition.

1.1 Data source and Data Description

The data set used for this study is a secondary data collected from the stroke patients. A population of 5110 people are involved in this study with 2995 females and 2115 males. The data set has been obtained from kaggle (https://www.kaggle.com/dataset) to predict whether a patient is likey to get stroke based on the following attribute information:

1. Id : unique identifier
2. Gender : Male , Female, or Other
3. Age : Age of the patient
4. Hypertension : 0- if patient doesn’t have hypertension, 1- if patient has hypertension
5. Heart_Disease : 0- if patient doesn’t have any heart disease, 1- if patient has a heart disease
6. Ever_Married : ”No” or ”Yes”
7. work_type : ”Children” , ”Govt job”, ”Never_worked”, ”Private” or ”Self-employed”
8. Residence_type : ”Rural” or ”Urban”
9. Avg gluclose_level : Average glucose level in blood
10. bmi : Body mass index
11. Smoking_status : ”Formerly smoked”, ”Never smoked”, ”Smokes”, ”Unknown”

Chapter 2

METHODOLOGY

2.1 Factor Analysis

Factor analysis can be considered as the extension of principal component analysis. both can be viewed as attempts to approximate the covariance matrix.

In 1904, charles spearman first used factor analysis in the field of psychology when he suggested that the performance of school children on a large number of subject was linearly related to a common underlying factor (which he called g, hence g theory) that defined general intelligence. Later Raymond cattel used factor analysis to formulate his famous ten factor model of personality to explain intelligence. He was also a strong beliver in the use of statistical tool and psychometrics to provide the base of theories in psychology rather than just basing them on verbal arguments and discussion. factor analysis is a statistical tool that measures the impact of a few un observed variables called factors on a large number of observed variables .It is used as a data reduction method .It may be used to uncover and establish the cause and effect relationship between variables or to confirm a hypothesis . It is often used to determine a linear relationship between variables before subjecting them to further analysis .principal factor analysis (PCA) is also called common factor analysis and it aim to identify the minimum number of factor analysis include image factoring , Alpha factoring , principal component analysis (PCA) and so on .

Factor analysis operates on the notation that measurable and observable variables can be reduced to fewer latent variables that share a common variance and are unobservable, which is known as reducing dimensionaly . These unobservable factors are not directly measured but are essentially hypothetical constructs that are used to represent variables . For example , scores on a oral presentation and an interview exam could be placed under a factor called communication ability ; in this case , the latter can be inferred from the former butis not directly measured itself .

Factor analysis is used for studies that involve a few or hundreds of variables, items from questionnaires or a battery of tests which can be reduced to a smaller set ,to get an underlying concept ,and to facilitate interpretations . It is easier to focus on some key factors rather than having to consider too many variables that may be trivial , and so factor analysis is useful for plac- ing variables into meaningful categories . Many other uses of factor analysis include data transformation , hypothesis – testing ,mapping and scaling .

To perform a factor analysis ,there has to be univariate and multivariate normality within the data . It is also important that there is an absence of univariate and multivariate outliers . Also a deferring factor is based on the assumption that there is linear relationship between the factors and variables when computing the correlation . The recommended sample size is atleast 300 participants and the variables that are subjected to factor analysis each should have atleast 5 to 10 observation . A larger sample size will diminishthe error in data and so exploratory factor analysis generally works better with larger sample size .

However Guadagnoli and velicer (1988) proposed that if the data set has severalhigh factor loading scores (¿0.80) then a smaller sample size (n¿150) should be sufficient . A factor loading for a variable contribute to the factor . Thus highfactor loading scores indicate that the dimensions of the factors are better accounted for the variables . If your data set containing missing value , cases withmissing value are deleted to prevent over estimation or we can replace it with mean values . Variables that have issue with singularity and multicollinearityshould be removed from data set.

