Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol. 9, No. 1, January 2021, pp.3-10

Analysis and testing of the most important factors affecting (COVID-19)

Iqbal Mahmood Alwan [1], Narjis Hadi Irhaif[2], Asma Najim Abdullah [1]
[1]Department of Statistic, College of Administration and Economics, University of Baghdad, Baghdad, Iraq
[2]Department of Public Relations, College of Media, University of Baghdad, Baghdad, Iraq

ABSTRACT

Factor analysis is distinguished by its ability to shorten and arrange many variables in a small number of linear components. In this research, we will study the essential variables that affect the Coronavirus disease 2019 (COVID-19), which is supposed to contribute to the diagnosis of each patient group based on linear measurements of the disease and determine the method of treatment with application data for (600) patients registered in General AL-KARAMA Hospital in Baghdad from 1/4/2020 to 15/7/2020. The explanation of the variances from the total variance of each factor separately was obtained with six elements, which together explained 69.266% of the measure's variability. The most important variable is cough, idleness, fever, headache, palpebral, and difficulty in breathing. In the first factor and the variable appetite, not smelling, not to taste and diarrhea in the second factor: variables (sex, work, smoking, drinking alcohol) in the third factor, variables (diabetes, age, pressure) in the fourth factor, variables (vomiting, heartburn) in the fifth factor, variables (Blood group, drinking alcohol) in the sixth factor. Therefore, we must wash hands and covering mouths, or wearing a face mask when sneezing or coughing. Social distancing, disinfecting surfaces, ventilation, air-filtering, monitoring, and self-isolation are exposed or symptomatic.

Keywords: Factor Analysis, Principal Components Method, The Communalities, Factor Loading Matrix

Corresponding Author:

Iqbal Mahmood Alwan
Department of Statistic, College of Administration and Economics, University of Baghdad
Baghdad, Iraq
Email:iqbal.alwan@coadec.uobaghdad.edu.iq

1. Introduction

Factor analysis is one of the essential statistical methods aimed to examine the analytical phenomena extracted is the analysis to find the most important factors that influenced it through the calculation of the correlation coefficients between the variables of the phenomenon studied, and characterized by its ability to reduce the many variables and arrange them in a small number of linear compounds. The real history of factor analysis begins since *Galton* began measuring mental skills in an experimental direction in 1869. Several researchers in the literature had been discussed factor analysis like Priyaand Shruti [3], Kavit Srividhya, and Muthuselvan [6], Grimaccia and Naccarato [4].

Factor analysis [11] is one of the advanced statistical methods and related to multidimensional variables and depends on the calculation of the correlation coefficients between the phenomena studied and describes and explains the phenomena and attributes of the variables based on the lowest possible number of factors.
Factor model [3] is the mathematical model of the factor analysis consists of a set of variables seen (n) based on a function, taken from a sample of its size and number (Observed Variables) from (p) and to (p>q), for instance:
If X stands for random vector for observed variables, $\acute{X} = [X_1 \quad , \dots \dots , \quad X_P]$
B stands for factor loading matrix, $B = \begin{bmatrix} \lambda_{11} & \dots & \lambda_{1q} \\ \vdots & \ddots & \vdots \\ \lambda_{p1} & \dots & \lambda_{pq} \end{bmatrix}$
D is random Vector of Common Factors, $\acute{D} = [D_1 \quad , \dots \dots , \quad D_q]$

U is random Vector of Unique Factors, $\acute{U} = [U_1 \quad , \ldots\ldots , \quad U_P]$

So, we can write the model as follows:

$$\underline{X} = B\underline{D} + \underline{U} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

The q variables of common factor variable in D are independent variables with normal distributed with means zero and variances one.
The variable U is independent with normally distributed with means zero, but the variance is;

$var(u_i) = \psi_i$

The factor analyses aim to determine the loading matrix A and covariance matrix ψ

$$\sum = B\acute{B} \ + \ \psi$$

There are two types of factor analysis:

1 - Exploratory Factor Analysis [3]: This type is used in cases where the relationships between the variables and the underlying factors are not known, and thus factor analysis aims to
discover the factors to which the variables are described.

2- Confirmatory Factor Analysis [6]: This type is used to test hypotheses related to the existence or absence of a relationship between variables and underlying factors. Confirmatory factor analysis is also used to evaluate the ability of the factor model to express the actual data set and compare several factor models in this field.

Principal component method has several advantages, including that it leads to minute saturations, and each factor extracts the maximum amount of variance, and it leads to the least possible number of residuals, and the correlation matrix is reduced to the smallest number of orthogonal factors that are not related [2].

