EPH - International Journal of Applied Science

ISSN (Online): 2208-2182
Volume 07 Issue 01-January-2021

DOI: https://doi.org/10.53555/eijas.v7i1.71

APPLYING FACTOR ANALYSIS TO STUDY THE MOST LIKELY FACTORS LEAD TO THE OSTEOPOROSIS DISEASE

Salem M. Algezeri*
Faculty of Science, Benghazi University, Benghazi-Libya

****Corresponding Author:- Email:***
salgzeri@yahoo.com

Abstract:-

The application of Biostatistics in biological and medical fields has shown a great contribution extremely benefit in analyzing different types of datasets. The developed statistical principles and techniques in the analysis have helped the researchers in the biology and medicine fields to reach to impressive results and beneficial conclusions. Recently, the Osteoporosis disease has taken a considerable interest from the medical and biological researchers. In this paper, a medical dataset related to this disease is analyzed using Factor Analysis via Principle Component approach. This statistical technique works to make a reduction of the insignificant explanatory variables and gives just main effected factories on the Osteoporosisdisease. The dataset used in this paper represent 180 real records selected randomly from the Osteoporosis patient files in Benghazi Central Hospital. The considered main variable in the statistical analysis is a rank variable represent the diagnosis patient state. The results achieved after this analysis show the main most likely factors lead to Osteoporosisdisease. The Vitamin D deficiency and patient gender have been found the most likely factors lead to this disease.

Keywords:- *Osteoporosis, bone metabolic disease, Factor Analysis, Vitamin D deficiency, Principle Component, Multivariate sampling and Best selection Modeling*.

INTRODUCTION

About half of all women and one fifth of all men aged 50 years or older will experience a fracture in their lifetime, due chiefly to underlying osteoporosis. Osteoporosis and its consequences place a significant burden on the health care systems of developed countries. Present therapeutic modalities are effective in reducing the risk of fractures caused by osteoporosis. However, we do not know whether the interventions introduced in the past 15 years have significantly reduced the number of osteoporotic fractures in real life, and if yes, how cost-effectively. Osteoporosis is characterized by decreased bone mass and typically presents with fractures of the wrist, spine and hip [(1)]. Biostatistics can be defined as the application of the mathematical tools used in statistics to the fields of biological sciences and medicine. Biostatistics is a growing field with applications in many areas of biology including epidemiology, medical sciences, health sciences, educational research and environmental sciences [(2)].

Osteoporosis occurs when there is an imbalance between new bone formation and old bone desorption. Two essential minerals for normal bone formation are calcium and phosphate. Throughout youth, the body uses these minerals to produce bones. Calcium is essential for proper functioning of the heart, brain, and other organs. To keep those critical organs functioning, the body reabsorbs calcium that is stored in the bones to maintain blood calcium levels. If calcium intake is not sufficient or if the body does not absorb enough calcium from the diet, bone production and bone tissue may suffer. Thus, the bones may become weaker, resulting in fragile and brittle bones that can break easily. Usually, the loss of bone occurs over an extended period of years. Often, a person will sustain a fracture before becoming aware that the disease is present. By then, the disease maybe in its advanced stages and damage may be serious [(7)]. The leading cause of osteoporosis is a lack of certain hormones, particularly estrogen in women and androgen in men. Women, especially those older than 60 years of age, are frequently diagnosed with the disease. The following are risk factors for osteoporosis:

- Women are at a greater risk than men, especially women who are thin or have a small frame, as are those of advanced age.
- Women who are white or Asian, especially those with a family member with osteoporosis, have a greater risk of developing osteoporosis than other women.
- Women who are postmenopausal, including those who have had early or surgically induced menopause, or abnormal or absence of menstrual periods, are at greater risk.

Factor Analysis

The concept used in *Factor Analysis* technique is to investigate the relationship among the group of variables and segregate them in different factors on the basis of their relationship. Thus, each factor consists of those variables which are related among themselves and explain some portion of the group variability. For example, disease infection of an individual can be assessed by the large number of parameters. The factor analysis may group these variables into different factors where each factor measure some dimension of disease infection. Factors are so formed that the variables included in it are related with each other in some way. The significant factors are extracted to explain the maximum variability of the group under study [(3)]

The *Principal Component Analysis* is a method provides a unique solution so that the original data can be reconstructed from the results. Thus, this method not only provides a solution but also works the other way round, i.e. provides data from the solution. The solution generated includes less than or as many factors as there are variables.

The *Common factor analysis* technique uses an estimate of common difference or variance among the original variables to generate the solution. The number of factors will always be less than the number of original factors. So, "factor analysis" commonly refers to common factor analysis.