2.2 Orthogonal Factors

Let$X = (X_1, X_2, ..., X_p)'$ be the variable in population with

Mean

$$E(X) = \mu$$

$$= \begin{matrix} \mu_1 \\ \vdots \\ \mu_p \end{matrix}$$

and variance

$$V\,ar(X) = \Sigma$$

$$=\begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & \vdots & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{pmatrix}$$

Orthogonal factor model is

$$Xp\times1-\mu p\times1 =Lp\times1\ Fm\times1+Ep\times1$$

$$m\leq p,\mu=E(X)$$

$$\mathrm{L} = \begin{pmatrix} l_{11} & \cdots & l_{1p} \\ \vdots & \ddots & \vdots \\ l_{p1} & \cdots & l_{pp} \end{pmatrix}$$

Is called the factor loading matrix (which is non random)

$$F = \begin{pmatrix} F_1 \\ \vdots \\ F_p \end{pmatrix}$$

Are called the factors or common factors.

And,

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_p \end{pmatrix}$$

are called the errors or specific errors.

The model is re-expressed as,

$$X_i - \mu_i = \sum_{j=1}^{m} l_{ij}F_j + \epsilon_i, \; i = 1, 2, ..., p$$

and l_{ij} is called the loading of X_i on the factor F_j

The assumptions of the orthogonal models are ;

i $E(F) = 0_{m\times1}$ and $Var(F) = I_m$

ii $E(\epsilon) = 0_{p\times1}$ and $Var(\epsilon) = \psi$, a diagonal matrix with daigonal elements $\psi_1 ... \psi_p$

iii $Cov(F, \epsilon) = 0_{m\times p}$

The unobservable random vectors F and ϵ satisfy the following conditions ;
F and ϵ are independent

$$E(F) = 0$$

and

$$Cov(F) = 1$$

Also

$$E(\epsilon) = 0$$

and

$$Cov(\epsilon) = \psi$$

The orthogonal factor model implies a covariance structure fo X. From the above model

$$\begin{aligned}(X-\mu)(X-\mu)' &= (LF+\epsilon)(LF+\epsilon)' \\ &= (LF+\epsilon)((LF)'+(\epsilon)') \\ &= LF(LF)' + \epsilon(LF)' + LF(\epsilon)' + \epsilon(\epsilon)'\end{aligned}$$

So that,

$$\begin{aligned}\Sigma &= Cov(X)\\ &= E(X-\mu)(X-\mu)'\\ &= LE(F(F)')(L)' + E(\epsilon(F)')(L)' + LE(F(\epsilon)') + E(\epsilon(\epsilon)')\\ &= L(L)'\end{aligned}$$

According to above independent conditions,

$$Cov(\epsilon, F) = E(\epsilon, (F)') = 0$$

Also, by the above model

$$\begin{aligned}(X-\mu)(F)' &= (LF+\epsilon)(F)'\\ &= LF(F)' + \epsilon(F)'\end{aligned}$$

$$\begin{aligned}Cov(X, F) &= E(X-\mu)(F)'\\ &= LE(F(F)' + E(\epsilon(F)')\\ &= L\end{aligned}$$

2.3 Estimation by principal Component Approach

The Principal Component factor analysis of sample covariance matrix Σ is specified in terms of eigen values-eigen vector pairs $(\hat{\lambda}1, \hat{e}_1)(\hat{\lambda}2, \hat{e}_2)$ … $(\hat{\lambda}k, \hat{e}_k)$ where $\lambda\hat{}\,1 \geq \lambda\hat{}\,2 \geq \lambda\hat{}\,3 \geq \ldots \geq \lambda\hat{}\,k$

The decomposition of Σ

Y = Y1 Yp

= e'X-μ

Then,

(X-μ) = eY

= e1,...,epY1 Yp

= j=1mjejYjj + j=m+1pejYj

= 1e1 … … … mem Y11 Ymm + j=m+1pejYj

= LF+ε

where, L = 1e1 … … … mem

F = Y11 Ymm

And ε = j=m+1pe

2.4 Factor Rotation

As remarked in orthogonal factor model is not identifiable up to a rotation of the common factors of factor loading matrix. In other words;

$$X - \mu = LF + \epsilon$$

$$= L^* F^* + \epsilon$$

For

$$L^* = LT$$

$$F^* = T'F$$

where T is any $m \times m$ orthonormal matrix. There, up to a rotation, it is legimate and often desirable to choose a pair (L, F) so that it may achieve better interpretability.

