International Journal of Advanced Statistics and Probability, 5 (1) (2017) 8-16

SPC

International Journal of Advanced Statistics and Probability

Website: www.sciencepubco.com/index.php/IJASP
doi: 10.14419/ijasp.v5i1.6931
Research paper

Identification of correlation structure using rotated factor loadings

Iwok, I. A * and Nwikpe, B. J

Department of Mathematics/Statistics, Faculty of Science, University of Port-Harcourt, Port Harcourt, Nigeria
**Corresponding author E-mail:ibywok@yahoo.com*

Abstract

This work seeks to identify the correlation structure of variables in terms of few underlying but unobservable factors. The method was applied to age and five different tests results obtained from 200 patients in a hospital. Two factors were identified using the scree plot and the Kaiser criterion. The factor loadings obtained by the method of principal components gave an inadequate fit to the data. An algebraic approach was applied using orthogonal rotation, and the loadings were found to give a clear and interpretable pattern. Consequently, the variables: age, fasting blood sugar and diastolic blood pressure were found to cluster about the first factor F_1 called Age-Cardiovascular factor. Similarly, the remaining variables malaria, typhoid and haemoglobin clustered about the second factor F_2 and the given name was Hemo-typhomalaria factor. Diagnostic checks were carried out and the factor model generated by the rotated loadings was found to be adequate.

Keywords: *Factor Loadings; Orthogonal Matrix; Orthogonal Rotation; Principal Component and Communality.*

1. Introduction

The beginning of factor analysis could be traced back to the twentieth century. According to Rechard and Dean [14], Karl Pearson is often considered to have laid the foundation of modern factor analysis. In several decades before the new era, the development of factor analysis was slow due to lack of fast and powerful computing devices. In the wake of recent technological advancement, several sophisticated computational devices were developed to aid in computation and this has renewed the interest in the computation and theoretical aspects of factor analysis. Historically, factor analysis was mostly used in psychology and education. In recent times, however, its use within various disciplines cannot be overemphasised. According to Brett [4], this increase is illustrated in recent surveys of health science electronic databases where articles reporting factor analysis increases by 16,000%.

Giving a set ofq −dimensional interrelated random variables, the whole idea of factor analysis is predicated on determining whether the variables are linearly related to a few underlying, but unobserved random quantities called factors. Factor analysis is a statistical technique that is used to identify a small number of unobserved variables (latent variables) called factors that can be used to represent the relationship among the variables. In factor analysis, latent variables represent unobserved constructs and are referred to as factors or dimensions.

Malaria, typhoid and high blood pressure are devastating diseases and are major cause of morbidity and mortality. They remain the predominant cause of illness and death. Research shows that malaria alone causes an estimated one million deaths annually. It is a life threatening disease transmitted from person to person by the female anopheles mosquito. These problems are hard and fundamental and therefore require proper, radical and continuous attention to avoid further debilitating impact. In addition, Hypertension and diabetes are two of the leading risk factors for atherosclerosis (arterial disease). Atherosclerosis a common arterial disease in which raised areas of degeneration and cholesterol deposits plaques form on the inner surfaces of the arteries obstructing blood flow. This sickness if not treated can result in heart attacks and other related diseases. The intent of this work is to identify how these diseases are interrelated and to regroup the variables under study into a fewer set of clusters based on the shared correlations.

Factor analysis is a branch of multivariate statistics that is used to simplify a large data se tin a way that the relationship between the variables can be easily be depicted, interpreted and understood.

Uneke [16] studied malaria and typhoid in the tropics simultaneously to identify some of the hindrances to effective diagnosis. Using Medline search, the study revealed that some of the factors impeding the effective diagnosis of malaria and typhoid in Nigeria are lack of resources, widespread of self-treatment for clinically suspected malaria and typhoid fever and insufficient access to trained health personnel. The study also showed that there is an appreciable rate of concurrent malaria and typhoid fever.

In an attempt to uncover the relationship between malaria and typhoid fever, Madukosiri [10] studied the illness pattern and relationship between malaria, typhoid fever and other infections. Using difference in mean and correlation, the result of the study showed that the mean of malaria infection was on the increase. The relationship between the illness types showed a positive correlation between malaria and typhoid fever. Malaria and upper respiratory tract infections were also found to be positively correlated.

Oscar and Prasanna [13] studied the relationship between typhoid fever, temperature and malaria and remarked that malaria rises quickly and attains high level while typhoid fever has a pattern that rises slowly during the second and third weeks. In their research, they notice that there exist a relative positive correlation between typhoid fever, malaria and temperature.

Ina et al [7] examined the connection between diabetes and malaria infection. In their study, a total of 495 diabetes patients and 451 hypertensive patients were used. Fasting blood sugar and haemoglobin of each patient were measured. Malaria parasite test was also conducted on each of the patients. The study adopted Mann Whitneyχ^2 and Fisher's exact tests. The result showed that 13 (0.9%) of all participants had malaria parasites at low density and the infected persons had reduced mean haemoglobin. The study also showed evidence for increased risk for malaria infection in patients with type II diabetes.

Haemoglobin (HB) is an iron based organic molecule in red blood cells that transports oxygen and gives blood its red colour. However, Sylvia [15] defines haemoglobin as the life substance of every red blood cell. The study of haemoglobin and other disease is motivated by the fact that every organ of the body depends on oxygen for growth and function.

Femke et al [5] studied the association between haemoglobin, systolic blood pressure and diastolic blood pressure in healthy persons. The study was made up of 101377 whole blood plasma donors who visited a blood bank. Using generalized estimating equations and mixed model, the result of the study revealed that both systolic blood pressure and diastolic blood pressure associates with haemoglobin level positively in healthy individuals. The work of Femke et al [5] clearly showed a relationship between blood pressure and haemoglobin.

Bernard [3] established a relationship between blood pressure and blood sugar. He found that 50% to 80% of patients that were diabetic were also hypertensive. The study further revealed that that diabetes and hypertension are mostly found on the same individual more frequently than would occur by probability.

Hopkins [6] studied the relationship between Haemoglobin and packed cell level of 21 adults within the ages of 40 to 67 using correlation analysis. The results showed a strong positive linear relationship between packed cell volumes and haemoglobin. He noted that haemoglobin and packed cell volume of some university students under study decreased simultaneous. This indicated a positive correlation between these variables.

Ani and Sean [2] highlighted that the relationships between health variables should not only be investigated using correlation analysis. They encouraged the use of factor analysis to identify the underlying structure that generates the observed data. Ani and Sean [2] studied factor analysis using mathematical procedure for the simplification of interrelated measures to discover pattern in a set of variables. They discovered that factor analysis is a better tool for investigating the principles of interaction and integration within the health system.

