Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

# Comparison of Some Statistical Criteria for Determining Factors Affecting Anemia

**Nawras Majid Zaki**
**Asst. Prof. Dr. Ameena KareemEssa**
Statistics Department/ College of Administration and Economics/ Al-Mustansiriyah University
ameena@uomustansiriyah.edu.iq

***Article Info***
***Page Number: 4906 - 4920***
***Publication Issue:***
***Vol 71 No. 4 (2022)***

***Article History***
***Article Received****: 25 March 2022*
***Revised****: 30 April 2022*
***Accepted****: 15 June 2022*
***Publication****: 19 August 2022*

**Abstract**

Anemia was a widespread disease, occurring in persons where there were not enough red cells to transport oxygen throughout the body. Anemia was diagnosed at the level of simple and severe i.e. temporary and permanent, the normal ratio in males was 13.5 g/dL, Then the use of exploratory factor analysis and two criteria for comparison, namely the criterion of the least partial average and the criterion of parallel analysis, it was found through this study that the criterion of the least partial average had four factors, while the standard of parallel analysis had three influential factors, and the proportion of the explained variance of the criterion of the least partial average was higher than The percentage of the explained variance of the parallel analysis criterion.

## 1. Introduction

Factor Analysis is a mathematical process that aims to simplify the correlations between the various variables involved in the analysis to the common factors, which describe the relationship between these variables and their interpretation, so it is a statistical methodology for analyzing multiple data that have been associated with each other with different associations, the method of practical analysis is essentially based on the correlation factors between the variables.

The principal components method of Hottelling (1933), one of the characteristics of the principal compounds method is that each factor in it extracts the maximum possible variance, that is, the

sum of squares reaches its maximum in each factor, so the correlation matrix is summarized at least on a number of orthogonal factors, the basic feature is characterized by its ability to reach a solution that agrees with the squares of the statistical correlation matrix of the relationships between variables.

## 2. Factor analysis

The factor analysis method aims to summarize the multiple variables in a smaller number called (factors) so that each of these factors has a function that links it to some or all of these variables, and through this function it is possible to give an explanation for this factor according to the variables that are strongly associated with it, and this has arisen The method is mainly for analyzing psychological experiments and measures so that a certain set of tests can be traced back to the factor of intelligence and another to the factor of memory, and so on, although this does not mean that this method is not used in other fields. The idea of factor analysis is based on extracting a set of factors related to the original variables, so that these factors explain the largest possible percentage of variance in the original variables, and factor analysis can be used to convert a related group of variables to another independent group linked to the first group by linear relationships. In all cases, the relationship represents the original variables and factors in the form of equations as follows:

$$\underline{X} = A\underline{F} + \underline{U} + \mu \quad \text{.....(1)}$$

whereas :

$\underline{X}$ : Random vector variables.

$\underline{A}$ : Factor Loading

$\underline{F}$ : Common Factors

$\underline{U}$ : Unique Factors

$\underline{\mu}$ : Vector median variables.

The vectors of the means of each of the common and lone factors are zero vectors according to the assumption that the vector of the means of the variables is also zero, meaning that:

$$E(\underline{X}) = \underline{\mu} = 0 \quad \text{.....(2)}$$

$$E\begin{bmatrix} \underline{F} \\ \underline{U} \end{bmatrix} = \begin{bmatrix} \underline{0} \\ \underline{0} \end{bmatrix} \quad \text{……(3)}$$

So, the factorial model will be in the form:

$$\underline{X} = A\underline{F} + \underline{u} \quad \text{……(4)}$$

In other words,

$$\begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ . \\ x_p \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & . & . & . & a_{1q} \\ a_{21} & a_{22} & . & . & . & a_{2q} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ a_{p1} & a_{p2} & . & . & . & a_{pq} \end{bmatrix} \times \begin{bmatrix} f_1 \\ f_2 \\ . \\ . \\ . \\ . \\ f_q \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ . \\ . \\ . \\ . \\ u_p \end{bmatrix}$$