A criterion called varimax criterion can be applied to find an optimal rotation. Let $\hat{L}^*$ be the $p \times m$ rotated factor loading matrix with elements $\tilde{l}^*_{ij}$

Define

$$\hat{l}^*_{ij} = \frac{\hat{l}^*_{ij}}{\hat{h}_i}$$

$$V = \sum_{j=1}^{m} [\frac{1}{p} \sum_{i=1}^{p} (\tilde{l}^*_{ij})^4 - (\frac{1}{p} \sum_{i=1}^{p} (\tilde{l}^*_{ij})^2)^2$$

which is the sum of the column-wise variance of the squares of scaled factor loadings. Find the optimal l^*_{ij} such that V achieves maximum. Then, the optimal rotated factor loading matrix is,

$$l^*_{ij} = \hat{h}_i(\text{optimal}(l^*_{ij})$$

$$/$$

2.5 Types of Factor Analysis

1. Exploratory Factor Analysis

The primary objectives of an Exploratory factor analysis are to determine the number of common factors influencing a set of measure and to evaluate the strength of the relationship between each factor and each observed measure.

- Summarizing data by grouping correlated variables.
- Investigating sets of measured variables related to theoretical constructs.
- Usually done near the onset of research

2. Confirmatory Factor Analysis

The primary objective of a confirmatory factor analysis is to determine the ability of a predefined factor model to fit an observed set of data.

- More advanced technique.
- When factor structure is known or at least theorized.
- Testing generalization of factor structure to new data.

2.6 Types of Factoring

There are different types of methods used to extract the factor from the data set:

Principal Component Analysis

This is the most common method used by researchers used to form uncorrelated linear combinations of the observed variables. 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. PCA is used to obtain the initial factor solution . It can be used when a correlation matrix is singular.

Common Factor Analysis

The second most preferred method by researchers; it extracts the common variance and puts them into factors. This method does not include the unique variance of all variables. This method is used in SEM.

Unweighted Least-Squares Method

A factor extraction method that minimizes the sum of the squared differences between the observed and reproduced correlation matrices (ignoring the diagonals).

Generalized Least-Squares Method

A factor extraction method that minimizes the sum of the squared differences between the observed and reproduced correlation matrices. Correlations are weighted by the inverse of their uniqueness. So that variables with high uniqueness are given less weight than those with low uniqueness.

Maximum likelihood method

A factor extraction method that produces parameter estimates that are most likely to have produced the observed correlation matrix if the sample is from a multivariate normal distribution . The correlations are weighted by the inverse of the uniqueness of the variables , and an iterative algorithm is employed.

Principal Axis Factoring

A method of extracting factors from the original correlation matrix , with squared multiple correlation coefficients placed in the diagonal as initial estimates of the communalities . These factor loadings are used to estimate new communalities from one iteration to the next satisfy the convergence criteria for extraction.

Image factoring

A factor extraction method developed by Guttman and based on image theory. This method is based on correlation matrix. OLS Regression method is used to predict the factor in image factoring. The common part of the variable, called the partial image , is defined as its linear regression on remaining variables , rather than a function of hypothet- ical factors.

Alpha

A factor extraction method that considers the variables in the analysis to be a sample from the universe of potential variables . This method maximizes the alpha reliability of the factors.

2.7 Eigen Values

Eigenvalues is also called characteristic roots. Eigenvalues 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.

2.8 Factor Score

The factor score is also called the component 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 item with this factor score, whatever analysis we will do, we will assume that all variables will behave as factor scores and will move.

2.9 Factor Loadings

Factor loading is basically the correlation coefficient for the variable and fac- tor. Factor loading shows the variance explained by the variable on that particular factor. In the SEM approach, as a rule of thumb, 0.7 or higher factor loading represents that the factor extracts sufficient variance from that variable.

2.10 Criteria for Determining the Number of Factors

According to the Kaiser Criterion, Eigenvalues is a good criteria 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 thsn 0.7,then we should not consider that a factor.