According to Williams et al [17], two types of factor analysis can be identified; the confirmatory and exploratory factor analysis. According to Aniand Sean [17] exploratory factor analysis (EFA) uncovers complex patterns by exploring the data set , whereas the confirmatory factor analysis (CFA) attempt to confirm hypothesis and uses path analysis to represent variables. The exploratory factor analysis helps in determining the nature and number of latent variables that is responsible for the variation among the observed data.

Alexander [1] applied factor analysis in environmental studies. He made use of nutrient distribution patterns under shrub live-oak in two contrasting soils. The objectives of the research were to identify underlying patterns in soil properties using factor analysis and analyze factor scores to determine how the factor patterns varied between soils, canopy covers, and depth. Factor analysis provided a statistical tool for grouping the 11 correlated soil variables into three uncorrelated factors. Analysis of factor scores allowed independent assessment of soils, shrub cover, depth, and their interactions on soil properties.

Factor analysis of the properties of volcanic soil constituents was first applied by Okuhara et al [12]. The variation of fourteen soil chemical and physical properties of twenty soil samples from Andosols was decomposed into the contributions of seven soil constituents or end-members. The samples were from the slopes of the andesitic Turrialba volcano in Costa Rica. The result showed that Factor analysis of the data explained 98% of the variance by six orthogonal factors.

Michael [11] proposed using factor analysis to environmental data with probabilistic neural networks. He analyzed observation data which consist of environmental factors as the explanatory variables and a population number of a creature (firefly) as the explained variable. The proposed system incorporated probabilistic neural networks which can acquire 60 known nonlinear mapping from the explanatory variables to the explained variable. The proposed system could estimate the effect of the explanatory variables on the explained variable. In other words, the system could solve the inverse problem. To realize the desired environment for the selected creature, the authors showed that the proposed system can suggest an adequate strategy for the controllable explanatory variables.

It is interesting to note that the aforementioned works in the health and environmental sectors in particular have been based oneither correlation analysis or factor analysis with loadings that the values are high and low enough to provide interpretations. However, this work considers a situation where the factor loadings do not give a clear picture of the relationship between the variables and the underlying factors. The work employs the principle of rotation in addressing this set back.

2. Methodology

2.1.The kaiser criterion

This method postulates that only factors whose eigen values are greater than one should be retained. This criterion was proposed by Kaiser [8] and it is probably the most widely used criterion. The idea behind this approach is that any component whose eigen value is more than one account for a meaningful amount of variance while component whose eigen values are less than one account for less variance that had been contributed by one variable.

2.2.Scree test

The Scree test is a graphical method use to obtain significant eigen values of the correlation matrix.In this approach, the eigen value associated with each component is plotted. To obtain the significant values, we look for a break between the component with relatively large eigen values and those with small eigen values. The components that appear before the break are considered meaningful and retained. When several breaks occur, the last big break should be sorted for and the components before the last big break should be retained.

2.3.The orthogonal factor model

Let$X_1, X_2, \dots, X_p$ be a set of mean zero random variables with each variable observed on n subjects. Then, the factor model states that each variableX_i $(i = 1, 2, \dots, p)$ can be expressed as a linear combination of few underlying unobservable random variables $F_1, F_2, \dots, F_m$ called common factors with an accompanied error termε_i associated only withX_i. This can be expressed as:

$$X_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{im}F_m + \varepsilon_i \,; i = 1,2,\dots,p; m < p \quad (1)$$

WhereX_i is theith random variable.

Thea_{ij}'s are unknown regression – type coefficients called factor loadings.

Considering the fact that $X_i(i = 1,2,\dots,p)$is a set of zero – mean variables; then for any observation vectorx_r $(r = 1,2,\dots,n)$, the common factor model (1) can explicitly be written as:

$$X_i - \mu_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{im}F_m + \varepsilon_i \quad (2)$$

Where μ_i is the mean of variable i.

2.3.1.Assumptions of the factor model

1) $\mathrm{E}[\mathrm{F_j}] = 0\ ;\ \mathrm{j} = 1, 2, \ldots, \mathrm{m}$

2) $\mathrm{Var}[\mathrm{F_j}] = 1\ ;\ \mathrm{j} = 1, 2, \ldots, \mathrm{m}$

3) $\mathrm{Cov}(\mathrm{F_j},\ \mathrm{F_k}) = 0\ ;\ \mathrm{j} \neq \mathrm{k}\ ;\ \mathrm{j} = 1, 2, \ldots, \mathrm{m}\ ;\ \mathrm{k} = 1, 2, \ldots, \mathrm{m}$

4) $\mathrm{E}[\varepsilon_\mathrm{i}] = 0\ ;\ \mathrm{i} = 1, 2, \ldots, \mathrm{p}$

5) $\mathrm{Var}[\varepsilon_\mathrm{i}] = \psi_\mathrm{i}\ ;\ \mathrm{i} = 1, 2, \ldots, \mathrm{p}$

6) $\mathrm{Cov}(\varepsilon_\mathrm{i},\ \varepsilon_\mathrm{k}) = 0\ ;\ \mathrm{i} \neq \mathrm{k}\ ;\ \mathrm{i} = 1, 2, \ldots, \mathrm{p}\ ;\ \mathrm{k} = 1, 2, \ldots, \mathrm{p}$

7) $\mathrm{Cov}(\varepsilon_\mathrm{i},\ \mathrm{F_j}) = 0\ ;\ \forall\, \mathrm{i}, \mathrm{j}.$

From (1), the assumptions that $Var[F_j] = 1$, $Var[\varepsilon_i] = \psi_i$,$Cov(F_j, F_k) = 0$ and$Cov(\varepsilon_i, F_j) = 0$yield:

$$Var[X_i] = a_{i1}^2 + a_{i2}^2 + \cdots + a_{im}^2 + \psi_i\ .$$

Whereψ_i is the ith specific variance.
In matrix notation, we consider a random sample$x_1, x_2, \ldots, x_n$from a homogenous population with mean vectorμ and covariancematrix . Let$X$ be a random vector with p components. Then, equation (2) can be expressed as:

$$\underset{(p\times1)}{X-\mu} = \underset{(p\times1)(p\times1)}{AF} + \underset{(p\times1)}{\varepsilon} \qquad (3)$$

Where

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix};\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix};F = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix};\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix};$$

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pm} \end{bmatrix}$$

2.4.Factor loadings

This describes the correlations between the factors and the original variables used in the construction of the factors.

2.5.Estimation of loadings and communalities

2.5.1.The principal component method

Let$x_1, x_2, \ldots, x_n$ be a random sample based on p correlated random variables.
Let$R$ be the sample correlation matrix.
In the factor model, it is assumed that the correlation matrix can be expressed as

$$R = AA' + \psi \qquad (4)$$

Whereψ is the matrix of specific variance given as:

$$\psi = \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \psi_p \end{bmatrix}$$

In this estimation method, we seek for an estimator$\hat{A}$ that will approximate (4).

