A Statistical Study to Determine the Most Important Factors Affecting (Heart Disease) using Factor Analysis

م.م. هوازن طه عبدالله — الباحث

أ.م.د. سميرة محمد صالح — جامعة السليمانية/ كلية الادارة والاقتصاد

م.م. زيان محمد عمر — جامعة السليمانية/ كلية الادارة و الاقتصاد

P: ISSN : 1813-6729
E : ISSN : 2707-1359

http://doi.org/10.31272/JAE.i129.67

تأريخ أستلام البحث : 2021/3/31

مقبول للنشر بتأريخ :2021/4/18

Abstract:

This study has conducted for the purpose of defining the most influential factors for the occurrences of the heart disease in a group of patient during the classification and specification of the most significant ones of the variables with the realization of each variable and its influence on each other in the study,as well as the classifying relation among these variables were described through practicing on the analytical part which is called the factor analysis by using the method of principal axis factor on the heart data. This work reached out and finished with some conclusions among the most prominent ones are the ability to classify the variables into six group or major heterogeneity among themselves that influence the occurrence of the diseases.We also conclude that those variables that cause heart failure, which are serum creatinine, serum sodium, platelets, ejection fraction (cardiac output), sex and smoking, and also monitoring the time of disease occurrence in patients with the doctor.

Keywords: **Factor analysis, common variance, Unique Variance, Principal Factor Method, heart disease.**

مجلة الادارة والاقتصاد
مجلد 46 / العدد 129 / ايلول/ 2021
الصفحات : 468-486

1.1: Introduction

Factor analysis is a multivariate tool that aims to study a set of variables associated with each other and transformed into a lesser group of unrelated factors, the wide spread of the factor analysis within the practical reality in the scientific and human studies was based on the characteristics and assumptions that helped spread. It reduces the number of studied variables to a smaller number of affective factors on the studied variables ,in practical reality, the study of any phenomenon requires an extrapolation of the variables that affect it, and these variables may have different units of measurement and this difference may sometimes prevent the achievement of the desired results,of Factor analysis and because of its dependence on the correlation matrix and not on the covariance matrix.

1.2: Literature Review

- In 2020, Wibowo, Utami, Nadia, Nizeyumukiza, and Setiawati purposed to analyze the scare of COVID-19 in the Indonesian residents. Exploratory Data Analysis and Confirmation Factor analysis utilized for this aim. An entire of 117 participants replied to set of scale. The outcomes of the EDA demonstrated that the pair of scales has two dimensions .Moreover the consequences of the CFA discovered that the Indonesian model of FCV-19 displayed great structure validity (factorial and convergent), and passable accuracy. These discoveries recommend that the Indonesian model of FCV-19 is a progressively appropriate instrument that can be utilized to check out fears of corona virus in Indonesia. Although the problem of worry and scare has not been a priority in the treatment of corona virus, the results of studies show that there is an increase in people's fear of corona virus.
- Also In 2020, Onyekachi and Olanrewaju aim to draw attention to the most effective extraction technique which will be investigate while they are using the common three of the foremost popular methods for selecting the quantity of factors: Principal Component Analysis, Maximum Likelihood Estimate and axis correlational analysis (PAFA), and comparison the performance of them in words of precision and accuracy. to realize this study objective, the analysis of the three methods was subjected to numerous research contexts. A Monte Carlo method was accustomed simulate data. It generates variety of datasets for the five organization considered during this study: the conventional, Exponential, Uniform, Gamma and Laplace distributions. the extent of improvement within the estimates was associated with the proportion of observed variables and therefore the entirety of the square being load the factors inside the dataset and across the studied distributions. Different combines of sample size and whole number of variables through the distributions were accustomed perform the examination on the three analyzed techniques. The results of analysis, by applying the strategies on the simulated data, emphasize that PC analysis is overall most fitted, in spite of the fact that the loadings from PCA and PAFA are maybe same and don't differ considerably.

- 2018, Maskey, Fei and Nguyen HO, their goal was to debate the approaches that are embraced whereas applying exploratory factor analysis (EFA) in maritime journals to achieve an element solution that fulfills the standards of EFA, obtains the project targets and creates explanation easy. to attain this aim, 35 papers from four maritime journals were reviewed. this may be come after an instance of EFA utilizing an empirical data set to emphasize the approaches which will be undertaken to create appropriate decisions on whether to keep or drop a thing from the analysis to achieve an interpretable factor solution. The conclusion of this paper explain that major part of maritime researches utilize EFA retain a factor solution supported the researchers' subjective judgment. However, the researchers don't provide sufficient information to permit readers to gauge the analysis. The bulk of the reviewed papers did not provide important information associated with EFA explaining how the ultimate factor structure has been acquired. Furthermore, some papers have did not justify their decisions, for instance, for removing an item or keeping factors with individual measured variable.
- In 2016,Rossoni L., EngelbertR.and Luiz B. N., their aim was to analyze how different methods of extraction, factor definition, and rotation of exploratory correlational analysis affect the fit of measurement scales. For this purpose, they use a meta-analysis of 23 papers. Their conclusion demonstrates that the Principal Components method gives major explained variance, while the most Likelihood method get larger reliability. Therefore the rotations methods, Varimax gives larger reliability otherwise Quartimax gives smaller correlation between factors. lastly, this paper focuses attention on implications for quantitative study and proposes possible novel studies.
- In 2014, Hamed T., Shamsul S.and Neda J. intended to supply a simplified collection of knowledge for researchers and practitioners undertaking exploratory factor analysis (EFA) and to form decisions about best practice in EFA. Particularly, the target of the paper is to supply practical and theoretical information on deciding of sample size, extraction, number of factors to retain and rotational methods.