2.11 Rotation Method

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. There are a number of rotation methods available:

i **Orthogonal rotation method**

In this method, axis are maintained at 90 degrees, thus the factors are uncorrelated to each other. In orthogonal rotation, the following three methods are available based on the rotation. The unrotated output maximizes variance accounted for by the first and subsequent factors, and forces the factors to be orthogonal. This data compression comes at the cost of having most items load on the early factors, and usu- ally, of having many items load substantially on more than one factor. Rotation serves to make the output more understandable, by seeking so-called

"Simple Structure" : A pattern of loadings where each item loads strongly on only one of the factors, and much more weekly on the other factors.

ii Varimax rotation method

Used to simplify the column of the factor matrix so that the factor ex- tracts are clearly associated and there should be some separation among the variables helpful to the research purpose . An orthogonal rotation method that minimizes the number of variables that have high load-ings on each factor. This method simplifies the interpretation of the factors. Varimax rotation is an orthogonal rotation of the factor axesto maximize the variance of the squared loadings of a factor (column) on all the variables (rows) in a factor matrix, which has the effect of dif- ferentiating the original variables by extracted factor. Each factor will tendto have either large or small loadings of any particular variable. A varimax solution yields results which make it as easy as possible to identifyeach variable with a single factor. This is a most common rota- tion option. However, the orthogonality (i.e. independence) of factors is often an unrealistic assumption. Oblique rotations are inclusive of orthogonal rotation, and for that reason, oblique rotations are preferred method.

iv Quartimax rotation method

Rows are simplified so that the variable should be loaded on a single factor. A rotation method that minimizes the number of factors needed to explain each variable. This method simplifies the Interpretation of the observed variables. Quartimax rotation is an orthogonal alternative. This type of rotation often generates a general factor on which most variables are loaded to high or medium degree. Such a factor structure is usually not.

v Equimax rotation method

The combination of the above two methods. This method simplifies row and column at a single time. A rotation method that is a combination of the varimax method, which simplifies the factors, and the quartimax method, which simplifies the variables. The number of variables that load highly on a factor and the number of factors needed to explain a variable are minimized. Equimax rotation is a compromise between varimax and quartimax criteria.

vi Direct oblimin method

A method for oblique (nonorthogonal) rotation. When delta equals 0 (the default), solutions are most oblique. As delta becomes more negative, the factors become less oblique. To override the default delta of 0, enter a number less than or equal to 0.8. Direct oblimin rotation is the standard method when one wishes a non-orthogonal (oblique) solution that is, one in which the factors are allowed to be correlated. This will result in higher eigen values but diminished interpretability of factors.

vii Promax rotation method

An oblique rotation, which allows factors to be correlated. This rota- tion can be calculated more quickly than a direct oblimin rotation, so it is useful for large datasets. Promax rotation is an alternative non orthogonal (oblique) rotation method and therefore is sometimes used for very large data sets

2.12 Principal Component Analysis

The principal component analysis consisted of the calculation of eigenvectors and eigenvalues from the covariance matrix of M. Eigenvectors are the vectors of coefficients corresponding to eigenvalues and were used to calcu- late the results. Thus, the coefficients represent the loading factors of each original variable to obtain the new transformed data, and the positive or neg- ative value represents a direct or inverse proportionality, respectively. The eigenvalues represented the variances of each component, so that the first eigenvalues retained the greater part of the variance. Then, the scores were calculated by multiplying the original data (centred in the mean value) by the eigevaectors. Finally, we selected the first and second components to epresent the new data.

Chapter 3

DATA ANALYSIS

3.1 Factor Analysis

The statistical software used in this analysis is IBM SPSS. The software name originally stood for Statistical Package for Social Science (SPSS). It is a software package used for logical batched and non-batched statistical analysis. It can perform standard analyses including descriptive statistics, exploratory data analysis, correlation, a variety of regression, general linear modeling (including ANOVA), time series modeling and analysis, forecasting etc.

We want to reduce the number of variables in the data using factor anal- ysis. That is, here we are going to reduce the 11 variables present in the data into appropriate number of factors by factor analysis method. From the cor- relation matrix of this data we found the eigenvalues and the eigenvectors. Since 11 variables were analysed, the 11 eigenvalues found for the principal components, representing the variance retained by each of them.