2. Basic assumption of factor analysis

This hypothesis is based on a correlation between a set of variables and that these correlations are the result of the presence of common factors among them. The factor analysis aims to interpret these correlations with factors that are less than the variables used and that this hypothesis takes the standard value for variables to obtain variables with normally distributed with mean zero and variance one to facilitate calculations as well as to get rid of the different units of measurement of variables if found. Under this hypothesis, there are three types of the total variation of variables [8]:

A - Common variation is the part of the variation associated with other variables through common factors and calculates general factors' coefficients.
B- Specific Variance is that part of the total contrast that is not associated with any variable but only with the variable itself.
C- Error Variance.
2- The assumption for orthogonal factors is the coefficient of correlation between two variables is equal to the sum of a load of variables multiplied by the factors in common between them, which is mean [1]
Kaiser Criterion [5] stands for the determining the number of factors. It is a mathematical criterion in its nature. His idea depends on the size of the variance expressed by the worker, and for the factor to act as a classification class, his variance or its underlying root must be greater or at least equal to the size of the original variance of the variable, and since we cannot Theoretically, extracting all the variance of the variable in one factor, if we obtain a root factor whose latent is not less than one correct, the source of the variance must be more than one variable and thus be a factor expressing a common variance between multiple variables.
Accordingly, this criterion requires reviewing the Eigenvalue root of the resulting factors and accepting the factors whose latent root is greater than the correct one and are general factors. The latter is very appropriate in many fields, mainly if the researcher uses the Harold Hotling essential components method. Therefore, this

method's indicative factors are the factors whose latent root is equal to or greater than one integer, provided that a correct one is placed in the diagonal cells.
Extraction factors [7], are drawn based on the selection of a set of variables that explain the most significant possible root of the total variation, which represents this as the first factor, and then the program selects a set of variables that describe the maximum possible variation after extracting the first factor, which constitutes this as the second factor... And so on.
The communalities [1], stand for the variable is a set of loading squares (saturation) of that variable and represents the contrast ratio explained by the common factors resulting from the analysis of the matrix R, i.e., it gives the extent of overlap between variables and common factors, the properties of communalities they are positive and between (0,1).
Factor loading [4], is the process of association of each variable with a particular factor, as the more significant the saturation of the factor than (0.3) the variable that has a relationship is well described, while the saturation of factors that are less than (0.3) are neglected.
Interpretation of factors [3] is the process of interpreting factors resulting from factor analysis is one of the most important problems facing researchers in the field of scientific research, as the idea of factor interpretation depends on the study variables, whether they are physical, skill, or functional tests, or the various aspects related to the factor and those that are not associated with it, by specifying the main, average, and zero saturation.
Stages of analysis [9],[10].
1- The R-matrix configuration stage is the matrix that contains correlation coefficients for all pairs of variables included in the analysis.
For the factor calculation stage, after completing the link matrix phase, factors are calculated, as principal components are one of the most commonly used methods.
For the rotating stage, factors are rotated for the relationships between variables. As some factors are possibly the most robust possible, and one of the most widely used methods is maximizing contrast (Vairmax), which is independent of the factors.

3. Application with real data

Coronavirus disease 2019 (COVID-19) was firstly recognized in Wuhan's Chinese city in 2019. The common symptoms of COVID-19 include fever, cough, fatigue, breathing difficulties, and loss of smell taste. Other symptoms include less common but may afflict several patients: the pains and aches, nasal congestion, headache, conjunctivitis, sore throat, diarrhea, loss of sense of taste or smell, and the appearance of a rash or color hands finger change or feet. Typically, these symptoms are mild and begin gradually. In addition, some people become infected without feeling only very mild symptoms [3]. Therefore, it was necessary to study, summarize and analyze these factors in the context of factor analysis to study a set of variables related realistic for patients registered into AL-KARAMA General Hospital in Baghdad for a period from (1-4-2020) to (15-7-2020) with a sample size (600) patients using factor analysis with (20) variables. SPSS (22) was used to run the data analysis as in Table 1.

Table 1. Adopted variables in SPSS simulator

Age = x1	Diarrhea = x11
Sex = x2	heartburn = x12
Work = x3	appetite = x13
Blood type = x4	Not smelling = x14
Palpebral = x5	No to taste = x15
Cough = x6	vomiting = x16
fever = x7	smoking = x17
headache = x8	Drinking alcohol = x18
Difficulty breathing = x9	Pressure blood = x19
idleness = x10	diabetes = x20

4. Result and discussion

Table 2 explains the outcomes of KMO and Bartlett Test of the dataset.