1.3: Factor Analysis

Factor analysis is a multivariate method used to study and analyze the internal relations between a large number of variables through common factors causing these relationships to find a new set of variables that are less in number than the original set of variables with the least loss of information. That these factors are orthogonal and reflect the common variance between variables. The factor analysis uses the Correlation Coefficients Matrix or the Covariance Matrix in the factor analysis (Shalal,404,2000). There is a main hypothesis of the factor analysis is that it is not possible to directly observe these factors, as the variables depend on the factors, but they are also subject to random errors .Factor analysis does not require making any assumptions about the nature of the variables or observations under study, and this is the reason for its flexibility, as it can be used on the broadest scale in scientific studies to analyze a large number of variables and return them to a smaller number of important factors that constitute the original variables, so that

Explain most of the differences in the data obtained and then give the appropriate model that represents the studied problem and an economic and objective description of the multivariate phenomena.(Azam,643,1998).

1.4: Factor Analysis Model

The factor analysis model interprets (P) of the variables for a sample of size (n) on the basis of a linear function that consists of (P) of the mean of the variables, (m) of the common factors and (P) of unique factors for each variable.(Andrson,551,1998) .Where (m <p), this linear model is as follows:

$$\underline{X}_{P\times 1} = \underline{\mu}_{P\times 1} + A_{P\times m}\,\underline{F}_{m\times 1} + \underline{U}_{P\times 1} \qquad (1)$$

whereas:-

X: represents the random vector of the observations.

μ: represents the vector of the variable mean.

A: represents the loading factor matrix of the variables.

F: represents the random vector of the Common Factors selected from (P) of variables.

U: represents the random vector of uniquefactors (specific variance) of the variables.

If the variable unit measurement are different, the standard value is used in the analysis of the correlation matrix by converting the variables into standard variables (having one mean and one variance),(Johnson,482,2007),(Azam,642,1998). In other words:

1- The vector of the variables mean will be a zero vector, that is,:
$$E(\underline{X}) = \underline{\mu} = \underline{0} \qquad (2)$$
2- The vector of covariance will be a unary vector, i.e. that:
$$Var(\underline{X}) = \underline{I} \qquad (3)$$
In this case the factor analyze formula will be as follow:
$$\underline{X} = A\underline{F} + \underline{U} \qquad (4)$$

The two directions of the averages of both the common and the unique factors are positive zero, depending on the assumption that the mean vector of the variables is also zero, i.e. that (Andrson,555,1984),(Alven, 306,2002)

$$E\left[\frac{F}{U}\right] = \left[\frac{0}{0}\right] \qquad (5)$$

The Covariance Matrix for U and F (assuming they are independent) is

$$E\left[\frac{\underline{F}}{\underline{U}}\right]\left[\underline{F}' \quad \underline{U}'\right] = \begin{bmatrix} E(\underline{F}\underline{F}') & E(\underline{F}\underline{U}') \\ E(\underline{U}\underline{F}') & E(\underline{U}\underline{U}') \end{bmatrix} = \begin{bmatrix} \Phi_{(m\times m)} & 0_{(m\times p)} \\ 0_{(p\times m)} & \Psi_{(p\times p)} \end{bmatrix} \qquad (6)$$

$$\begin{aligned} \mathrm{E}(\underline{U}) &= \underline{0}_{(p\times 1)} \\ \mathrm{E}(\underline{F}) &= \underline{0}_{(m\times 1)} \\ Cov(U,F) = \mathrm{E}(UF') &= \underline{0}_{(p\times m)} \\ \mathrm{E}(FF') &= \Phi \\ \mathrm{E}(UU') &= \Psi \end{aligned} \qquad (7)$$

whereas:

Φ: Represents the covariance matrix of common factors (F).

Ψ: Represents the diagonal matrix of single factor covariance (U)

And the covariance matrix of x is

$E[X:X'] - [E(X)]^2 = \Sigma_{(p\times p)}$ (8)

Where Σ is a symmetric positive matrix of order p

And the standard linear model for factor analysis to explain the value of the item i of the variable j for m of the factors

The basic model of factor analysis is based on the interpretation of the singular value (i) of the variable (j) as the weighted sum of (m) values of the factors, and this linear model can be formulated with the standard value as well as follows(shalal,406,2000)

$$Z_{ji} = a_{j1}F_{1i} + a_{j2}F_{2i} + \cdots + a_{jm}F_{mi} + U_{ji} \qquad (9)$$

whereas:

Z_{ji}: the standard value of the variable i and j.

F_{1i}, ……, F_{mi}: standard value of the singular item i of the given common factor.

U_{ji}: the standard value of the singular i for the unique factor of the variable j.

a_{j1}, ……, a_{jm}: Loading Factors, which are weights associated with the common factor values.