3.1.1 Kaiser Meyer Olkin (KMO) and Bartlett’s Test

Kaiser-Meyer-Olkin (KMO) Test measures the strength of relationship among the variables. It is a measure of how suited your data is for factor analysis. The test measures sampling adequacy for each variable in the model and for the complete model. The statistic is a measure of the proportion of variance amongvariables that might be common variance. The lower the proportion, the more suited your data is to factor analysis.

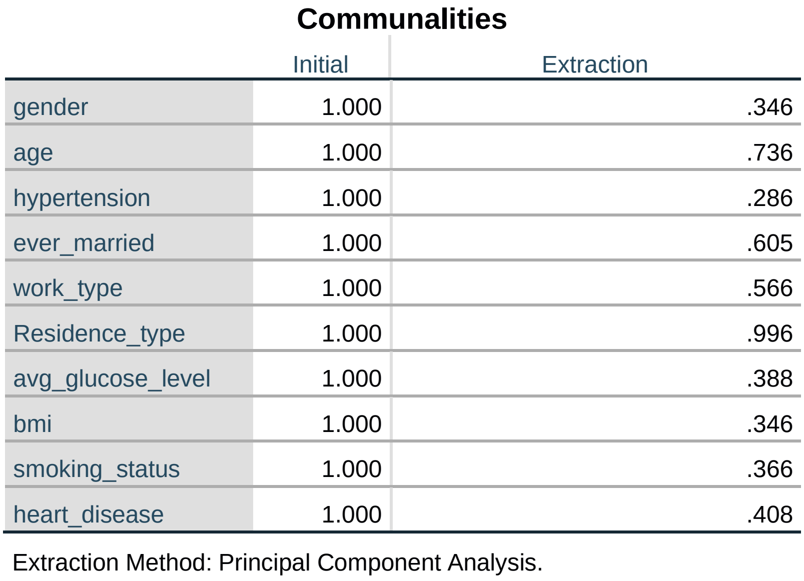
KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.760
Bartlett's Test of Sphericity	Approx. Chi-Square	8072.331
	df	45
	Sig.	.000

The KMO measures the sampling adequacy (which determines if the responses given with the sample are adequate or not) which should be close than 0.5 for a satisfactory factor analysis to proceed. Kaiser (1974) recommend 0.5 (value for KMO) as minimum (barely accepted), values between 0.6-0.8 acceptable, and values above 0.9 are superb. Looking at the table, the KMO measure is 0.760, which is acceptable and therefore factor analysis can be done.

The Bartlett's Test indicates the strong relationship among variables. This teststhe null hypothesis that the correlation matrix is an identity matrix. An identity matrix is matrix in which all of the diagonal elements are 1 and all off diagonal elements are close to 0. We want to reject this null hypothesis. From the same table, we can see that the Bartlett's Test of Sphericity is significant (0.000). That is, significance is less than 0.05. The significance level is small enough to reject the null hypothesis. This means that correlation matrix is not an identity matrix.