Table 2. KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.677
Bartlett's Test of Sphericity	Approx. Chi-Square	5191.358
	df	171
	Sig.	.000

Based on above, we observe that:

a- The Kaiser- Meyer-Olkin index of sampling adequacy value (KMO=0.677) verified the proposed analysis's sampling adequacy [6].

b- For Bartlett's Test of Sphericity, it indicates the strength of the relationship between the variables.
95% of the significance level $\alpha = 0.05$ $0.000 < 0.05$
If p-value =0.000 < 0.05, then the analysis is valid.
The correlations between the variables are all zero.
The approximate value of chi-square is 5191.358 with 171 degrees.
For further analysis, factor analysis is an appropriate technique.

The following Table 3 depicts the communalities of the dataset. Extraction method is based on principal component analysis.

Table 3. Communalites of the COVID-19

Variables	Extraction	Initial
Age = x1	.701	1.000
Sex = x2	.861	1.000
Work = x3	.821	1.000
Blood type = x4	.766	1.000
Palpebral = x5	.461	1.000
Cough = x6	.702	1.000
fever = x7	.638	1.000
headache = x8	.614	1.000
breathing = x9	.458	1.000
ldle = x10	.684	1.000
Diarrhea = x11	.448	1.000
heartburn = x12	.851	1.000
appetite = x13	.935	1.000
Not smelling = x14	.961	1.000
Not to taste = x15	.961	1.000
vomiting = x16	.814	1.000
smoking = x17	.549	1.000
Drinking alcohol = x18	.479	1.000
Pressure blood = x19	.496	1.000
diabetes = x20	.672	1.000

If the value's communality should be more than 0.5, then proceed to the further step for factor analysis; otherwise, these variables are removed from the additional factor analysis step. The above table shows that all the variables are above 0.5 except the function of COVID 19 Pedigree variable. So, we can proceed with further action for factor analysis. Table 4 discusses the explanation of the total variance of the dataset. Extraction method is based on principal component analysis.

Table 4. Total variance

Component	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	3.509	17.545	17.545	3.509	17.545	17.545	3.374	16.872	16.872
2	3.492	17.459	35.004	3.492	17.459	35.004	3.200	16.000	32.872
3	2.409	12.046	47.050	2.409	12.046	47.050	2.219	11.094	43.966
4	1.909	9.544	56.593	1.909	9.544	56.593	2.059	10.294	54.260
5	1.467	7.335	63.928	1.467	7.335	63.928	1.908	9.542	63.801
6	1.088	5.438	69.366	1.088	5.438	69.366	1.113	5.565	69.366
7	.900	4.502	73.869						
8	.749	3.745	77.614						
9	.724	3.621	81.236						
10	.678	3.389	84.625						
11	.584	2.921	87.546						
12	.568	2.842	90.388						
13	.468	2.338	92.726						
14	.439	2.196	94.923						
15	.317	1.585	96.507						
16	.297	1.485	97.992						
17	.186	.929	98.921						
18	.162	.809	99.730						
19	.054	.270	100.000						
20	-9.476E-17	-4.738E-16	100.000						

The eigenvalue reflects the number of extracted factors from the number of items included in the factor analysi]. The above table is divided into three segments like Sum of Squared Loadings Extraction, Initial Eigenvalues, and Sum of Squared Loadings. The first six variables' values are 17.545% of variance, 17.459 % of the variance, 12.046 % of the variance, 9.544% of the variance, 7.335 % of variance, and 5.438% of variance. So the first six components are taken for further analysis. The remaining variances are not significant.

From Figure 1, the components and eigenvalues are on X-axis and Y-axis, respectively, in the Scree Plot. The graph of figure 1 is to determine and to retain the factors. It is used to find the points in the curve where it is to start to flatten. The curve begins to flatten in 6. The eigenvalue is less than 1.0 component from 6 to the end component. Therefore, six factors above from 4 are removed from the factors.