Let us for now suppose that the specific factors in the model (3) are of minor importance so that ψcan be neglected in (4). Diagonalizing $\hat{R}$ using spectral decomposition;$\hat{R}$ is factored into

$$\hat{R} = \hat{A}\hat{A}' \qquad (5)$$

Where

$$\hat{A}' = VD^{1/2} \qquad (6)$$

is the estimated unrotated factor loading matrix and

$$D = \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sqrt{\lambda_p} \end{bmatrix}$$

is a diagonal matrix with eigen values $\lambda_1, \lambda_2, \ldots, \lambda_p$ of$\hat{R}$on the diagonal.
$V$is the matrix having normalized significant eigen vectors as columns.

2.5.2.Extraction of the factors

However, the goal of factor analysis is to summarize the information contained in the correlation matrix with few factors as possible. Since each eigen value corresponds to a different potential factor, usually only factors with large eigen values are retained. Thus, if the factor analysis is good, the few factors should almost duplicate the correlation matrix.
Now, suppose it is found that the first$m$ factors are large enough to be retained in subsequent analysis; then the dimension of$\hat{A}'$is$(p\times m)$with$m < p$.
The estimated specific variances $\widehat{\psi_i}$ ‘s are provided by the diagonal elements of the matrix$\hat{R} - \hat{A}\hat{A}'$. That is,

$$\widehat{\psi_i} = s_{ii} - \sum_{j=1}^{m} \widehat{a_{ij}^2} \qquad (7)$$

2.5.3.Communalities

The$i$th communality is estimated by

$$\widehat{h_i^2} = \sum_{j=1}^{m} \widehat{a_{ij}^2} = \widehat{a_{i1}^2} + \widehat{a_{i2}^2} + \cdots + \widehat{a_{im}^2}$$

2.5.4.Variance of $\mathrm{X_i}$

The variance of the ith random variable is given by:

$$s_{ii} = \widehat{h_i^2} + \widehat{\psi_i} = \sum_{j=1}^{m} \widehat{a_{ij}^2} + \widehat{\psi_i} = \widehat{a_{i1}^2} + \widehat{a_{i2}^2} + \cdots + \widehat{a_{im}^2} + \widehat{\psi_i} \quad (8)$$

2.6.Rotation

If we consider a rectangular xy -coordinate system in the xy −plane, we can obtain a new $x'y'$-coordinate system if the original$xy - coordinatesystem$ is rotated anticlockwise about their origin through an angleθ .
In algebra, the relationship between such transformed coordinate is given by

$$x' = xcos\theta + ysin\theta \qquad (9)$$

$$y' = -xcos\theta + ysin\theta \qquad (10)$$

and

$$\begin{pmatrix} \mathrm{x} \\ y \end{pmatrix} = \begin{pmatrix} x'cos\theta & - & y'sin\theta \\ x'sin\theta & - & y'cos\theta \end{pmatrix} \qquad (11)$$

Algebraically, the relationship between the coordinates in the two coordinate systems is

$$X = MX' \qquad (12)$$

Where$M$is an orthogonal matrixsatisfying

$$M \quad M' = M'M \quad = 1$$

and

$$M = \begin{pmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{pmatrix}$$

If we consider a rotation as a change from an old basis$B$ to a new basis B', then M is the transformation matrix.

In factor analysis, rotation is a method of altering the initial factor loadings in order to achieve more interpretability while still preserving the essential properties of the initial loadings. Rotation is ordinarily used after extraction of factors to maximize high correlations and minimize low ones.This is usually accomplished by multiplying the unrotated factor loading matrix$\underline{A}$ by a transformation matrix$\underline{M}$to obtain the rotated loading matrix$\underline{A}^*$. That is,

$$\underline{AM} = \underline{A^*}$$

Usually, for a two factor model, the transformation

$$M = \begin{pmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{pmatrix}$$

is used.

3. Diagnostic checks

After fitting the factor model, we need to examine whether the model is adequate or not.One of the ways of checking the adequacy of the model is by examining the behaviour of the residuals matrices. According to Lutkepohl [9], if ρ_{ij} is the true correlation coefficients corresponding to the sample correlations r_{ij}, then we have the following hypothesis test at 5% level to check whether or not a given factor model is adequate.

$$H_0: \rho_{ij} = 0$$

Against

$$H_1: \rho_{ij} \neq 0$$

Decision

RejectH_0 if$\left|\sqrt{N}r_{ij}\ \right| > 2$orEquivalently$\left|r_{ij}\right| > \frac{2}{\sqrt{N}}$

Thus in practical sense, we obtain the residual correlations to be tested and compare their absolute value with $\frac{2}{\sqrt{N}}$

4. Data analysis and results

The data used for this research consist of the medical record of 200 patients upon which five different tests were conducted. The ages of the patient were also recorded.The age and the various tests are considered as variables. The data was obtained from UTH, Nigeria.

The six variables used are: age(X_1), malaria(X_2), typhoid(X_3), fasting blood sugar(X_4), haemoglobin(X_5) and diastolic blood pressure (X_6) . Typhoid was measured at three levels $(60{,}180\ and\ 360\,)$ representing mild, moderate and severe growth respectively. Since the level of malaria is usually recorded as$+, ++, and +++$; the different levels were represented in figures as 1, 2 and 3 respectively. Haemoglobin was measured in grams/decilitre, its normal range is $(13-18) gram/dl.$ Blood pressure was measured using the sphygmomanometer. Only the diastolic blood pressure was considered. The data for the six variables is displayed in table 3 of the appendix.

The variables under consideration have different scales of measurement. Thus, Correlation matrix is therefore appropriate since it is scale invariant. The sample correlation matrix$\hat{R}$ of the six variables is displayed in expression (13) below:

$$\hat{R} = \begin{bmatrix} 1.000 & -0.047 & -.039 & 0.665 & .011 & 0.686 \\ -0.047 & 1.000 & 0.617 & -.004 & -.584 & -.012 \\ -0.039 & 0.617 & 1.000 & -.008 & -.851 & -.123 \\ 0.665 & -.004 & -.008 & 1.000 & .001 & .552 \\ 0.011 & -.584 & -.851 & .001 & 1.000 & .080 \\ 0.686 & -.012 & -.123 & .552 & .080 & 1.000 \end{bmatrix} \quad (13)$$

The eigen values are:

$$\lambda_1 = 0.1457,\ \lambda_2 = 0.2691, \lambda_3 = 0.4223, \lambda_4 = 0.5124, \lambda_5 = 2.2025, \lambda_6 = 2.4480.$$

and the corresponding eigen vectors are:

$$\widehat{e_1} = \begin{pmatrix} .0140 \\ .0641 \\ -.7295 \\ .0307 \\ -.6767 \\ -.0690 \end{pmatrix}, \widehat{e_2} = \begin{pmatrix} .7687 \\ .1913 \\ -.0581 \\ -.3227 \\ .1287 \\ -.4985 \end{pmatrix}, \widehat{e_3} = \begin{pmatrix} -.1868 \\ .4013 \\ -.1410 \\ .6855 \\ .2655 \\ .1492 \end{pmatrix}, \widehat{e_4} = \begin{pmatrix} -.1138 \\ .7232 \\ -.2802 \\ -.3272 \\ .3104 \\ .4267 \end{pmatrix}$$

$$\widehat{e_5} = \begin{pmatrix} .51257 \\ .2871 \\ .3071 \\ .4929 \\ -.3286 \\ .4632 \end{pmatrix}, \widehat{e_6} = \begin{pmatrix} .3135 \\ -.4391 \\ -.5213 \\ .2737 \\ .5042 \\ .3286 \end{pmatrix}$$

The six eigen values and the corresponding eigen vectors above corresponds to six factors representing the six variables. The factor model seeks to determine k such that $k < q$. Several methods exist to achieve this. This research adopted the scree plot testand the Kaiser criterion.