1.5: Basic Assumptions of Factor Analysis

The first hypothesis: the existence of a correlation between a group of variables, which is sometimes known as (inter-correlation), and that these correlations are the result of the presence of common factors affecting them and that the amount of these correlations is due to the reality of those factors, where the factor analysis

seeks to explain the correlations between the variables with less factors of the variables used, and from these factors the standard value can be represented in the case of assuming the presence of (m) of the factors,(Rezan,21,2004) as indicated by the following equation:

$$S_{ji} = a_{j1}Z_{1i} + a_{j2}Z_{2i} + \cdots + a_{jm}Z_{mi} \quad (10)$$

S_{ji}:Represents the standard value of the measure (or observations) i with respect to the variable j

a_{jm}:The load (saturation) factor (m) with respect to the variable j

Z_{mi}: Standard value of observation i relative to the factor m

In light of the first hypothesis, there are three types of variation, which are:-

1- Common Variance

This part of the variance is explained by common factors, i.e., involving by other variables, and it is called (Communalities), which is the sum of the squares of the common factor loads and is denoted by (h_j^2) (Tharwat,405,2011)(Seber,219,1984)

$$h_j^2 = a_{j1}^2 + a_{j2}^2 + \cdots + a_{jm}^2 \quad (11)$$

All of a_{jm}^2, …… , a_{j2}^2 , $a_{j1}^2$is the square of the coefficient of correlation between the variable j and k factor where, $J = 1, 2, \ldots\ldots, p$ ' $k = 1, 2, \ldots\ldots, m$'$m < p$

It is a ratio of the total variance of the variable (j) that the factor (k) contributed to the determinion and it is called as Coefficient of Determination.

2- Unique Variance

It represents the extent to which the unique factor (U_j) contributes to the variance of the variable (j).

This part of the overall variance is divided into two parts:

A- Specific variance: It is the proportion of total variance that explained by the factors of the same variable, i.e., it is not related to the rest of the variables.

B- Error Variance: It is the variance resulting from the occurrence of errors in drawing or measuring the sample or any changes that may occur to the data, i.e., it is not explained by the common factors.(Johnson,501,2007).

$$U_j^2 = b_j^2 + e_j^2 \quad (12)$$

U_j^2: the variance of unique factor
b_j^2: specific variance of variable j
e_j^2: error variance

The variance components for any variable are according to the following equation (Johnson ,484,2007),(Shalal,409,2000)

$$\sigma_{jj} = a_{j1}^2 + a_{j2}^2 + \cdots + a_{jm}^2 + U_j^2 \quad (13)$$

Var (x_j) = Communality + Unique variance

Whereas:

$$h_j^{\,2} = a_{j1}^{\,2} + a_{j2}^{\,2} + \cdots + a_{jm}^{\,2} \qquad (14)$$

$$\therefore \sigma_{jj} = h_j^{\,2} + U_j^{\,2} \qquad (15)$$

As $X_j \sim N(0,1)$ That is

By using both of equations (12) and (15), we obtain the value of the error variance for any variable as follows:

$$e_j^{\,2} = 1 - \left(h_j^{\,2} + b_j^{\,2}\right) \qquad (16)$$

The second hypothesis:

Factor analysis also assumes the existence of a correlation between the variables (j, k) so that it can be calculated on the nature and effect of the common factor loads (saturations).

The correlation values between these variables can be found as follows:

$$R = AA'$$

R: represents the correlations matrix.

A: represents the loading factors matrix.

1.6: Communalities

The communalities amount of any variable is the sum squares of loading variables, and it represents the proportion of variance explained by the extracted factors for these variables and is denoted by the symbol (h_j^2) .(Tharwat,400,2011).

$$h_j^{\,2} = a_{j1}^{\,2} + a_{j2}^{\,2} + \cdots + a_{jm}^{\,2}$$

$$h_j^{\,2} = \sum_{p=1}^{m} a_{jp}^{\,2} \qquad (17)$$

$$0 \le h_j^{\,2} \le 1$$

Whereas:

a_{jp} Represents the weight of the factor p with respect to the variable j and are the coefficients of factor matrix and known as factors loading or factor saturations.

1.7: Factor analysis methods

There are several methods of factor analysis and the main factor method was adopted to analyze the research sample data, so our focus is on this method and its applications in this study.

1.7.1: Principal Factor Method

The principal factor method is an application of principle component method but by using reduced correlation matrix.

As the principal components method is one of the most accurate and popular factor analysis methods in research, this method has several advantages, including that it leads to accurate saturations and leads to the least possible amount of residues also, the reduced correlation matrix to the smallest number of orthogonal (unrelated) factors, the equation (9) can be written in a matrix form (Rezan,22,2000).

$$Z_{(p\times 1)} = A_{(p\times m)}F_{(m\times 1)} + U_{(p\times 1)} \qquad (18)$$

Whereas:

$$\begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_p \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & . & . & a_{1m} \\ a_{21} & a_{22} & . & . & a_{2m} \\ . & . & . & . & . \\ . & . & . & . & . \\ a_{p1} & a_{p2} & . & . & a_{pm} \end{bmatrix} \times \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix} + \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_p \end{bmatrix} \qquad (19)$$

In order for estimating the loading factors matrix, the following steps are taken:

1- We calculate the correlation coefficient matrix from standard valuefor the variables with different measurement units, if the variables have the same measurement units,we use the variance-covariance matrix.(Hair,chapter3,2010),(Azam,200,1998).