If the observed variables are $X_1, X_2 \ldots X_n$, the common factors are $F_1, F_2 \ldots F_m$ and the unique factors are $U_1, U_2 \ldots U_N$, the variables may be expressed as linear functions of the factors:

$$X_1 = a_{11}F_1 + a_{12}F_2 + a_{13}F_3 + \ldots + a_{1m}F_m + a_1U_1$$

$$X_2 = a_{21}F_1 + a_{22}F_2 + a_{23}F_3 + \ldots + a_{2m}F_m + a_2U_2$$

$$\ldots.$$

$$X_n = a_{n1}F_1 + a_{n2}F_2 + a_{n3}F_3 + \ldots + a_{nm}F_m + a_nU_n \quad (1)$$

Each of these equations is a regression equation; factor analysis seeks to find the coefficients $a_{11}, a_{12} \ldots a_{nm}$ which best reproduce the observed variables from the factors. The coefficients $a_{11}, a_{12} \ldots a_{nm}$ are weights in the same way as regression

coefficients (because the variables are standardized, the constant is zero, and so is not shown). For example, the coefficient a_{11} shows the effect on variable X_1 of a one-unit increase in F_1. In factor analysis, the coefficients are called loadings (a variable is said to 'load' on a factor) and, when the factors are uncorrelated, they also show the correlation between each variable and a given factor. In the model above, a_{11} is the loading for variable X_1 on F_1, a_{23} is the loading for variable X_2 on F_3, etc.

When the coefficients are correlations, i.e., when the factors are uncorrelated, the sum of the squares of the loadings for variable X_1, namely $a_{11}\,2 + a_{12}\,2 + \ldots + a_{13}\,2$, shows the proportion of the variance of variable X_1 which is accounted for by the common factors. This is called the communality. The larger the communality for each variable, the more successful a factor analysis solution is. By the same token, the sum of the squares of the coefficients for a factor -- for F_1 it would be $[a_{11}\,2 + a_{21}\,2 + \ldots + a_{n1}\,2\,]$ -- shows the proportion of the variance of all the variables which is accounted for by that factor. Equation (1) above, for variable 2, say, may be written explicitly for one subject i as

$$X2i = a21F1i + a22F2i + a23F3i + \ldots + a2mFmi + a2U2i \qquad (2)$$

This form of the equation makes it clear that there is a value of each factor for each of the subjects in the sample; for example, F_{2i} represents subject i's score on Factor 2. Factor scores are often used in analyses in order to reduce the number of variables which must be dealt with. However, the coefficients a_{11}, a_{21}, …. a_{nm} are the same for all subjects, and it is these coefficients which are estimated in the factor analysis.[4]

DATA DESCRIPTION

The data are collected through a questionnaire designed to cover the most expected variables related to the Osteoporosis disease and randomly distributed, these variables are:

Variable	Description
x1	Patient's age group
x2	Vitamin D rate
x3	Diet
x4	**Family history**
x5	Patient's gender
x6	Physical activity

STATISTICAL ANALYSIS

It is difficult to decide which variables should adopted to be the main factors lead to osteoporosis, and hence the plan in this study will be based on two main stages: In the first stage, a group of variables have been selected under the assumption that all these variables are related to the osteoporosis disease. The statistical analysis is applied in the second stage to identify the most efficient variables **(**factors) causes this disease [8] The correlation matrix of the above variables is given by the following table

Correlations

		Patient's Age Group	Vitamine D Rate	Diet	Family History	Patient's Gender	Physical Activity
Patient's Age Group	Pearson Correlation	1	.842**	.322	.412	.766**	.348
	Sig. (2-tailed)		.002	.364	.237	.010	.325
	N	10	10	10	10	10	10
Vitamine D Rate	Pearson Correlation	.842**	1	.451	.610	.843**	-.116
	Sig. (2-tailed)	.002		.191	.061	.002	.749
	N	10	10	10	10	10	10
Diet	Pearson Correlation	.322	.451	1	.466	.641*	.005
	Sig. (2-tailed)	.364	.191		.174	.046	.989
	N	10	10	10	10	10	10
Family History	Pearson Correlation	.412	.610	.466	1	.811**	.067
	Sig. (2-tailed)	.237	.061	.174		.004	.854
	N	10	10	10	10	10	10
Patient's Gender	Pearson Correlation	.766**	.843**	.641*	.811**	1	.102
	Sig. (2-tailed)	.010	.002	.046	.004		.778
	N	10	10	10	10	10	10
Physical Activity	Pearson Correlation	.348	-.116	.005	.067	.102	1
	Sig. (2-tailed)	.325	.749	.989	.854	.778	
	N	10	10	10	10	10	10

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Now by using the PCA we can reduce the number of variables one such use is to simplify a regression analysis by reducing the number of predictor variables predict a dependent variable using the first few PC's determined from the predictors. But it is not clear how many factors to extract**.**

Component Matrix[a]

	Component	
	1	2
Patient's Age Group	.827	.361
Vitamine D Rate	.903	-.152
Diet	.659	-.230
Family History	.790	-.128
Patient's Gender	.977	-.037
Physical Activity	.134	.955

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

Suppose that two factors are extracted using the PCs, the patient's age group loads more highly (0.827) on factor 1than on factor 2 (0.361) but the loading on factor 2 is not that small so maybe patient's age group is distinctly related to both factors. The loadings are usually rotated and ordered to be better able to allocated them to factors

Rotated Component Matrix[a]

	Component	
	1	2
Patient's Gender	.968	.140
Vitamine D Rate	.915	.015
Family History	.801	.017
Patient's Age Group	.748	.505
Diet	.690	-.107
Physical Activity	-.041	.963

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

The first 5 variables load more highly on factor 1 than on factor 2 only physical activity loads more highly on factor 2 than factor 1 but factors with only 1 associated variable are usually under suspect. However, patient's age group loads highly on both factors, maybe it should be discarded since it is not unidimensional?