4.1.The scree plot test

This is a plot of eigen values against the factors. The graph is displayed below.

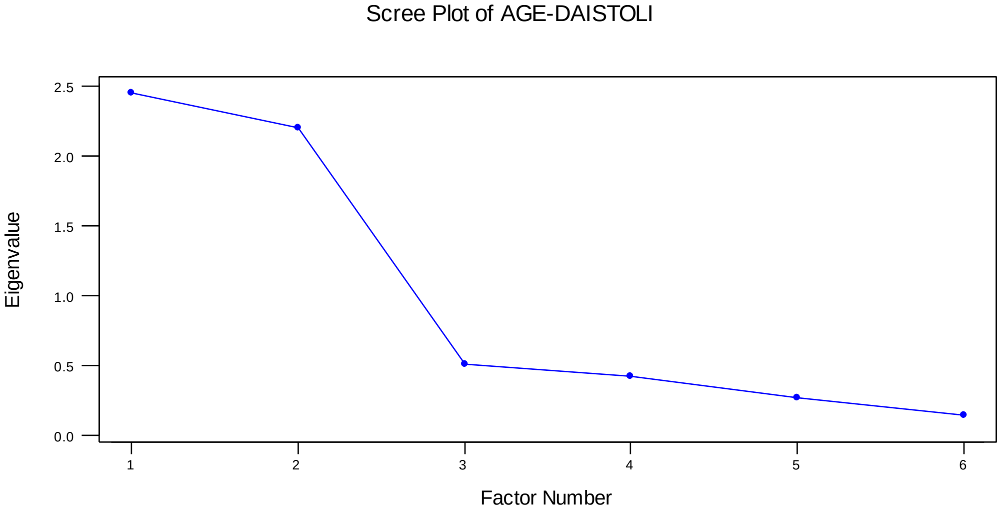

Fig. 1: Scree Plot of the Data.

In figure1, the scree test shows a break between the factors with relative large eigenvalues and those with small eigen values. The factors that appear before the point where the curve makes an elbow (the break) are considered meaningful. This suggest that all the factors that appear from the point of break further should be discarded. By this criterion, only the first two factors are to be retained.

4.2.Kaiser criterion

Using the Kaiser criterion discussed in section 3.1; only factors with eigen values greater than one are retained. Thus, from the list of eigen values; $k = 2$because only the eigen values2.2025 and 2.4480 will be utilized and the corresponding eigen vectors are $\hat{e_5}$and$\hat{e_6}$. This means 78%$[(2.2025\times2.4480/6)\times100]$of the total variation will be explained by the two factors to be extracted.

4.3.Estimation of the factor loadings and communalities

The factor loadings are estimated using equation (6) as shown below:

$$\hat{A} = \begin{pmatrix} .5127 & .3135 \\ .2871 & -.4391 \\ .3071 & -.5213 \\ .4929 & .2737 \\ -.3208 & .5042 \\ .4267 & .3286 \end{pmatrix} \begin{pmatrix} \sqrt{2.2025} & 0 \\ 0 & \sqrt{2.4480} \end{pmatrix} = \begin{pmatrix} .7609 & .4945 \\ .4261 & -.6926 \\ .4558 & -.8222 \\ .7315 & .4317 \\ -.4761 & .7951 \\ .6874 & .5141 \end{pmatrix} \tag{14}$$

Expression (14) gives the loadings of the variables on the factors. In the matrix (14), each row corresponds to each of the variables under study. The communalities were obtained by summing the squares of the rows of the loadings. Table 1 below shows the factor loadings and communalities of the observed variables.

Table1: Factor Loadings of Diseases

Variables	F_1	F_2	Communalities
Age	.7609	.4945	0.8235
Mal.	.4261	-.6926	0.6613
Typh.	.4558	-.8222	0.8838
FBS.	.7315	.4317	0.7215
HB	-.4761	.7951	0.8589
Dias.	.6874	.5141	0.7368
---------	----------	-----	---------
Var	2.447	2.207	4.6858
%Var	.408	.367	0.7810

The specific variances were estimated using equation (7) and is expressed in matrix form as shown below.

$$Cov(\varepsilon) = \hat{\psi} = \begin{bmatrix} .1765 & 0 & 0 & 0 & 0 & 0 \\ 0 & .3344 & 0 & 0 & 0 & 0 \\ 0 & 0 & .1163 & 0 & 0 & 0 \\ 0 & 0 & 0 & .2786 & 0 & 0 \\ 0 & 0 & 0 & 0 & .1414 & 0 \\ 0 & 0 & 0 & 0 & 0 & .2633 \end{bmatrix}$$

Thus, $\hat{A}\widehat{A}' + \hat{\psi}$ will produce a matrix similar to $\hat{R}$ as postulated by the factor model.
Now,

$$\hat{A}\widehat{A}' + \hat{\psi} =$$

$$\begin{bmatrix} 1.000 & -0.018 & -0.060 & 0.770 & 0.031 & 0.777 \\ -0.018 & 1.000 & 0.764 & 0.013 & -0.754 & -0.062 \\ -0.060 & 0.764 & 1.000 & -0.022 & -0.871 & -0.109 \\ 0.770 & 0.013 & -0.022 & 1.000 & -0.005 & 0.725 \\ 0.031 & -0.754 & -0.871 & -0.005 & 1.000 & 0.082 \\ 0.777 & -0.063 & -0.109 & 0.725 & 0.815 & 1.000 \end{bmatrix} \tag{15}$$

Clearly, expression (15) gives a fair approximation of the original correlation matrix $\hat{R}$.

4.4.The vagueness of the unrotated factor loadings

Observantly, table 1 does not give a clear interpretable pattern. To ascertain this fact,
below is the plot of the loadingpairsfor each variable.