Where

$$R = \begin{bmatrix} 1 & r_{12} & . & . & r_{1p} \\ r_{21} & 1 & . & . & a_{2m} \\ . & . & . & . & . \\ . & . & . & . & . \\ r_{p1} & r_{p2} & . & . & 1 \end{bmatrix} \qquad (20)$$

2. We compute the square of the multiple correlation coefficient for each variable with the rest of the variables, $R^2_{j,\ \text{rest}}$,as a first estimate of the values of the communality to replace the diagonal elements to obtain the Reduced correlation matrix (Rr).

3. In the reduced correlation matrix, Eigen values are extracted according to the following characteristic equation:

$$|Rr - \lambda I| = 0 \ (21)$$

4. We choose the distinct values whose value is greater than one, as their number represents the number of factors we use in our analysis.

5. Extract the Eigen vector (L) - associated with each distinct value, so choose it and start with the largest value, according to the following system of equations:

$$|Rr - \lambda I|\underline{a} = \underline{0}(22)$$

As the characteristic value is the amount of the factor's contribution to the sum of the communality values. The vector of the characteristic associated with the largest characteristic value represents the estimated first factor loads, the vector of the second largest characteristic value represents the estimated second factor loads and so on. In this way, we obtain a matrix of the first estimated factor loads, i.e., that: (Seber ,202,1984).

$$L_1 = \begin{bmatrix} L_{11} & L_{12} & \dots\dots & L_{1m} \\ L_{p1} & L_{p2} & \dots\dots & L \end{bmatrix} \qquad (23)$$

6. we obtain the communality value from L_1 And as it follows:

$$\begin{aligned} h^2{}_1 &= a^2{}_{11} + a^2{}_{12} + \cdots + a^2{}_{1m} \\ &\cdots\cdots \\ &\cdots\cdots \\ h^2{}_p &= a^2{}_{p1} + a^2{}_{p2} + \cdots + a^2{}_{pm} \end{aligned} \qquad (24)$$

These values are illustrated in the reduced correlation matrix Rr as the diagonal elements.

7. To obtain the second factor estimated loading matrix A_2 we will repeat the 3, 4, 5, 6 steps and so on, until the differences between $h^2{}_j$ of two consecutive matrices are very small, so this matrix is the final estimated factorial matrix (L) which is the (Initial Solution). The purpose of the main factor method is based on reducing the number of variables and distributes them in a linear corresponding form so that the number of factors is less than the total rank of the matrix.

1.8: The rotated factor matrix

There are several methods for obtaining the matrix of the rounded factors and since we have relied in our research on the method (Varimax), and this method was proposed in 1958 by Kaiser and it is the most common orthogonal rotation method and depends on simplifying the composition of factors through the loading square variances.(Tharwat,405,2011).

$$S^2{}_p = \frac{1}{n}\sum_{j=1}^{n}(a^2{}_{jp})^2 - \frac{1}{n^2}\sum_{j=1}^{n}(a^2{}_{jp})^2 \qquad (25)$$

Whered_{jp} is the row element and (j) of the column (P) in the inverse matrix and when the variance is the greater than possible, the factor has the ability to explain and simplify on the basis that its loads are directed around zero and one, so if you collect (25) with all the factors, then:

$$S^2{}_p = \sum_{p=1}^{m} S^2{}_p = \frac{1}{n}\sum_{p=1}^{m}\sum_{j=1}^{n} a^4{}_{jp} - \frac{1}{n^2}\sum_{p=1}^{m}(\sum_{j=1}^{n} a^2{}_{jp})^2 \qquad (26)$$

To maximize in equation (26) is called the raw varimax criterion, and equation (26) oscillates in terms of the communality values $(h^2{}_j)$ so the approach criterion for

maximizing variance is to make its value (24) the greatest possible to obtain the best loading factors, as follows:

$$V = n\sum_{p=1}^{m}\sum_{j=1}^{n}(a_{jp}/h_j)^4 - \sum_{p=1}^{m}(\sum_{j=1}^{n} a^2{}_{jp}/h^2{}_j)^2 \qquad (27)$$

So, the equation (24) was called the normal varimax measurement.

2.1: Application Data & analysis

2.1.1: Description of Data & analysis

The data that analyzed contains medical annals of 299 heart failure patients. The patients contained of 105 female and 194 male, and their ages scope between 40 and 95 years old. The dataset embodies 13 countenance, which report clinical, body, and lifestyle information, that we compendiously describe here. Some features are binary: anemia, high blood pressure, diabetes, sex, and smoking. The data has been collected during year 2020. We took our data from https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records, where this site belongs to Center for Machine Learning and Intelligent Systems. Since there are 13 attributes from this dataset as follow:

1- X_1: Represents the Age Variable
2- X_2: Represents the Anaemia Variable
3- X_3: Represents the Creatinine Phosphokinase Variable
4- X_4: Represents the Diabetes Variable
5- X_5: Represents the Ejection Fraction Variable
6- X_6: Represents the High Blood Pressure Variable
7- X_7: Represents the Platelets Variable
8- X_8: Represents the Serum Creatinine
9- X_9: Represents the Serum Sodium Variable
10- X_{10}: Represents the Sex (Gender) Variable
11- X_{11}: Represents the Smoking Variable
12- X_{12}: Represents the Time
13- X_{13}: Represents the Death Event Variable

After describing the data by using the statistical package (SPSS 26) ,we get these results :

Table (1): Represents the Total Variance Explained

Factor	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	2.031	15.623	15.623	1.528	11.754	11.754	1.168	8.986	8.986
2	1.661	12.776	28.399	1.110	8.538	20.291	1.006	7.737	16.723