3.1.2 Communalities



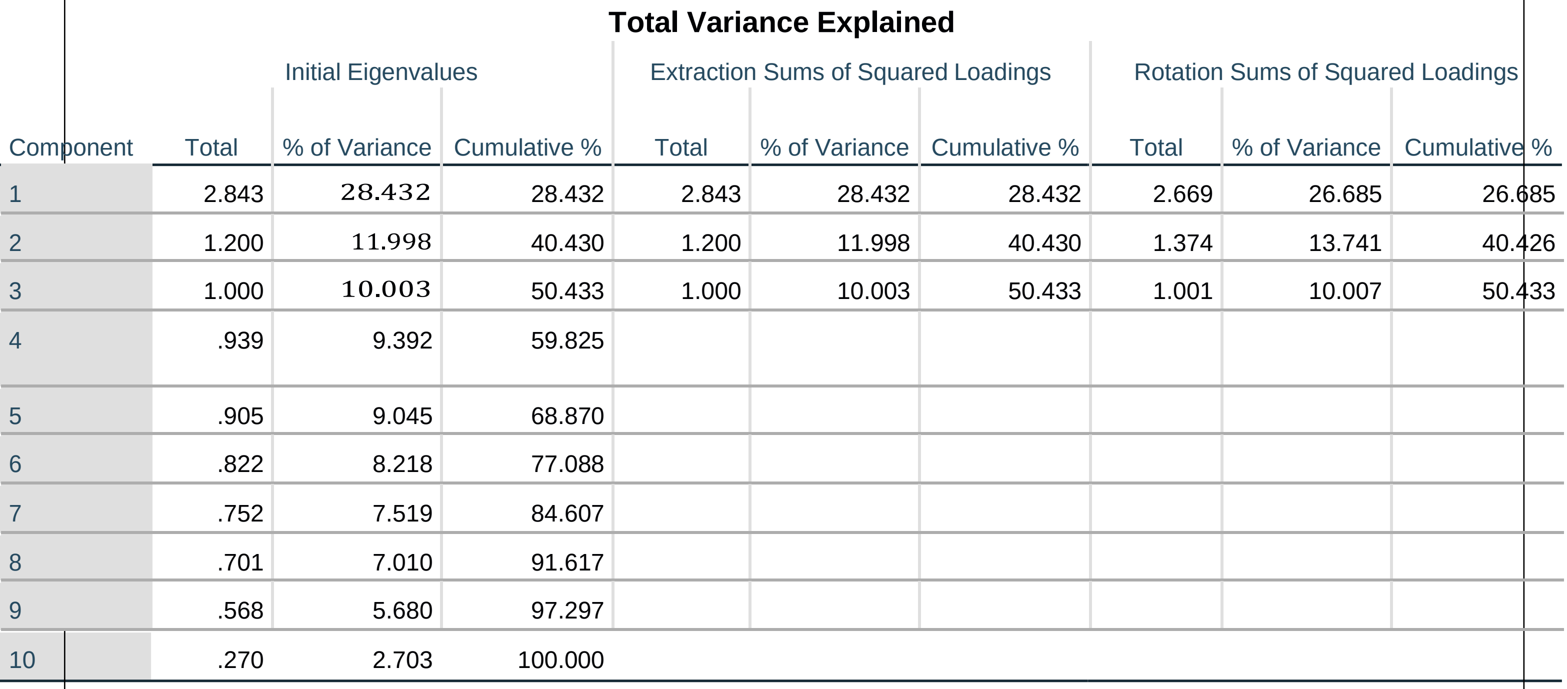
Communalities

	Initial	Extraction
gender	1.000	.346
age	1.000	.736
hypertension	1.000	.286
ever_married	1.000	.605
work_type	1.000	.566
Residence_type	1.000	.996
avg_glucose_level	1.000	.388
bmi	1.000	.346
smoking_status	1.000	.366
heart_disease	1.000	.408

Extraction Method: Principal Component Analysis.

Communalities indicate the amount of variance in each variable that is accounted for. Initial communalities are estimates of the variance in each variable accounted for by all components or factors. For principal compo- nents extraction, this is always equal to 1.0 for correlation analysis.

Extraction communalities are estimates of the variance in each variable accounted for by the components. The communalities in this table are the high values and low values . The high value indicates that the extracted components represent the variables are well.