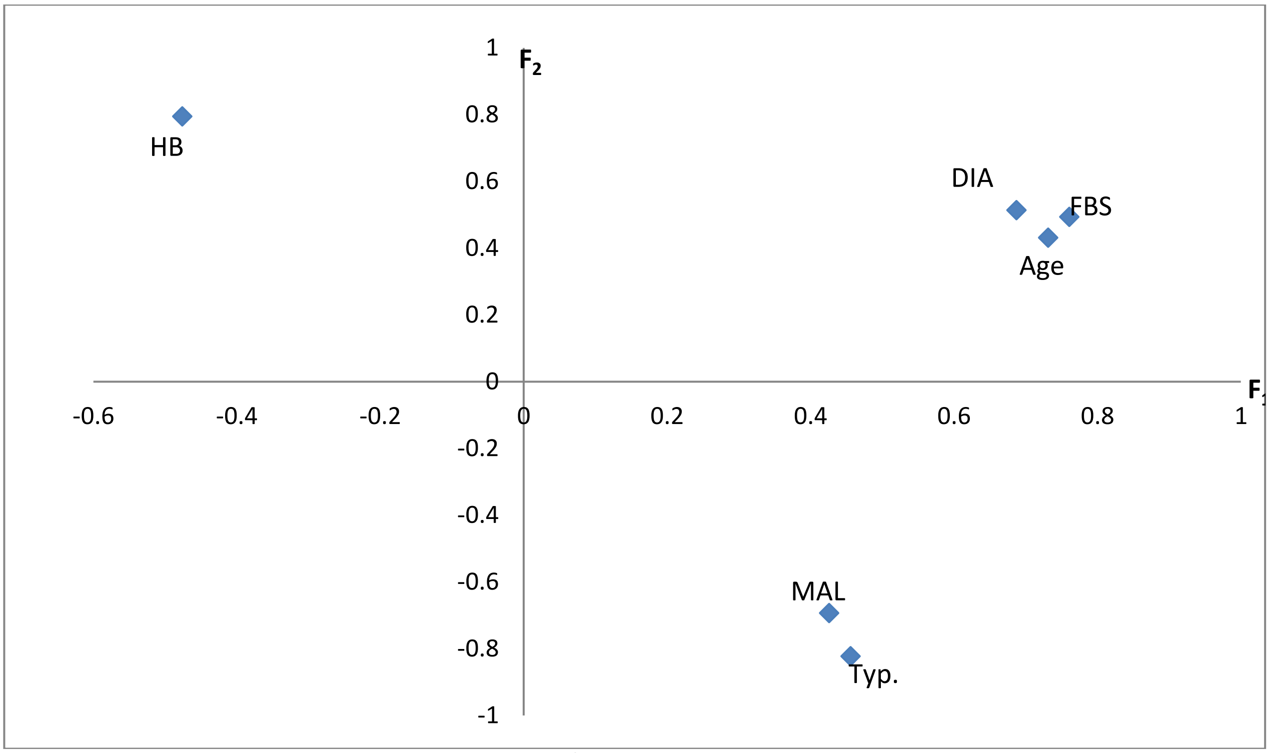

Fig. 2: Plot ofUnrotated Factor Loadings.

In figure 2, the x-axis represent the first factor (F_1) while the y-axis represents the second factor (F_2). As seen in figure 2, the variables do not cluster well about the factors. The loading matrix and the plot in figure 2 does not give a clear picture and interpretation of the correlation structure between the variables and the factors. Usually, a factor is most interpretable when a few variables are highly correlated with it and the rest are not. To maximize high correlations and minimise low ones, an orthogonal rotation is necessary.

4.5.Orthogonal rotation

In this method of rotation, we use the transformation matrix

$$M = \begin{pmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{pmatrix},$$

Such that the rotated loadings $\widehat{A^*}$ is

$$\widehat{A^*} = \hat{A} \ M$$

Where$\theta = 28°$is the angle of rotation and $\hat{A}$ is unrotated factor loadings.

The angle of 28^o was chosen such that one axis passes through a cluster of points. The factor loadings for the counter clockwise orthogonal rotation of 28^o were obtained as follows:

$$M = \begin{bmatrix} 0.8829 & -0.4695 \\ 0.4695 & 0.8829 \end{bmatrix} (16)$$

$$\widehat{A^*} = \begin{pmatrix} 0.902 & 0.076 \\ 0.054 & -0.807 \\ 0.020 & -0.934 \\ 0.847 & 0.035 \\ -0.050 & 0.920 \\ 0.848 & 0.131 \end{pmatrix} (17)$$

The factor loadings and the communalities are shown in the table 2 below:

Table 2: Rotated Factor Loadings Variable

	F_1	F_2	Communalities
Age	0.902	0.076	0.819
Mal.	0.054	−0.807	0.654
Typh.	0.020	−0.934	0.873
FBS	0.847	0.035	0.719
HB	−0.050	0.920	0.849
Dias.	0.848	0.131	0.736
Var	2.256	2.394	4.650
%Var	0.376	0.399	0.775

In the above table, the loadings in $\widehat{A^*}$ give a clear picture of the relationship between the variables and the factors. The high loadings are extremely high while the low ones are extremely low.It should be noted that the signs of the loadings do not count in this case. The plot of the rotated loadings is shown below.

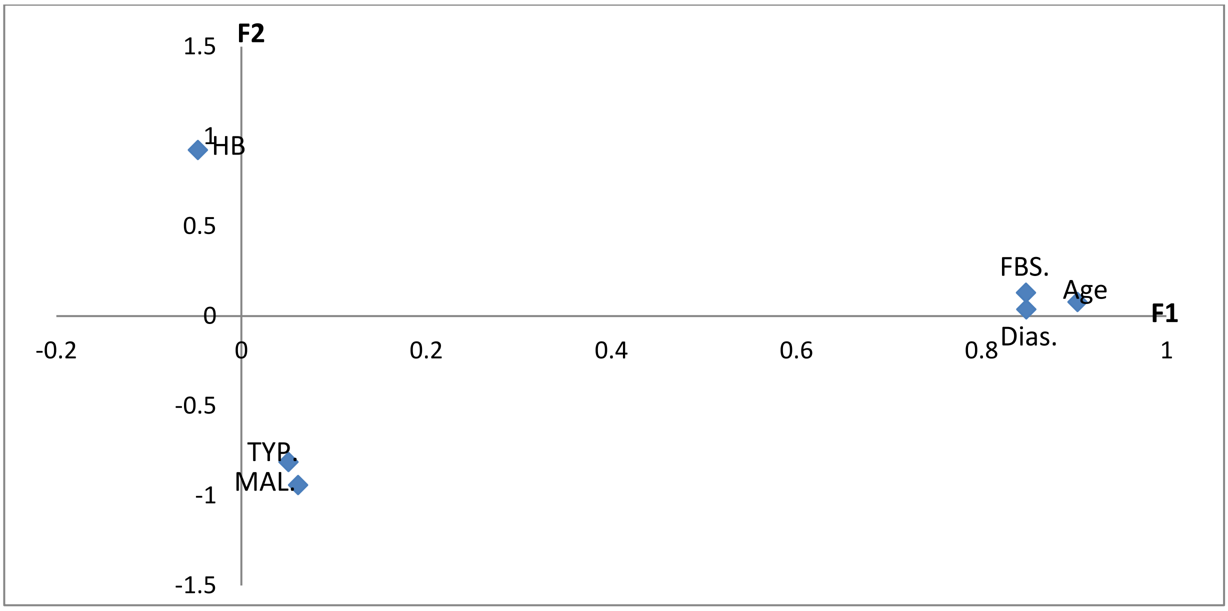

Fig. 3: Plot of Rotated Factor Loading.