3	1.312	10.093	38.492	0.631	4.855	25.147	0.645	4.959	21.682
4	1.159	8.913	47.405	0.496	3.818	28.965	0.599	4.607	26.289
5	1.031	7.928	55.333	0.324	2.491	31.456	0.573	4.409	30.699
6	1.001	7.699	63.032	0.266	2.047	33.503	0.365	2.805	33.503
7	0.908	6.986	70.018						
8	0.845	6.502	76.520						
9	0.768	5.908	82.428						
10	0.708	5.447	87.875						
11	0.686	5.277	93.152						
12	0.514	3.951	97.103						
13	0.377	2.897	100.000						
	13.001	100							

The table above represent the number of factors whose distinctive value exceeds the correct one, and these six factors explain (63.032%) of the total variance of the variables, and that these extracted factors explain a different percentage of the variance, but each factor has its importance in diagnosing the influencing variables. In the event of heart failure.These six factors explain each of them respectively (15.623%, 12.776%, 10.093%, 8.913%, 7.928%, 7.699%) of the total variance.

From the table above it is obvious that the sum of the variations of the factors (the Initial Eigenvalues) equals (13.001), which represents the total variance of all the studied variables (the variables in their standard form), and the shaded parts indicate the presence of six main factors (significant) that lead to the occurrence of heart failure in the patients, which

Table (2): Represents the Rotation Varimax Method and Communalities which (Rotation converged in 7 iterations.)

Rotated Factor Matrix							**Communalities**	
	Factor						Initial	Extraction
	1	2	3	4	5	6		
X_1	0.321	0.069	0.189	0.146	-0.101	-0.184	0.125	0.209
X_2	0.179	-0.130	-0.026	0.041	-0.280	-0.126	0.080	0.146
X_3	0.007	-0.001	-0.058	-0.021	0.663	-0.016	0.069	0.444
X_4	-0.071	-0.199	0.053	-0.097	-0.007	0.273	0.061	0.132
X_5	-0.031	-0.080	-0.076	0.615	-0.051	0.066	0.148	0.398
X_6	0.267	-0.063	-0.089	0.051	-0.104	0.078	0.064	0.103
X_7	0.035	-0.005	-0.072	0.097	0.041	0.368	0.044	0.153
X_8	0.163	-0.019	0.468	0.059	0.008	-0.107	0.125	0.261
X_9	0.017	-0.001	-0.441	0.235	0.054	-0.052	0.100	0.256
X_{10}	-0.061	0.617	0.026	-0.119	0.107	-0.271	0.256	0.484
X_{11}	0.000	0.742	-0.019	-0.025	0.014	0.058	0.222	0.555
X_{12}	-0.689	-0.038	-0.123	0.053	0.007	0.014	0.329	0.494
X_{13}	0.671	-0.031	0.392	-0.317	0.121	-0.057	0.417	0.722

% of Variance	15.623	12.776	10.093	8.913	7.928	7.699		

From the **rotated factor matrix** table, it is obvious that we have six main effective factors that affect the heart failure variables with 63.032% of the total variance where:

In the first factorthere is 15.623% of variance which describes the effect of this factor over the variables especially the (X_{12}:time and X_{13}:death event) variables with (0.689 and 0.671).Also in the second factor it explains 12.776% of total variance we can see that it describes both of (X_{10}:smoking and X_{11}:sex) among all the variables by (0.742 and 0.617). then we have the third factorwhere 10.093% of total variance explains the X_8: serum creatinine and X_9:serum sodium respectively by 0.468 and -0.441 among all the other variables. As well as fourth factor has the X_5:ejection fraction it is the variable that described by 0.615 among the other variables since the total variance in this factor is describing the other variables about 8.913%. in the last both of fifth factor and sixth factor, in these two factors their total variance are respectively 7.928% and 7.699% so those variables that affected in these two factors are X_3:creatinine phosphokinase by 0.663 in the fifth factor and X_7:platelets by 0.368 in the sixth factor.

The communalities after the iteratives of six factors solutions, they are now sum to 2.04 which means 16% of variance is common and 84% is unique. From the iteration of communalities in principal axis factoring with the residual (R^2)'s it takes the communalities from the first step and insert them into the main diagonal correlation matrix instead of the (R^2) and repeats the analyzation again.

Table (3): Represents the Correlation Matrix for the (X_1:Age, X_2:Anaemia, X_3:Creatinine Phosphokinase, X_4:Diabetes, X_5:Ejection Fraction, X_6:High Blood Pressure, X_7:Platelets, X_8:Serum Creatinine, X_9:Serum Sodium, X_{10}:Sex, X_{11}:Smoking, X_{12}:Time and X_{13}:Death Event).