The covariance matrix for $\underline{F}$ and $\underline{u}$ (assuming they are independent) is:

$$E\begin{bmatrix} \underline{F} \\ \underline{u} \end{bmatrix} = \begin{bmatrix} \phi_{q.q} & 0_{q.p} \\ 0_{q.p} & \varphi_{p.p} \end{bmatrix}$$

Since:

$\phi$ : Covariance matrix for F

$\varphi$ : The diagonal matrix of covariance for u

And the covariance matrix for is:

$$E\left[\underline{X} : \underline{X}\right] - \left[E\left(\underline{X}\right)\right]^2 = \sum_{p.p}$$

Since:

$\sum$ : Symmetric Positive Matrix of the *p* order.

There are two hypotheses for factor analysis, the first is based on the existence of correlations between the variables in question as a result of the presence of common factors between them, and the factor model for *p* can be formulated from the observed variables from a sample of size *n* on the basis that it is a linear function of *q* of factors and as follows:

$$S_{ij} = a_{j1}X_{1i} + a_{j2}X_{2i} + ..... + a_{jq}X_{qi} \quad .....(5)$$

Since:

$S_{ij}$ : The standard value of the observation *i* with respect to the variable *j*.

$a_{jq}$ : Load the factor *q* with respect to the variable *j*.

$X_{qi}$ : Represents the real viewing value.

The other hypothesis of the factor analysis depends on the correlation coefficient between the variables (i, j) due to the nature of their saturation with common factors and the extent of this saturation. This hypothesis can be represented by the following equation for orthogonal factors:

$$r_{ij} = a_{i1}a_{j1} + a_{i2}a_{j2} + .... + a_{iq}a_{jq} \quad ……(6)$$

That is, the correlation coefficient between two variables is equal to the sum of the sum of the variables' loadings with their common factors. The previous equation can be rewritten as follows:

$$R = A.A'$$

Since:

*R*: correlation matrix.

*A*: Factor loading matrix.

**The Communalities:**

It is symbolized by $hj^2$, and represents the sum of squares of the saturations of each variable, which is the percentage of variance that is explained by the common factors resulting from the analysis of the correlation matrix *R*, that is, it gives the extent of overlap between the variables and the common factors. If the $hj^2$ of the variable is large and approaches one, it will show that this variable completely overlaps with the extracted factors. But if the $hj^2$ of one of the variables is equal to zero, the factor loadings (saturations) for that variable will be zero, meaning that the extracted factors could not explain any part of the variance of that variable. But if it falls between zero and one, it indicates a partial overlap between variables and factors.

## 3. Factor analysis methods

In this paper, we will discuss one of the methods of factor analysis, which is:

### 3.1 Principal component's method:

It is one of the most accurate, common and used factor analysis methods for the accuracy of its results compared to the rest of the methods. This method has many advantages, including that it leads to accurate saturations and that each factor extracts the maximum variance and that it leads to the least possible amount of residuals, and the correlation matrix reduces the amount of variance, to the least number of orthogonal and independent factors.

### 3.2 Objectives of the main component method:

1- Provide the information contained in the questionnaire in a simplified form.

2- Explanation of the largest percentage of variance for the original variables.

3- Representing the quantitative variables of geometric vocabulary based on the data table.
4- Determine the factors (components) that best explain the dispersion of variables.