In the above plot, the variables age(X_1), fasting blood sugar(X_4) and diastolic blood pressure(X_6) cluster about the F_1 axis while malaria(X_2), typhoid(X_3) and haemoglobin(X_5) cluster about the F_2 axis. Thus a clear picture of the factor loadings is achieved.

From the value of the communalities in the above table,81.9% of the variance in X_1 (Age) is accounted for by Factor 1 plus Factor 2. The communalities of other variables can be interpreted in the same way.

The value 0.376 means that 37.6% of the variance in the variables is accounted for by the first Factor (i.e. Factor 1)

The value 0.399 means that means that the second Factor accounts for39.9% of the variance in the variables.

Since the rotation is orthogonal, the two Factors together account for 78%of the variance in the variables.

4.6.The factor model

Since, a satisfactory factor loadings have been achieved, the factor model becomes

$$X - \mu = \widehat{A^*} \qquad F + \varepsilon (18)$$

Where

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{pmatrix}; \ \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \\ \mu_5 \\ \mu_6 \end{pmatrix}; \widehat{A^*} = \begin{pmatrix} 0.902 & 0.076 \\ 0.054 & -0.807 \\ 0.020 & -0.934 \\ 0.847 & 0.035 \\ -0.050 & 0.920 \\ 0.848 & 0.131 \end{pmatrix}; \ F$$

$$= \begin{pmatrix} F_1 \\ F_2 \end{pmatrix}; \ \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}$$

4.7.The Reproduced Correlation Matrix $\widehat{\mathrm{R}^*}$

As noted previously, the two factors account for 78% of the total variance and therefore represent the six variables well. To see how well the rotated two factor model reproduces the correlation matrix, we proceed as follows. The reproduced correlation matrix is obtained as

$$\widehat{\mathrm{R}^*} = \widehat{\mathrm{A}^*}\widehat{\mathrm{A}^{*\prime}} + \widehat{\psi}$$

$$= \begin{bmatrix} 1.000 & -0.041 & -0.033 & 0.687 & 0.023 & 0.675 \\ -0.041 & 1.000 & 0.655 & -0.007 & -0.545 & -0.016 \\ -0.033 & 0.655 & 1.000 & -0.006 & -0.860 & -0.125 \\ 0.687 & -0.007 & -0.006 & 1.000 & 0.003 & 0.523 \\ 0.023 & -0.545 & -0.860 & 0.003 & 1.000 & 0.078 \\ 0.675 & -0.016 & -0.125 & 0.523 & 0.078 & 1.000 \end{bmatrix} (19)$$

Comparing the closeness of the matrices (15) and (19) to the original correlation matrix $\widehat{\mathrm{R}}$; it is clear that the valuesestimated by (19) are closer to the values of$\widehat{\mathrm{R}}$than the elements of (15).

This simply means that the factor model generated by the rotated loadings is better than the one produced by the unrotated factor loadings.

4.8.The error (residual) matrix $\widehat{\mathrm{E}}$

The Residual (Error) correlation matrix $\widehat{\mathrm{E}}$ is obtained as

$$\widehat{\mathrm{E}} = \widehat{\mathrm{R}} - \widehat{\mathrm{R}^*}$$

$$= \begin{bmatrix} 0.000 & -0.006 & -0.006 & -0.022 & -0.012 & 0.011 \\ -0.006 & 0.000 & -0.038 & 0.003 & -0.039 & 0.004 \\ -0.006 & -0.038 & 0.000 & -0.002 & 0.009 & 0.002 \\ -0.022 & 0.003 & -0.002 & 0.000 & -0.002 & 0.029 \\ -0.012 & -0.039 & 0.009 & -0.002 & 0.000 & 0.002 \\ 0.011 & 0.004 & 0.002 & 0.029 & 0.002 & 0.000 \end{bmatrix} (20)$$

Observantly, the above error matrix $\widehat{\mathrm{E}}$almost results in a null matrix. In fact, a correction of the values in $\widehat{\mathrm{E}}$to one decimal place absolutely gives a null matrix. This is an evidence that the fitted factor model (18) using rotated loadings is adequate. In other words, there is no significant difference between the original correlation matrix and the correlation matrix generated by the factor analysis. It is now left to test statistically whether this difference is negligible.

4.9.Diagnosis

After obtaining the factor model, the next step is to carry out diagnostic checks to ascertain whether the model is adequate or not. This is achieved by examining the residual matrix as done in section 5.8 and using the hypothesis stated in section 4 of the methodology.

Since

$$\mathrm{N} = 200 \Rightarrow \frac{2}{\sqrt{200}} = 0.1414$$

Then, under this hypothesis,

$H_0$is rejected if $|r_{uv,i}| > \frac{2}{\sqrt{N}} = 0.1414$.

Now, examining the residual correlation matrix $\widehat{E}$ above; it clearly shows that none of the residual autocorrelations $|r_{uv,i}|$is greater-than0.1414. This provides evidence that the fitted factor model is adequate.

5. Discussion and conclusion

The aim of factor analysis is to represent the variables of interest as a linear combination of a few random variables called factors. The factors are underlying constructs or latent variables that generate the variables of interest. If these variables are correlated, the goal of the factor analysis is to reduce the redundancy among the variables by using a smaller number of factors (Brett, [4]). In this work, six variables of interest were involved: age(X_1), malaria(X_2), typhoid(X_3), fasting blood sugar(X_4), haemoglobin(X_5) and diastolic blood pressure(X_6). The correlation matrix of these variables clearly shows that some variables are highly correlated while some have low correlations. By this research, it was believed that this pattern of correlations is being generated by an underlying structure called factors. If the factors identified truly represent the pattern, a larger percentage of the variance in the variables will be accounted for by these factors. In this work two factors were identified to represent these groups of variables using the Kaiser and scree plot method. The method of principal component was used in the estimation of factor loadings. However, the loadings obtained did not give a satisfactory picture for interpretation. To enhance interpretation of the correlation structure, an orthogonal transformation matrix was multiplied with the original loadings to give satisfactory and interpretable results. The new loadings and their graphical representation gave a clear picture of the correlation structure defined by the two underlined factors. It was observed that age, fasting blood sugar (FBS) and blood pressure clustered into one group and loaded high on the first factor, F_1 (figure 3). This factor might be called Age-Cardiovascular factor. Similarly, malaria, typhoid and haemoglobin clustered and loaded high on the second factor. This factor can be labelled as Hemo-typhomalaria factor. It is also observed that though malaria, typhoid and haemoglobin are loaded high in one factor (F_2); malaria and typhoid load in opposite sign with haemoglobin. Perhaps, this is due to the fact that, medically, malaria and typhoid are inversely related to haemoglobin. However, in factor analysis, it is the absolute values of the loadings that are considered. The signs do not play significant role on how high or low the loadings are. It is also noted that the elements of the first factor all loaded with the same sign. The result of Bernard [3] showed that there exist a positive relationship between blood pressure and blood sugar. In this work, age(X_1), fasting blood sugar(X_4) and diastolic blood pressure(X_6) loads high on the Age-Cardiovascular factor. This affirms the conclusion drawn by Bernard [3].