		X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}
Correlation	X_1	1.000	0.088	-0.082	-0.101	0.060	0.093	-0.052	0.159	-0.046	0.065	0.019	-0.224	0.254
	X_2	0.088	1.000	-0.191	-0.013	0.032	0.038	-0.044	0.052	0.042	-0.095	-0.107	-0.141	0.066
	X_3	-0.082	-0.191	1.000	-0.010	-0.044	-0.071	0.024	-0.016	0.060	0.080	0.002	-0.009	0.063
	X_4	-0.101	-0.013	-0.010	1.000	-0.005	-0.013	0.092	-0.047	-0.090	-0.158	-0.147	0.034	-0.002
	X_5	0.060	0.032	-0.044	-0.005	1.000	0.024	0.072	-0.011	0.176	-0.148	-0.067	0.042	-0.269

	X_6	0.093	0.038	-0.071	-0.013	0.024	1.000	0.050	-0.005	0.037	-0.105	-0.056	-0.196	0.079
	X_7	-0.052	-0.044	0.024	0.092	0.072	0.050	1.000	-0.041	0.062	-0.125	0.028	0.011	-0.049
	X_8	0.159	0.052	-0.016	-0.047	-0.011	-0.005	-0.041	1.000	-0.189	0.007	-0.027	-0.149	0.294
	X_9	-0.046	0.042	0.060	-0.090	0.176	0.037	0.062	-0.189	1.000	-0.028	0.005	0.088	-0.195
	X_{10}	0.065	-.095	0.080	-0.158	-0.148	-0.105	-0.125	0.007	-0.028	1.000	0.446	-0.016	-0.004
	X_{11}	0.019	-.107	0.002	-0.147	-0.067	-0.056	0.028	-0.027	0.005	0.446	1.000	-0.023	-0.013
	X_{12}	-0.224	-.141	-0.009	0.034	0.042	-0.196	0.011	-0.149	0.088	-0.016	-0.023	1.000	-0.527
	X_{13}	0.254	.066	0.063	-0.002	-0.269	0.079	-0.049	0.294	-0.195	-0.004	-0.013	-0.527	1.000

In table (3) we can see the correlations between each variables and the determinant = 0.293 for the Correlation Matrix. It is necessary since the determinant is not equal to zero since it will produce a computational issue with the factor analysis.

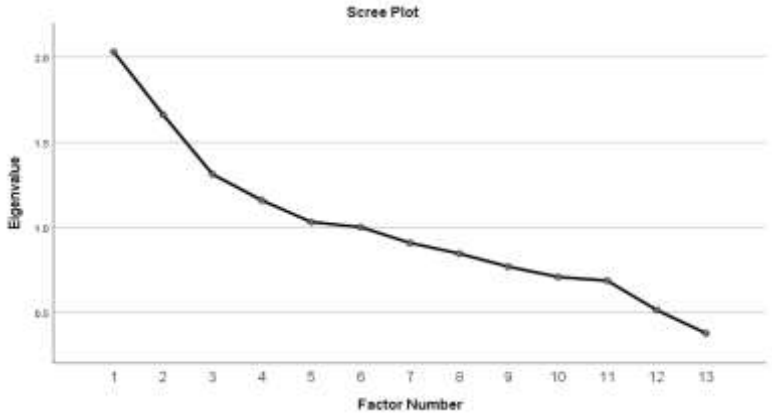

Figure (1): Represents the Scree Plot for the thirteen factors.

Table (4): Represents the Factor Matrix

	Factor					
	1	2	3	4	5	6
X_1	0.339	-0.017	0.262	0.036	0.154	-0.009
X_2	0.125	-0.201	0.237	-0.105	-0.029	-0.148
X_3	-0.008	0.165	-0.423	0.462	0.154	-0.034
X_4	-0.057	-0.206	-0.192	-0.070	-0.131	0.165
X_5	-0.260	-0.217	0.368	0.257	0.219	0.183
X_6	0.136	-0.157	0.151	0.110	-0.157	-0.016
X_7	-0.094	-0.094	-0.011	0.160	-0.200	0.263
X_8	0.362	-0.044	0.010	-0.100	0.299	0.169
X_9	-0.268	-0.021	0.192	0.315	-0.080	-0.203
X_{10}	0.086	0.676	0.083	-0.033	0.073	-0.080

X_{11}	0.041	0.670	0.206	0.035	-0.154	0.190
X_{12}	-0.630	0.085	-0.155	-0.216	0.138	0.018
X_{13}	0.831	-0.055	-0.149	0.064	-0.046	0.001

In the table above we can see the unrotated factors where there are correlation between the variables and the factors since their values are between $+1$ and -1. Also the correlation under 0.3 is considered as a weak correlation.

Table (5): Represents Factor Score Coefficient Matrix

	Factor					
	1	2	3	4	5	6
X_1	0.101	0.036	0.076	0.155	-0.064	-0.132
X_2	0.072	-0.048	-0.051	0.012	-0.160	-0.110
X_3	0.003	-0.036	-0.058	0.033	0.607	-0.021
X_4	-0.044	-0.059	0.047	-0.072	-0.004	0.207
X_5	0.072	-0.004	0.060	0.504	0.021	0.039
X_6	0.127	-0.006	-0.087	0.023	-0.051	0.057
X_7	0.047	0.028	-0.023	0.050	0.031	0.294
X_8	-0.032	-0.004	0.326	0.133	0.007	-0.069
X_9	0.114	-0.001	-0.323	0.116	0.046	-0.081
X_{10}	-0.051	0.352	0.023	-0.070	0.065	-0.281
X_{11}	0.028	0.567	-0.012	0.026	-0.033	0.199
X_{12}	-0.398	-0.043	0.097	-0.058	0.040	-0.040
X_{13}	0.473	-0.047	0.294	-0.270	0.149	-0.009

From the factor score coefficient matrix, the main diagonal matrix are for each six factors. So the $R^2$is between each factor and observed variables, hence values under 0.7 are undesirable. The variance of factor scores are the squared multiple correlation coefficients.