## 4. Criteria for determining the number of factors

The problem of estimating the number of factors to be determined in the study is one of the problems facing researchers, because the possibility of extracting factors from the relational matrix to the extent that the last matrix becomes a zero-residue of possible things and where a number of factors can be extracted equal to the number of variables that we started with, Among the criteria that were used in this research are:

### 4.1 Minimum average partial:

Which was introduced as a test for determining the number of factors by Velicer in 1976 and is based on a matrix of partial correlation coefficients, where each factor that is derived from the matrix is excluded and then averaged, partial correlation coefficients, and the number of factors that are kept as the output of the analysis is determined by the point At which the least average of the square of the partial correlation coefficients is reached, and the rationale for this criterion can be clarified by knowing the relationship between the general or common variance and the unique or special variance in the correlation coefficients matrix. It provides an unambiguous criterion for determining the number of factors that are derived by isolating the general variance from the matrix about the special variance, and this can be explained in the following: -

- If we succeeded in identifying one of the factors on which a group of variables are saturated and the covariance of this factor was partially excluded from the matrix of correlation coefficients, then the average square of the partial correlation coefficients decreases.
- If this process continues until we reach the lowest mean of the partial correlation coefficients square, then we will have excluded the general variance - the joint - from the matrix of correlation coefficients.

Upon reaching the previous point, the derivation of any increase factors leads to the start of excluding the special (unique) variance from the matrix, which is followed by an increase in the average squares of the partial correlation coefficients. Partiality between variables will be high. The procedural steps to apply this test start with the formation of a matrix of correlation coefficients between the variables entered into the analysis, then the formation of a matrix of partial correlation coefficients for these variables, then calculating the average of squares of the partial correlation coefficients that make up this matrix, which is known as the least zero partial mean (MAP0) where no factors have been derived, and the next step It consists in excluding the covariance of the first common factor that can be derived and then calculating the average, the partial correlation coefficients that make up the matrix and this process continues for a number (P-1) of times where P represents the number of variables, and the number of factors in this case is corresponding to the lowest value of the mean square of the correlation coefficients It should be noted that if the mean square of the zero partial correlation coefficients is less than the mean of the partial correlation coefficients after excluding the covariance of the first factor (MAP0 < MAP1), this does not mean that it is square in size, there are common factors that can be derived from this matrix in the sense that the variables are not saturated on No common factors. In evaluating the performance of this criterion, the results of the study (1982, Zwick&Velicer) confirmed its accuracy compared to other criteria. In a study (1986, Zwick&Velicer) it was confirmed that this criterion is distinguished by its performance from the performance of the Keizer criterion and from the criterion of accumulation of latent roots, but it is less good than the parallel analysis criterion, and the results of this study confirmed the same results as the previous study with regard to the impact of the performance of this criterion by ramifications, and that Its performance tends to reduce the number of factors in the case of a small number of variables relative to the factor, and in another study, its results confirmed the previous results, where this criterion is biased in the case of low saturation, which is less than 0.40, as well as in the case of medium saturation, which is less than 0.55 in the case of a small number of saturated variables On the worker for 7 variables, and its results also confirmed that the performance of this criterion was not affected by the size of the sample. (Zwick&Velicer, 1982) confirms that the idea of this criterion is consistent with the idea on which the factor analysis is based - isolating the general

variance from the private variance and the variance of errors - and since the partial correlation coefficients estimate is not affected by the diagonal values of the correlation coefficients matrix, the performance of this criterion does not differ in factor analysis than in principal component analysis.

### 4.2 Parallel Analysis Standard:

This criterion was postulated by "Horn" in 1965 and his idea is to compare the latent roots of the factors resulting from the analysis of the matrix of correlation coefficients between the actual scores, with the average of the latent roots of a number of the correlation matrices of random scores for the same number of variables and the same number of individuals, and the factor in the case of actual scores. whose latent root is greater than the average of the latent roots of the corresponding factor in the case of random grades is an essential factor and is maintained, and this criterion appeared to overcome the problem of increasing the number of factors in the Keizer criterion. Factors whose latent roots exceed (1) appear as a result of random errors, and relying on the parallel analysis criterion and comparing the latent roots with the random latent roots controls this issue. Sample, number of variables, and rationale for parallel analysis although the variance explained by the factor must be greater than expected By chance, and that this criterion was initially designed for use in the analysis of principal components based on the original correlation coefficients matrix, and it needs further evaluation if it is used in factor analysis; Which is one of the objectives of the current study.