References

[1] Alexander, S. (2004). Factor Analysis in Environmental Studies. HAIT Journal of Science and Engineering Volume 2 issues 1-2, pp. 54-94.

[2] Ani, G. and Sean, P. (2013).A Beginner's Guide to Factor Analysis. Tutorials in Quantitative Methods for Psychology vol. 9(2), pp. 79-94. https://doi.org/10.20982/tqmp.09.2.p079.

[3] Bernard, P. M. (2012). The Effect of Diabetes on High Blood Pressure. Journal of Medical Science. Vol. 7, Issue 6, pp. 51-59.

[4] Brett W. (2012).Exploratory Factor Analysis, A first-step guide for Novice Unpublished. Naim Publishers, New York. ISBN: 0-2533-344-7.

[5] Femke A., Ingrid V., Win D., and Marian V. (2012). On the relationship between blood pressure and haemoglobin. Journal of Medical Statistics, Vol. 10, No. 7; pp. 72-81.

[6] Hopkin J. (2009) Insulin resistance and Hypertension. American Journal of Psychology H1597-Hi602:2009.

[7] Ina D., George B. and Frank P. (2010). Type 2 Diabetes mellitus and increased Risk for Malaria infection: Emerging Infectious Diseases. Vol. 16, No. 10, October 2010.

[8] Kaiser, S. (1960). On Determination of Number of Factors in Factor Analysis. International Journal of Statiatics. Vol. 3; No. 8; pp. 34-47.

[9] Lutkepol H. (2005): New introduction to multiple Time Series Analysis. Springer Berlin Heidebelg New York. ISBN 3-540-40172-5. SPIN 10932797.https://doi.org/10.1007/978-3-540-27752-1.

[10] Madukosiri, G.M. (2012). Illness Pattern and Relationship between the Prevalence of Malaria and other infection. Agriculture and Biological Journal of North America (ABJNA), 2012.3.10.4B.https://doi.org/10.5251/abjna.2012.3.10.413.426.

[11] Michael W. (2007) an Overview of Analytical Rotation in Exploratory Factor Analysis. International Conference paper on Artificial Neural Network.Vol.4, pp. 57-66.

[12] Okuhara, K; Titu, S. and Williams, M. (2000). Application of Factor Analysis on volcanic soil constituents. Journal of Geological Sciences. Vol. 8; No. 4; pp. 13-21.

[13] Oscar, S. and Prasanna, K. (2012). Co-infection of Typhoid and Malaria. Journal of Medical Laboratory and Diagnosis vol. 2(3) pp. 22-26.

[14] Rechard, K. and Dean, H (1992). The Origin of Factor Analysis. CBMS-NSF Regional Conference Series in Applied Statistics. vol. 64, No. 3, pp. 52-58.

[15] Sylvia, M. (1997).On vital functions of Haemoglobin. J. Health inst., 19: 61-63.

[16] Uneke, P. N. (2002). Medline search method for the study of malaria and typhoid. Journal of Health Science. Vol. 5, Issue 5, pp. 332-340. DOI: 103923/jhsci. 2002 332-340.

[17] Williams, K.; John, M. and Benedict T. (2010). On the identification of factor Analysis. Journal of Research in Physical Sciences. Vol. 6; No. 3; pp. 18-27.

Appendix

Table 3: Age and Test Results of 200 Patients

S/No.	AGE (X_1)	MAL. (X_2)	TYPH. (X_3)	FBS (X_4)	HB (X_5)	DAIS. (X_6)	S/No.	AGE (X_1)	MAL. (X_2)	TYPH. (X_3)	FBS (X_4)	HB (X_5)	DAIS. (X_6)
1	65	1	60	5.7	15.5	90	51	44	3	60	4.6	12.6	83
2	23	3	60	4.5	15.6	79	52	49	2	60	4.3	16.1	82
3	21	2	180	4.3	14.5	77	53	49	3	180	4.6	11.2	84
4	20	1	180	3.5	14.7	75	54	67	3	360	5.6	9.8	90
5	49	3	360	5.2	7.5	79	55	68	2	180	5.5	12.2	88
6	54	3	360	5.5	7.6	85	56	62	3	360	4.8	8.3	89
7	22	3	180	4.8	14.5	89	57	63	1	60	5.3	15.1	91
8	57	2	60	5.6	16.2	95	58	54	2	60	4.7	14.7	85
9	24	2	180	4.7	14.7	80	59	50	2	180	4.6	14.2	85
10	23	2	180	4.6	14.4	85	60	69	1	60	5.2	15.6	91
11	44	1	60	5.2	15.1	80	61	55	2	60	3.9	15.4	86
12	55	2	180	3.3	14.3	88	62	29	2	180	4.3	14.3	77
13	65	3	60	3.8	15.9	91	63	27	2	180	4.2	14.9	78
14	49	3	360	5.8	10.1	85	64	66	3	360	4.7	10.1	89
15	43	2	180	4.6	14.7	83	65	69	3	360	5.6	9.8	90
16	63	1	60	5.5	15.9	88	66	54	2	360	4.6	10.2	85