Table (6): Represents the Reproduced Correlations with Residuals

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}

Reproduced Correlation	X_1	0.487	0.176	-0.126	-0.250	0.203	0.120	-0.032	0.381	-0.038	0.074	0.088	-0.364	0.311
	X_2	0.176	0.506	-0.456	-0.073	0.029	0.200	-0.239	-0.006	0.031	-0.169	-0.189	-0.157	0.065
	X_3	-0.126	-0.456	0.774	-0.090	-0.078	-0.101	-0.004	-0.007	0.171	0.042	-0.065	-0.048	0.143
	X_4	-0.250	-0.073	-0.090	0.495	-0.131	0.026	0.299	-0.036	-0.270	-0.317	-0.228	0.084	-0.001
	X_5	0.203	0.029	-0.078	-0.131	0.702	0.001	0.234	0.110	0.310	-0.230	-0.121	0.168	-0.335
	X_6	0.120	0.200	-0.101	0.026	0.001	0.502	0.232	-0.177	0.205	-0.149	-0.027	-0.385	0.202
	X_7	-0.032	-0.239	-0.004	0.299	0.234	0.232	0.709	-0.057	0.023	-0.138	0.127	-0.032	-0.081
	X_8	0.381	-0.006	-0.007	-0.036	0.110	-0.177	-0.057	0.616	-0.367	-0.015	-0.065	-0.216	0.364
	X_9	-0.038	0.031	0.171	-0.270	0.310	0.205	0.023	-0.367	0.607	-0.052	-0.024	0.069	-0.297
	X_{10}	0.074	-0.169	0.042	-0.317	-0.230	-0.149	-0.138	-0.015	-0.052	0.675	0.645	0.003	0.016
	X_{11}	0.088	-0.189	-0.065	-0.228	-0.121	-0.027	0.127	-0.065	-0.024	0.645	0.733	-0.010	-0.047
	X_{12}	-0.364	-0.157	-0.048	0.084	0.168	-0.385	-0.032	-0.216	0.069	0.003	-0.010	0.653	-0.621
	X_{13}	0.311	0.065	0.143	-0.001	-0.335	0.202	-0.081	0.364	-0.297	0.016	-0.047	-0.621	0.734
Residual	X_1		-0.088	0.045	0.149	-0.143	-0.027	-0.021	-0.221	-0.008	-0.008	-0.069	0.140	-0.058
	X_2	-0.088		0.265	0.060	0.003	-0.161	0.195	0.058	0.011	0.074	0.082	0.015	0.002
	X_3	0.045	0.265		0.080	0.034	0.031	0.028	-0.010	-0.111	0.038	0.067	0.038	-0.080
	X_4	0.149	0.060	0.080		0.126	-0.039	-0.207	-0.011	0.181	0.160	0.081	-0.051	-0.001
	X_5	-0.143	0.003	0.034	0.126		0.023	-0.161	-0.121	-0.134	0.082	0.054	-0.126	0.066
	X_6	-0.027	-0.161	0.031	-0.039	0.023		-0.182	0.172	-0.168	0.044	-0.029	0.189	-0.123
	X_7	-0.021	0.195	0.028	-0.207	-0.161	-0.182		0.016	0.039	0.013	-0.098	0.043	0.032
	X_8	-0.221	0.058	-0.010	-0.011	-0.121	0.172	0.016		0.178	0.022	0.037	0.067	-0.069
	X_9	-0.008	0.011	-0.111	0.181	-0.134	-0.168	0.039	0.178		0.024	0.029	0.019	0.102
	X_{10}	-0.008	0.074	0.038	0.160	0.082	0.044	0.013	0.022	0.024		-0.199	-0.018	-0.021
	X_{11}	-0.069	0.082	0.067	0.081	0.054	-0.029	-0.098	0.037	0.029	-0.199		-0.013	0.035
	X_{12}	0.140	0.015	0.038	-0.051	-0.126	0.189	0.043	0.067	0.019	-0.018	-0.013		0.094
	X_{13}	-0.058	0.002	-0.080	-0.001	0.066	-0.123	0.032	-0.069	0.102	-0.021	0.035	0.094	

The Residual values in this part of the table represent the differences between original correlations (shown in the Table (3) the Correlation Matrix) and the reproduced correlations, which are shown in the table above. For example, the original correlation betweenthe X_1:Ageand X_2: Anaemia is 0.088, and the reproduced correlation between these two variables is 0.176. The residual is $-0.088 = 0.088 - 0.176$.

Conclusion and Recommendation

Conclusion:

From this study we conclude that:

1- The number of factors whose distinctive value exceeds the correct one, and these six factors explain (63.032%) of the total variance of the variables.
2- The rotated factor matrixtablethe most effect factors over the variables are (X_{12}:time and X_{13}:death event) variables with (0.689 and 0.671) values. As well as and X_7:Platelets by 0.368 in the sixth factor. On the other hand, the communalities issum to 2.04 which means 50% of variance.
3- In the reproduced matrix by subtracting the correlation values and the reproduced correlation values we found the residuals as follow: the original correlation betweenthe X_1:Ageand X_2: Anaemia is 0.088, and the reproduced correlation between these two variables is 0.176. The residual is $-0.088 = 0.088 - 0.176$.
4- The X_{12}:time variable is the most effective variable that affect the heart failure disease with 15.623% of variance in the first factor by 0.689 to compare with the less effective variable in the sixth factor with 7.699% of variancewhich is X_7:Platelets by 0.368.