## 5. Data description

Here, (10) variables were taken for (80) people suffering from anemia, and below is a definition of the study variables:

X1 Serum folic acid ng/ml liter

X2 Blood viscosity pa

X3 Rheumatism RF factor in blood unit/ml liter

X4 Vitamin B12 in the blood pg/ml liter

X5 Patient weight kg

X6 Cumulative blood sugar percentage

X7 Vitamin C in the blood mg

X8 Serum cholesterol mg/dL

X9 Blood urea mg/dL

X10 Serum creatine mg/dL

Table (1)Descriptive roots statistics for the studied data

| Factor | Mean | Variance | Minimum | Median | Maximum | Skewness |
|---|---|---|---|---|---|---|
| X1 | 15.943 | 54.164 | 3.3 | 14.68 | 29.38 | 0.17 |
| X2 | 16.93 | 79.219 | 1.81 | 17.355 | 34.3 | 0.13 |
| X3 | 447 | 38668 | 102.1 | 425.1 | 816.6 | 0.11 |
| X4 | 98.52 | 358.68 | 65.88 | 96.53 | 129.8 | -0.02 |
| X5 | 7.896 | 3.725 | 4.62 | 7.825 | 11.45 | 0.21 |
| X6 | 63.14 | 260.94 | 36.35 | 62.48 | 89.82 | 0.05 |
| X7 | 136.75 | 1887.38 | 60.28 | 143.26 | 197.67 | -0.22 |
| X8 | 25 | 124.06 | 8.02 | 24.88 | 44.64 | 0.1 |
| X9 | 1.0385 | 0.016 | 0.8 | 1.02 | 1.25 | -0.05 |
| X10 | 1.7267 | 0.1504 | 0.96 | 1.715 | 2.51 | -0.16 |

Table (2)The normal distribution test for the study factors

| Variable | Statistic | Sig. | Variable | Statistic | Sig. |
|---|---|---|---|---|---|
| X1 | 0.095 | 0.07 | X6 | 0.09 | 0.167 |
| X2 | 0.075 | 0.214 | X7 | 0.098 | 0.055 |
| X3 | 0.089 | 0.176 | X8 | 0.076 | 0.227 |
| X4 | 0.081 | 0.257 | X9 | 0.089 | 0.178 |
| X5 | 0.082 | 0.265 | X10 | 0.085 | 0.286 |

Table (2) represents the results of the (Kolomogrov-Samernov test) (K.S.), which aims to find out whether the studied data in its standard form follow a normal distribution at the level of significance (5%), by testing the following hypothesis:

$H_0$: The data follows a normal distribution

$H_1$: The data does not follow a normal distribution

It turns out that the (p-value) of the (K.S) test for all factors of the study is greater than the level of significance, and this prompts us not to reject the above null hypothesis, and then all study factors follow a normal distribution.

Table (3)Correlation Matrix for Study Factors

| | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 | X10 |
|---|---|---|---|---|---|---|---|---|---|---|
| X1 | 1 | 0.004 | 0.516 | 0.133 | -0.607 | 0.561 | -0.09 | -0.019 | 0.043 | -0.603 |
| X2 | 0.004 | 1 | -0.016 | -0.095 | 0.068 | -0.094 | -0.089 | 0.15 | -0.164 | 0.189 |
| X3 | 0.516 | -0.016 | 1 | 0.133 | -0.653 | 0.644 | -0.055 | -0.077 | -0.121 | -0.539 |
| X4 | 0.133 | -0.095 | 0.133 | 1 | -0.099 | 0.188 | 0.033 | 0.049 | 0.249 | -0.199 |
| X5 | -0.607 | 0.068 | -0.653 | -0.099 | 1 | -0.64 | 0.064 | 0.041 | -0.087 | 0.664 |
| X6 | 0.561 | -0.094 | 0.644 | 0.188 | -0.64 | 1 | 0.028 | 0.005 | -0.028 | -0.586 |
| X7 | -0.09 | -0.089 | -0.055 | 0.033 | 0.064 | 0.028 | 1 | 0.164 | -0.089 | 0.063 |
| X8 | -0.019 | 0.15 | -0.077 | 0.049 | 0.041 | 0.005 | 0.164 | 1 | 0.019 | 0 |
| X9 | 0.043 | -0.164 | -0.121 | 0.249 | -0.087 | -0.028 | -0.089 | 0.019 | 1 | 0.071 |
| X10 | -0.603 | 0.189 | -0.539 | -0.199 | 0.664 | -0.586 | 0.063 | 0 | 0.071 | 1 |

