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 Osteoporosis disease. 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 Osteoporosis disease. 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 (3). 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 X₁, X₂, ... Xₙ, the common factors are F₁, F₂, ... Fₘ and the unique factors are U₁, U₂, ... Uₙ, the variables may be expressed as linear functions of the factors:

\[
X_1 = a_{11}F_1 + a_{12}F_2 + a_{13}F_3 + \ldots + a_{1m}F_m + a_{11}U_1
\]

\[
X_2 = a_{21}F_1 + a_{22}F_2 + a_{23}F_3 + \ldots + a_{2m}F_m + a_{22}U_2
\]

\[
\vdots
\]

\[
X_n = a_{n1}F_1 + a_{n2}F_2 + a_{n3}F_3 + \ldots + a_{nm}F_m + a_{nU_1}(1)
\]

Each of these equations is a regression equation; factor analysis seeks to find the coefficients a₁₁, a₁₂, ... aₘₙ which best reproduce the observed variables from the factors. The coefficients a₁₁, a₁₂, ... aₘₙ are weights in the same way as in regression coefficients (because the variables are standardized, the constant is zero, and so is not shown). For example, the coefficient a₁₁ shows the effect on variable X₁ of a one-unit increase in F₁. 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, a₁₁ is the loading for variable X₁ on F₁, a₂₃ is the loading for variable X₂ on F₃, 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 \( [a_{11}^2 + a_{21}^2 + \ldots + a_{m1}^2] \) -- 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

\[
X_{2i} = a_{21}F_{1i} + a_{22}F_{2i} + a_{23}F_{3i} + \ldots + a_{2m}F_{mi} + a_{2U2i}
\]

(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 \( a_{11}, a_{21}, \ldots, a_{nm} \) are the same for all subjects, and it is these coefficients which are estimated in the factor analysis.

### 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:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>Patient's age group</td>
</tr>
<tr>
<td>x2</td>
<td>Vitamin D rate</td>
</tr>
<tr>
<td>x3</td>
<td>Diet</td>
</tr>
<tr>
<td>x4</td>
<td>Family history</td>
</tr>
<tr>
<td>x5</td>
<td>Patient's gender</td>
</tr>
<tr>
<td>x6</td>
<td>Physical activity</td>
</tr>
</tbody>
</table>

### 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. The correlation matrix of the above variables is given by the following table.

<table>
<thead>
<tr>
<th></th>
<th>Patients Age Group</th>
<th>Vitamin D Rate</th>
<th>Diet</th>
<th>Family History</th>
<th>Patient's Gender</th>
<th>Physical Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patients Age Group</td>
<td>Sig. (2-tailed)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Vitamin D Rate</td>
<td>Sig. (2-tailed)</td>
<td>.542**</td>
<td></td>
<td>.222</td>
<td>.412</td>
<td>.796**</td>
</tr>
<tr>
<td></td>
<td>.002</td>
<td>.364</td>
<td>.237</td>
<td>.016</td>
<td>.325</td>
<td></td>
</tr>
<tr>
<td>Diet</td>
<td>Sig. (2-tailed)</td>
<td>.843**</td>
<td></td>
<td>.610</td>
<td>.002</td>
<td>-.116</td>
</tr>
<tr>
<td></td>
<td>.002</td>
<td>.191</td>
<td>.061</td>
<td>.002</td>
<td>.749</td>
<td></td>
</tr>
<tr>
<td>Family History</td>
<td>Sig. (2-tailed)</td>
<td>.468</td>
<td></td>
<td>.174</td>
<td>.046</td>
<td>.986</td>
</tr>
<tr>
<td></td>
<td>.008</td>
<td>.191</td>
<td></td>
<td>.046</td>
<td>.986</td>
<td></td>
</tr>
<tr>
<td>Patient's Gender</td>
<td>Sig. (2-tailed)</td>
<td>.811**</td>
<td></td>
<td>.004</td>
<td>.857</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.004</td>
<td>.237</td>
<td></td>
<td>.004</td>
<td>.857</td>
<td></td>
</tr>
<tr>
<td>Physical Activity</td>
<td>Sig. (2-tailed)</td>
<td>.102</td>
<td></td>
<td>.004</td>
<td>.857</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.102</td>
<td>.018</td>
<td></td>
<td>.004</td>
<td>.778</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.778</td>
<td>.018</td>
<td></td>
<td>.004</td>
<td>.778</td>
<td></td>
</tr>
</tbody>
</table>

**. 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.
Suppose that two factors are extracted using the PCs, the patient's age group loads more highly (0.827) on factor 1 than 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.

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. 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).

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 $\mathbf{R}$. 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 $\text{EV}_1 \geq \text{EV}_2 \geq \text{EV}_3 \geq \ldots \geq \text{EV}_k$. Eigenvalues at the start have the more important factors (or PCs) 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:
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

![Graphical representation of eigenvalues](image)

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 component 1 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**


[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