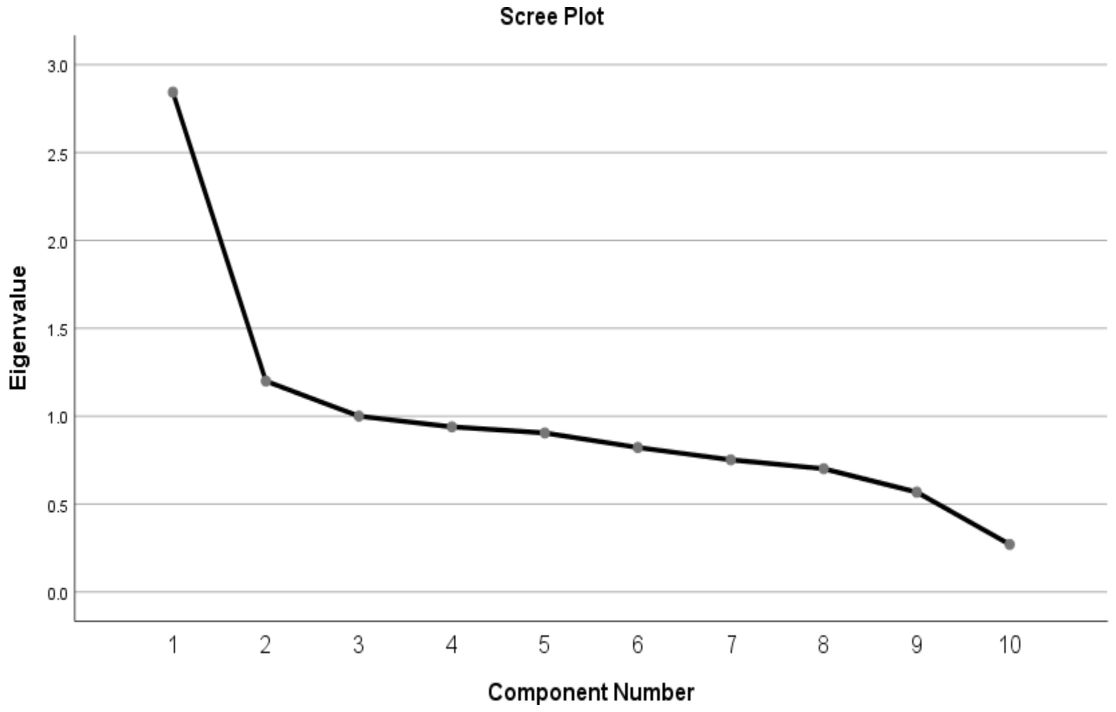
Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.843	28.432	28.432	2.843	28.432	28.432	2.669	26.685	26.685
2	1.200	11.998	40.430	1.200	11.998	40.430	1.374	13.741	40.426
3	1.000	10.003	50.433	1.000	10.003	50.433	1.001	10.007	50.433
4	.939	9.392	59.825						
5	.905	9.045	68.870						
6	.822	8.218	77.088						
7	.752	7.519	84.607						
8	.701	7.010	91.617						
9	.568	5.680	97.297						
10	.270	2.703	100.000						

Eigenvalue actually reflects the number of extracted factors whose sum should be equal to number of items which are subjected to factor analysis. The next item shows all the factors extractable from the analysis along with their eigenvalues.

The Eigenvalue table has been divided into three sub-sections, i.e. Initial Eigen Values, Extracted Sums of Squared Loadings and Rotation of Sums of Squared Loadings. For analysis and interpretation purpose we are only concerned with Extracted Sums of Squared Loadings. Here one should notice that the first factor accounts for 28.432 % of the variance, the second 11.998 % of the variance, and the third 10.003 % of the variance. all the remaining factors are not significant.