The matrix of correlations for study factors and results was also calculated in the table (3), and it is clear from it that the correlations between the studied factors range from medium and weak correlations, and this indicates the existence of a correlation between some factors and the absence of the problem of multicollinearity in the studied data.

Table (4)KMO test to measure the adequacy of the sample for factor analysis

| | |
|---|---|
| KMO test | 0.765 |
| Chi-Square test | 226.714 |
| df | 45 |
| P-value | 0.000 |
| Correlation Matrix Determinant | 0.04833 |

The researcher conducted statistical analyzes on the study data based on the ready-made software packages available in the R program, and before starting to analyze the results, it was necessary to study the appropriateness of the sample size and the correlation matrix in order to complete the exploratory factor analysis procedures, and to achieve this, a specific correlation matrix was calculated and a test ( Bartlett) on the studied data and the KMO test, and it was clear from the results in the table (4), that the value of the determinant of the correlation matrix was (0.04833), which indicates that there is a linear relationship between the rows, or between the columns of the correlation matrix, and the value of the chi-square statistic reached Convergence (226.714), and the P-value is less than the level of significance, i.e. the correlation matrix used is different from the zero matrix, in addition to the KMO value has reached (0.765), and this indicates the sufficiency of the sample to perform the factor analysis.

Table (5)Criterion test less average partial

| | The mean square of the partial correlation coefficients | Link strength rate | | The mean square of the partial correlation coefficients | Link strength rate |
|---|---|---|---|---|---|
| 1 | 0.06500802 | 0.017188458 | 6 | 0.22510694 | 0.180344840 |
| 2 | 0.03403946 | 0.004141198 | 7 | 0.20155115 | 0.090081786 |
| 3 | 0.10390558 | 0.016482049 | 8 | 0.29883929 | 0.169747229 |
| 4 | 0.05600735 | 0.056875786 | 9 | 0.48623131 | 0.357758564 |

| | | | | | |
|---|---|---|---|---|---|
| 5 | 0.01785561 | 0.001055091 | 10 | 1 | 1 |

It is clear from the results of the table that the mean of the square of the partial correlation coefficients reached (0.017) at the fifth, and the largest value of the correlation strength rate was (0.0568) at the fourth compound, so the number of compounds included in the analysis is four.

Table (6)Distinctive roots, total variance ratios, and total variance ratios for least partial average criterion

| | Total | % of Variance | Cumulative % | | Total | % of Variance | Cumulative % |
|---|---|---|---|---|---|---|---|
| 1 | 3.129 | 31.29 | 31.29 | 6 | 0.794 | 7.94 | 81.62 |
| 2 | 1.216 | 12.16 | 43.46 | 7 | 0.572 | 5.72 | 87.34 |
| 3 | 1.124 | 11.24 | 54.70 | 8 | 0.473 | 4.73 | 92.08 |
| 4 | 1.029 | 10.29 | 64.99 | 9 | 0.423 | 4.23 | 96.31 |
| 5 | 0.868 | 8.68 | 73.76 | 10 | 0.368 | 3.68 | 100 |