Recommendation

1- Since from the result of our data one of the effective factors was smoking, we recommend the patients to quit smoking.
2- We also recommend the patient to keep their following up with doctors during the feeling of illness in their heart's.
3- Lipid profile test which means both of the (cholesterol and triglycerides) tests, blood pressure test, ECGtest, blood sugar test, CBC test, thyroid tests and chest x-ray. We recommend these tests for those patients with heart failure disease.

References

1) Anderson, T. W. (1984)," An introduction to Multivariate Statistical Analysis ", John Wily & Son, New York –USA.
2) Hamed T., Shamsul S., Neda J.(2014). "Exploratory Factor Analysis:Concepts and Theory". World Scientific and Engineering Academy and Society, 27, pp.375- 382 , Mathematics and Computers in Science and Engineering Series.
3) Hair JR.JosephF.,Black William C .,Babin,BarryJ,AndersonRolph E.(2010) ,7th edition "Multivariate Data Analysis "perason Prentice Hall.
4) MaskeyR. , Fei J., Nguyen H.O. (2018) "Use of exploratory factor analysis in maritime research". The Asian Journal of Shipping and Logistics 34(2):91–111.
5) Johnson Richard A.,Wichern Dean W.(2007),8th Edition "Applied Multivariate Statistical Analysis "perason Prentice Hall .
6) Onyekachi A. M., Olanrewaju S. O. ,(2020) "A Comparison of Principal Component Analysis, Maximum Likelihood and the Principal Axis in Factor Analysis", American Journal of Mathematics and Statistics, 10(2), pp. 44-54.
7) Rencher , Alvin C.,(2002),2nd edition "Methods of Multivariate Analysis" John Wily & Son, New York –USA.
8) Rossoni L., Engelbert R., Luiz B. N. (2016), "Normal science and its tools: Reviewing the effects of exploratory factor analysis in

management". Revista de Administração ., Vol. 51 Issue 2, p198-211. 14p.
9) Seber ,G.A.F .(1984)," Multivariate Observations. John Wiley & sons,new York –USA.
10) Wibowo, Y., Utami, R., Nadia, Y., Nizeyumukiza, E., &Setiawati, F. (2020). "The fear of coronavirus scale: exploratory and confirmatory factor analysis". Konselor, 9(2), 75-80.

المصادر العربية :

11) جونسون ،ريتشارد ،دين وشر ،تعريب عزام ،عبدالرحمن حامد (1998)" التحليل الاحصائي للمتغيرات المتعددة من الوجهة التطبيقية " دارالمريخ للنشر ، السعودية .

12) د. شلال حبيب الجبوري ، صلاح حمزة عبد (2000) ، تحليل متعدد المتغيرات ، دار الكتبللطباعة و النشر – بغداد ..

13)عبدالمنعم ،د.ثروت محمد (2011)")" التحليل الاحصائي للمتغيرات المتعددة " ، مكتبة الانجلو المصرية ، القاهرة .

14)كريم، ريزان حمه خورشيد (2003)، " د راسة احصائية لأهم العوامل المؤثرة على ظاهرة الانتحار" رسالة ماجستير، مقدمة الى مجلس كلية الأدارة و الأقتصاد ، جامعة السليمانية .

A Statistical Study to Determine the Most Important Factors Affecting (Heart Disease) using Factor Analysis

دراسة إحصائية لتحديد لأهم العوامل المؤثرة على (أمراض القلب) باستخدام تحليل العاملي

م.م. هوزان طه عبدالله
كلية الادارة و الاقتصاد
جامعة السليمانية
hozan.abdulla@univsul.edu.iq

أ.م.د. سميرة محمد صالح
كلية الادارة و الاقتصاد
جامعة السليمانية
Samira.muhamad@univsul.edu.iq

م.م.زيان محمد عمر
كلية الادارة و الاقتصاد
جامعة السليمانية
zhyan.omer@univsul.edu.iq

المستخلص:

أجريت هذه الدراسة بهدف تحديد أكثر العوامل تأثيراً في حدوث أمراض القلب لدى مجموعة من المرضى خلال تصنيف وتحديد أهم المتغيرات مع إدراك كل متغير وتأثيره على الآخر. في الدراسة وكذلك تم وصف علاقة التصنيف بين هذه المتغيرات من خلال التمرين على التحليل الذي يسمى العامل التحليلي باستخدام طريقة عامل المحاور الرئيسي على البيانات المتعلقة بالقلب.توصلت الى وتنتهي هذه الأعمال ببعض الاستنتاجات بين المتغيرات. أبرزها القدرة على تصنيف المتغيرات إلى ستة مجموعات أو عدم تجانس رئيس فيما بينها مما يؤثر على حدوث الأمراض.وكذلك نستنتج أن تلك المتغيرات التي تسببتة مرض قصور القلب وهي مصل الكرياتينين ، المصل صوديوم ، الصفائح الدموية ،النتاج القلبي (ejection fraction) ، الجنس والتدخين ، و ايضا مراقبة زمن حدوث المرض لدى المصابين من قبل الطبيب .

الكلمات الافتتاحية: تحليل العاملي ، التباين المشترك ، التباين المنفرد ، طريقة العامل الرئيسي ، أمراض القلب.