The initial communalities are usually estimated using the squared multiple correlations**.**[(5)]

In Equation (2), a part of each X is explained by the common factors, the communality for X is the amount of its variance explained by the common factors (hence its name).

Communalities

	Initial	Extraction
Patient's Age Group	1.000	.814
Vitamine D Rate	1.000	.838
Diet	1.000	.488
Family History	1.000	.641
Patient's Gender	1.000	.956
Physical Activity	1.000	.930

Extraction Method: Principal Component Analysis.

The communalities started out as all 1's since the PC method was used to extract factors.

Now, each factor F (or PC) has an associated eigenvalue also called a characteristic root since by definition it is a solution to the so-called characteristic equation for the correlation matrix **R.** The sum of the eigenvalues over all factors equals the total variance, and hence the eigenvalue measures how much of the total variance of the X's is accounted for by its associated factor (or PC). In other words, factors with larger eigenvalues contribute more towards explaining the total variance of the X's, they are generated in decreasing order $EV_1 \geq EV_2 \geq EV_3 \geq \ldots \geq EV_k$.[(6)]

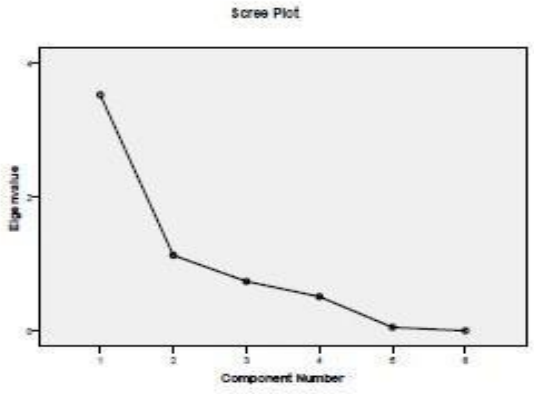
Eigenvalues at the start have the more important factors (orPC's) but they were reestimated based on loadings for the 2 extracted factors. The new values are less than 1 as they should be when the number of factors less than the number of items**.** The total variance analysis based on the eigenvalues is given by the following table:

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sum of Square Loading		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.531	58.844	58.844	3.531	58.844	58.844
2	1.136	18.927	77.770	1.136	18.927	77.770
3	0.746	12.432	90.202			
4	0.519	8.642	98.844			
5	0.061	1.010	99.855			
6	0.009	0.145	100.000			

Extraction Method : Principal Component Analysis.

The analysis of the eigenvalues says to extract 2 factors, since these two factors explain about 78% of the total variance. The graphical representation of these six components according to their corresponding eigenvalues is



Again the above graph supports the idea that most of variability has been explained in the first two components (factors) as the large change in slope biggest change is between components1 and 2.

DISCUSSION

There was a sequence of debates related to the main factors lead to Osteoporosis disease. However, all those debates did not based on scientific evidences either medical or statistical. All the previous attempts did not give a satisfactory confidence to put plans and rules to perform useful researches to study this disease. In this paper, the factor analysis through the principal component has been used to extract the main factors and hence to point out the most efficient variables on the Osteoporosis disease which are:

Gender The Females are more likely to be attacked with this disease compared to males.

Vitamin D Deficiency as the level of vitamin D decreases, the prognosis of the diseases increases. Age as the person become older his probability of having this disease increases.

ACKNOWLEDGEMENT

The author thank the medical administration of Benghazi Medical Center (BMC) for their cooperation, and in particular Dr. Samir Marghany whom provide us with the actual data from patient records.

REFERENCES

[1].Sambrook P, Cooper C (2006) Comparative statistical analysis of osteoporosis treatment based on Hungarian claims data and interpretation of the results in respect to cost-effectiveness. Lancet 367:2010–2018

[2].Armitage, P.; Berry, G.; Matthews, J.N.S. (2002), Statistical Methods in Medical Research, Blackwell, ISBN 978-0632-05257-8

[3].Cattell, R. B. (1952). Factor analysis. New York: Harper.

[4].Child, D. (2006). The Essentials of Factor Analysis, 3rd edition. Bloomsbury Academic Press.

[5].Gaddis, M. (1990) Introduction to biostatistics: Part 1, basic concepts. Departments of Surgery and Emergency Health Services, Truman Medical Center, University of Missouri, Kansas City, USA

[6].Maindonald, J. and Braun, W. (2009) Data Analysis and Graphics data and functions, R package version 1.01.

[7].Vanadin S.K. And Jerilynn C. P. (2010) Progesterone and Bone: Actions Promoting Bone Health in Women. Journal of Osteoporosis Volume 2010 (2010), Article ID 845180, 18 pages

[8].Zar, J. (1999) Biostatistical Analysis, Upper Saddle River, New Jersey: Prentice Hall.