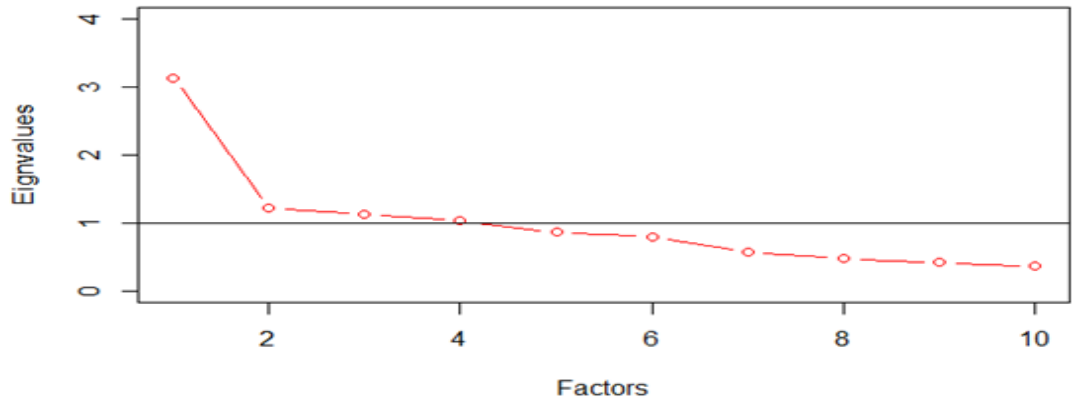


Figure (1)Distinctive roots of factor analysis using a minimum partial average criterion

The results in the previous table (6) and the subsequent figure (1) represent the application of factor analysis with a criterion less partial average, indicating that to build the inclusion of four

factors, whose distinctive root exceeded one, with an explanatory aggregate variance rate of 64.99%, as the first factor obtained On 31.299% of the total variance, while the second factor got 12.161% of the total variance, and the third factor got 11.242% of the total variance, while the fourth factor got 10.29% of the total variance.

To apply the parallel analysis criterion, 500 empirical correlation matrix was generated with the same characteristic roots and dimensions of the original correlation matrix, and the characteristic roots of each correlation matrix were calculated and then the rates of the characteristic root for those matrices were found, and compared with the characteristic roots of the original correlation matrix.

Table (7)Comparing the original characteristic roots with the averages of the characteristic roots of the empirical correlation matrices

| Factor | Eigen Values | Eigen Values Mean | Factor | Eigen Values | Eigen Values Mean |
|---|---|---|---|---|---|
| 1 | 3.473 | 1.4971324 | 6 | 0.691 | 0.8891770 |
| 2 | 1.336 | 1.2963115 | 7 | 0.507 | 0.7930879 |
| 3 | 1.195 | 1.1484800 | 8 | 0.377 | 0.6931162 |
| 4 | 1.066 | 1.1040676 | 9 | 0.332 | 0.5883591 |
| 5 | 0.787 | 1.0019960 | 10 | 0.237 | 0.4782724 |

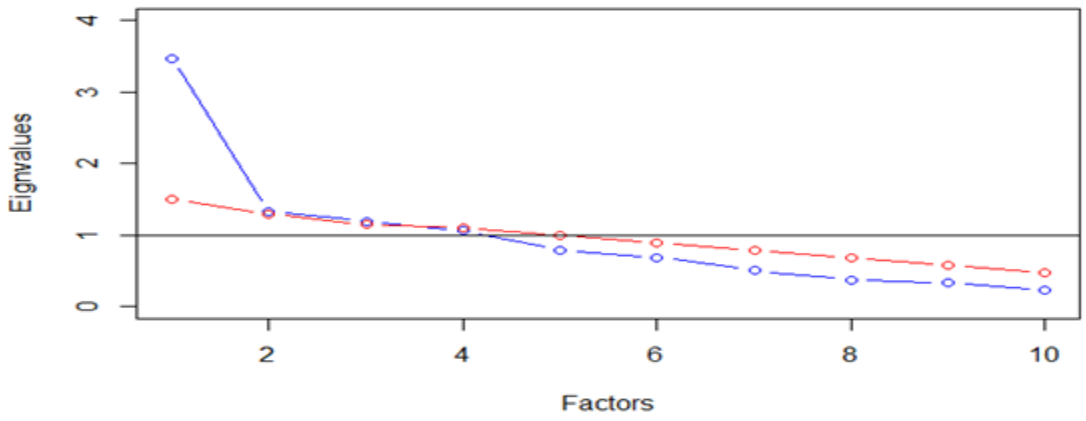


Figure (2) Distinctive roots of factor analysis using parallel analysis criterion