17	64	2	180	5.4	14.7	86	67	60	3	180	5.2	14.3	91
18	34	3	360	3.9	9.5	75	68	32	2	60	3.9	15.7	75
19	30	3	180	3.5	14.2	83	69	41	1	60	3.8	10.8	83
20	20	2	60	5.6	15.3	90	70	60	3	180	5.8	7.4	83
21	64	1	60	5.5	15.8	90	71	60	3	360	4.5	9.3	80
22	61	2	60	5.6	15.4	86	72	63	2	180	5.5	14.6	90
23	33	2	180	4.7	14.3	80	73	63	3	360	5.4	8.9	88
24	29	2	180	4.6	14.4	85	74	33	3	180	3.9	14.3	77
25	53	2	180	5.2	14.6	91	75	29	1	180	4.3	13.9	50
26	49	1	60	3.3	15.8	83	76	24	1	60	4.2	15.8	79
27	50	2	180	4.5	14.1	82	77	28	1	60	4.6	16.3	81
28	36	1	60	4.3	16.2	80	78	40	3	360	4.8	11.1	79
29	30	3	360	4.5	9.8	80	79	44	2	360	4.6	10.4	85
30	62	2	180	5.7	14.3	90	80	59	2	180	5.2	14.4	80
31	60	2	360	5.3	7.9	85	81	62	1	60	5.5	15.5	90
32	49	1	360	4.2	13.5	76	82	34	2	180	4.8	14.3	80
33	46	3	360	4.3	7.6	85	83	64	2	180	5.6	14	89
34	42	2	180	4.2	13.9	83	84	34	3	360	4.7	10.1	80
35	61	1	60	5.6	15.5	88	85	51	2	180	4.6	14.3	85
36	64	2	180	4.7	14.2	90	86	63	3	360	5.5	8.4	90
37	55	2	360	4.6	9.9	85	87	46	2	180	4.8	14.2	80
38	30	3	180	5.2	13.9	77	88	63	1	60	5.6	16.1	86
39	25	2	180	3.3	14.1	79	89	52	2	180	4.5	14.2	80
40	31	1	180	3.8	14.4	84	90	62	2	60	5.5	15.5	90
41	37	3	360	4.1	10.4	83	91	35	2	360	4.3	11.6	73
42	25	3	180	4.5	8.6	80	92	51	3	360	4.9	7.2	85
43	62	2	60	5.5	15.9	90	93	30	3	360	3.5	9.4	81
44	59	2	180	5.4	14.3	90	94	60	1	360	5.5	10.5	90
45	29	2	180	3.9	10.5	77	95	65	2	180	5.6	14.4	91
46	24	3	180	4.3	11.6	75	96	22	2	180	4.5	13.9	75
47	22	2	180	4.2	13.9	75	97	41	1	180	4.4	10.1	83
48	38	3	180	4.6	14.5	84	98	24	3	360	4.3	8.4	79
49	35	1	60	4.6	15.9	72	99	30	3	180	4.5	14.6	82
50	66	2	180	5.7	14.1	92	100	29	2	60	3.5	15.7	78

S/No.	AGE (X_1)	MAL.	TYPH.	FBS	HB (X_5)	DAIS.	S/No.	AGE	MAL.	TYPH.	FBS (X_4)	HB	DAIS.
		(X_2)	(X_3)	(X_4)		(X_6)		(X_1)	(X_2)	(X_3)		(X_5)	(X_6)
101	61	2	180	5.1	14.3	90	151	65	3	180	5.5	14.8	90
102	60	2	180	5.5	14.5	81	152	61	2	60	5.4	15.6	90
103	35	3	180	4.8	11.5	85	153	45	2	180	4.5	14.5	76
104	64	2	180	5.6	14.7	90	154	38	2	180	4.3	14.4	75
105	52	3	180	4.7	14.5	85	155	29	1	60	4.2	15.6	80
106	55	1	60	4.6	15.9	90	156	62	2	180	5.5	15.3	91
107	63	2	180	5.5	13.7	89	157	63	1	60	5.6	15.7	87
108	49	2	60	4.8	15.4	89	158	46	3	360	4.6	14.9	80
109	64	2	60	5.6	15.1	89	159	24	2	180	4.6	14.8	77
110	27	3	180	4.7	14.3	80	160	60	1	60	5.2	15.5	91
111	50	3	360	5.1	7.5	77	161	23	2	180	3.3	14.3	89
112	59	2	180	5.2	13.9	88	162	31	2	180	3.8	14.3	75
113	34	3	360	4.9	7.8	85	163	24	2	360	3.9	10.1	76
114	55	1	60	4.8	15.6	89	164	40	3	180	4.5	14.2	80
115	63	2	60	5.6	15.3	80	165	63	2	60	5.5	16.2	90
116	29	2	180	4.6	14.3	80	166	58	1	60	5.4	15.5	90
117	33	1	60	4.6	15.9	85	167	49	2	60	3.9	15.2	87
118	61	2	180	5.5	14.3	88	168	23	2	180	4.3	14.6	75
119	40	2	60	4.8	15.9	79	169	40	2	180	4.7	13.5	80
120	62	2	180	5.6	14.1	86	170	40	1	180	4.6	14.2	85
121	45	3	180	4.7	14.5	80	171	60	3	360	5.2	10.1	91
122	60	3	360	5.5	9.2	90	172	34	3	360	4.3	9.8	75
123	45	2	360	4.3	8.2	80	173	20	3	180	4.5	14.3	75
124	50	3	180	4.6	14.2	85	174	32	1	60	4.1	15.4	79
125	39	2	60	3.5	10.2	84	175	35	2	180	4.5	14.3	80
126	59	1	60	5.2	15.4	90	176	64	2	180	5.5	14.6	90
127	64	3	180	5.5	13.9	91	177	58	1	60	5.4	15.9	90
128	54	3	360	4.8	9.3	89	178	29	2	180	3.9	16.1	77
129	63	2	180	5.6	14.3	91	179	39	1	60	4.3	15.5	78
130	43	3	360	4.7	9.1	80	180	46	3	180	4.2	14.4	85
131	41	3	180	4.6	13.9	85	181	40	3	360	4.6	10.2	76
132	62	2	180	5.2	14.3	91	182	53	2	180	4.6	14.5	81
133	23	2	60	3.2	15.4	83	183	60	3	360	5.6	10.1	90
134	39	1	60	4.1	15.9	78	184	39	1	60	4.6	15.8	87
135	30	3	360	4.2	7.2	78	185	50	1	60	4.6	16.2	85
136	39	3	360	4.5	10.1	80	186	64	3	60	5.2	14.1	91
137	61	2	180	5.5	13.9	86	187	46	2	180	3.3	14.3	84
138	40	1	60	4.7	15.6	80	188	33	1	180	3.8	11.9	84
139	32	2	180	4.6	13.7	85	189	55	3	360	4.2	9.3	86
140	60	3	360	5.2	10.2	90	190	34	3	360	4.5	10.4	80
141	61	3	360	5.5	7.3	89	191	62	3	180	5.5	14.4	90
142	30	2	180	4.8	14.3	89	192	23	2	60	4.7	15.1	80
143	31	3	360	5.6	9.6	95	193	22	2	180	4.6	13.6	77
144	29	2	360	4.7	10.4	80	194	59	2	180	5.2	14.4	90
145	41	2	180	4.6	14.3	85	195	63	1	60	5.5	15.4	91
146	63	1	60	5.2	15.1	91	196	41	2	180	4.8	14.3	84
147	44	3	60	3.3	15.5	84	197	63	1	60	5.6	15.9	87
148	21	1	180	3.8	14.3	77	198	49	3	360	4.8	10.1	85
149	44	3	360	4.3	10.2	81	199	44	2	180	4.6	14.2	85
150	55	3	360	4.5	9.7	86	200	59	1	60	5.2	15.8	91