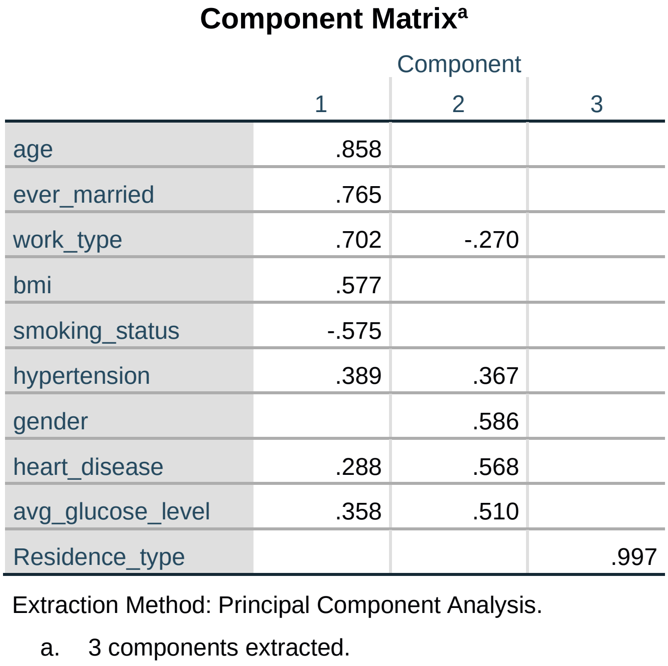
3.1.3 Screen plot



The scree plot is a graph of the eigenvalues against all the factors. The graph is useful for determining how many factors to retain. The point of interest is where the curve starts to flatten. The point where the slope of the curve is clearly levelling off indicates the number of factors that should be generatedby the analysis. It can be seen that factor 4 onwards have an eigenvalue of less than 1, so only three factors have been retained.

3.1.4 Component Matrix

The component Matrix shows the Pearson correlations between the items and the components. For some dumb reason, these correlations are called factor loadings. This table contains component loadings, which are the corre- lations between the variable and component. Because these are correlations, possible values range from -1 to +1.



Component Matrix[a]

	Component		
	1	2	3
age	.858		
ever_married	.765		
work_type	.702	-.270	
bmi	.577		
smoking_status	-.575		
hypertension	.389	.367	
gender		.586	
heart_disease	.288	.568	
avg_glucose_level	.358	.510	
Residence_type			.997

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

from the above table shows the loadings (extracted values of each item un- der 3 variables) of the 10 variables on the three factors extracted. The value of the loadings lie between -1 and 1. The higher the absolute value of the loading, the more the factor contributes to the variable(We have extracted three variables wherein the 10 items are divided into 3 variables according to most important items which are similar responses in component 1 and simultaneously in component 2 and 3). By default SPSS extracts those components with Eigen value greater than 1.

3.1.5 Rotated Component Matrix

The rotated component matrix, helps you to detemine what the components represent. sometimes referred to as the loadings, is the key output of prin- cipal components analysis. It contains estimates of the correlations between each of the variables and the estimated components. Rotation does not ac- tuallychange anything but makes the interpretation of the analysis easier.

Rotated Component Matrix[a]

	Component		
	1	2	3
age	**.808**	.286	
ever_married	**.768**		
work_type	**.752**		
smoking_status	-.604		
bmi	.482		

heart_disease		**.632**	
avg_glucose_level		**.598**	
gender		.474	
hypertension		**.538**	
Residence_type			**.998**

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.

from the above table we can see that, The first component is most highly correlated with Age,evermarried,worktype(Demographic characteristics). The second component is most highly correlated with Heart Disease ,Avg Glucose Level,hypertension(health conditions) . The third component is most highly correlated with Residence Type. This suggest that you can focus on Age, Heart Disease, Avg Glucose level, Residence Type in further analysis.

3.1.6 Component Transformation Matrix

from the table we can see that, by factor analysis we have reduced the dimension to 3 factors.

Component Transformation Matrix

Component	1	2	3
1	.945	.326	.007
2	-.325	.945	-.036
3	-.019	.032	.999

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser
Normalization.

CONCLUSION

In this analysis, the initial system cosisting of 10 variables was reduced to a system consisting of 3 factors using factor analysis. That is we reduce the 10 variables present in the data to 3 factors by using Factor Analysis with a principle components extraction. The variance explained by first factor is 28.432 and by the second factor is 11.998 and by the third factor is 10.003. The total variance explained by these three factors is 50.433 .
so from this analysis I conclude that, the more effecting factors of stroke are Age,evermarried,worktype(demographiccharacteristics),Heart disease ,hypertension, Avg Glucose Level(Health conditions), Residence Type using factor analysis. This is because carrying too much weight increases risk of High blood pressure, Heart disease, High cholesterol which all contribute to higher risk of stroke. Aging is the strongest non modifiable risk factor for stroke.

Bibliography

1. Anderson T.W (1958), “An Introduction to Multivariate Statistical Anal- ysis ”, *Wiley*

2. Richard A. Johnson, Dean W. Wichern (2007), “Applied Multivariate Statistical Analysis”, *Pearson Education*