and the obtained results were put In the table (7) and Figure (2), it is clear from it that three characteristic roots, were greater than the rates of the empirical characteristic roots, and thus the number of compounds included in the analysis is three.

Table (8) Distinctive roots, total variance ratios, and total variance ratios for the criterion for parallel analysis

| Factor | total | %of variance | Cumulative% | Factor | total | %of variance | Cumulative% |
|---|---|---|---|---|---|---|---|
| 1 | 1.473 | 17.860 | 17.860 | 6 | 0.807 | 9.790 | 80.539 |
| 2 | 1.331 | 16.134 | 33.995 | 7 | 0.764 | 9.261 | 89.799 |
| 3 | 1.204 | 14.591 | 48.587 | 8 | 0.596 | 7.235 | 97.035 |
| 4 | 0.932 | 11.298 | 59.885 | 9 | 0.183 | 2.226 | 99.261 |
| 5 | 0.896 | 10.863 | 70.748 | 10 | 0.061 | 0.739 | 100 |

Among the results in the table (8), which represent the application of factor analysis with the parallel analysis standard, it can be said that three factors exceeded the characteristic roots of one with an explained aggregate variance rate of 48.587%, as the first factor obtained 17.86% of the total variance, while the second factor It obtained 16.134% of the total variance, and the third factor obtained 14.591% of the total variance.

**Conclusion**

1. The correlation matrix for the study variables shows that there is no multicollineartiy problem.
2. The factors of the least partial mean criterion are four (F1, F2, F3, F4) that exceed one distinct root, while the criterion for parallel analysis are three factors (F1, F2, F3) for distinct roots that exceed one.
3. Through the study, it was found that the proportion of the cumulative variance explained in the criterion of least partial mean was greater than the criterion of parallel analysis.
4. It is clear from the study that the best criterion is the least partial average criterion.

**References**

1. RabeaAbda Ahmed Rashwan. (2015). “The Performance of Criteria for Determining the Number of Factors in the Exploratory Factor Analysis of Measurement Tools in Psychological Research”.
2. Maysoon Abdel Hussein. (2020). College of Administration and Economics/ Al-Mustansiriya University. “Using the Main Compounds in Determining the Variables Affecting Autism in Children under Ten Years Old”.
3. Mohamed BouzianeTeghza. (2016). King Massoud University. “Exploratory and Confirmatory Factor Analysis”.
4. Anderson, T.W. (1984), “An Introduction to Multivariate Statistical Analysis”, Printed in the United States of America.
5. Ding and Xiaofeng (2004), " K-means Clustering via Principal Component Analysis" Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720.
6. Milewski, Acacio, Więsak and Jankowska (2014), "The Use of Principal Component Analysis and Logistic Regression in Prediction of Infertility Treatment
7. Sahalia and Xiu (2015), " Principal Component Analysis of High Frequency Data",Department of Economics Princeton University and NBER.
8. Saini, Tiwari and Gupta (2013), “A Simplified Approach for Interpreting Principal Component Images “, Physical Research Laboratory, Ahmedabad, India.
9. Smith (2002), "A tutorial on Principal Components Analysis" John wiley and Sons.
10. Zou,Hastie and Tibshirani (2006), " Sparse Principal Component Analysis " ,American Statistical Association, Institute of Mathematical Statistics, and Interface Foundation of North America
11. Iqbal,Asma,andNarjis,(2021) (Analysis and testing of the most important factors affecting (covid-19)),periodicals of engineering and Natural Sciences(vol. 9 ,No. 1),pp.(3-10).