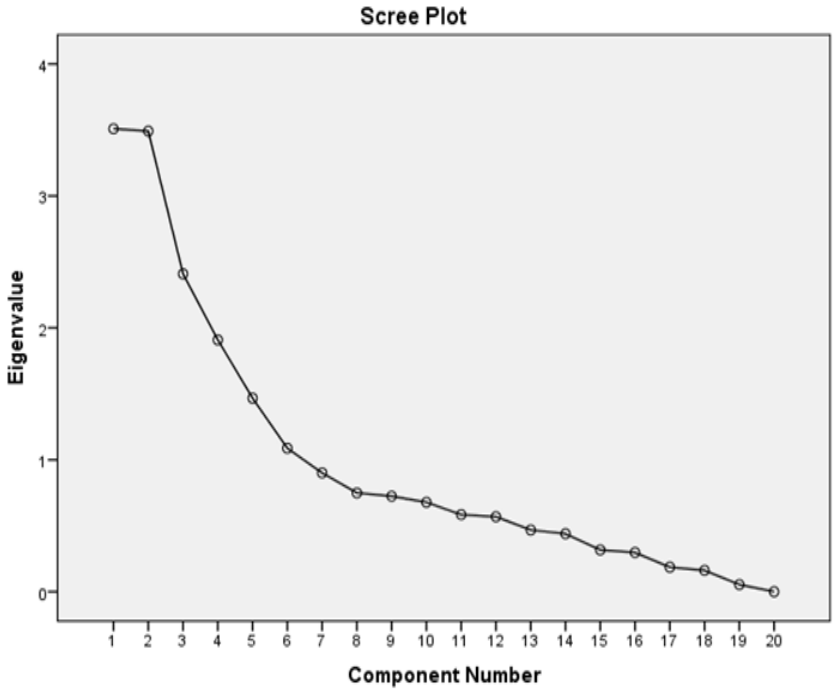

Figure 1. Plot of eigenvalue for the 20 items of the instrument to measure the COVID 19 using the main analysis components.

The component matrix presents the correlation of the component matrix about the variables in the dataset. Table 5 contains component loading and correlations between variable and component. The -1 to +1 ranges are the possible correlation values range.

Table 5. Component matrix

	Components					
Variables	1	2	3	4	5	6
Cough = x6	.762	.267	-.148	-.117	-.106	
ldle = x10	.740	.236	-.122	-.176	-.135	-.129
headache = x8	.694	.238	-.102	-.167		-.180
fever = x7	.664	.363	-.116		-.230	
breathing = x9	.570	.252				.242
Palpebral = x5	-.543	-.392				
Not to taste = x15	-.325	.905			.162	
Not smelling = x14	-.325	.905			.162	
appetite = x13	-.298	.874	.145		.238	
Sex = x2	.140		.798	.376	-.195	-.148
smoking = x17	.142		.712			.140
Work = x3	.145		.662	.534	-.208	-.161
diabetes = x20		-.121	.452	-.665		
Age = x1		.148	-.538	.621		
Pressure blood = x19			.329	-.551	-.219	.169
Drinking alcohol = x18		.153	.189	.500	-.221	.335
vomiting = x16	.437	-.301	.175	.139	.677	.157
heartburn = x12	.508	-.481	.219		.554	
Diarrhea = x11	.110	.378			.538	
Blood type= x4		-.126			-.155	.847

Table 6 shows the correlation of the rotated component matrix about variables in the dataset. Extraction method is based on principal component analysis. Table 7 gives details about the component transformation matrix. Rotation method has done by Varimax with Kaiser Normalization. Rotation has been converged in 5 iterations.

Table 6. Rotated component matrix

Variables	Components					
	1	2	3	4	5	6
Cough = x6	.832					
Idleness = x10	.807					
Fever = x7	.786					
Headache = x8	.750					
Palpebral = x5	-.645					
Breathing = x9	.607					
Appetite = x13		.959				
Not smelling = x14		.959				
Not to taste = x15		.959				
Diarrhea = x11		.497			.398	
Sex = x2			.912			
Work = x3			.899			
smoking = x17			.550	.402		
Drinking alcohol = x18			.472			.409
diabetes = x20				.818		
Age = x1				-.810		
Pressure blood = x19				.664		
vomiting = x16					.885	
heartburn = x12		-.339			.843	
Blood type = x4						.860

Table 7. Component transformation matrix

Components	1	2	3	4	5	6
1	.879	-.288	.130	.000	.357	-.035
2	.385	.876	.074	-.074	-.271	-.018
3	-.163	.116	.784	.550	.207	-.022
4	-.117	-.041	.543	-.821	.069	.106
5	-.199	.360	-.250	-.106	.853	-.175
6	.012	.077	-.080	.082	.158	.978

5. Conclusions

The diagonal of the anti-correlation matrix was also inspected for any values smaller than 0.5. Retention of a factor with eigenvalue greater than 1 to determine the optimal number of factors from the factor analysis, as well as the explanation of the variances from the total variance of each factor separately, and the six factors were obtained, which together explained 69.266% of the variability of the measure.

Six factors were highlighted and, together, explained 69.266% of the measurement variance as follows:

Factor 1 has strong relationships with 6 variables (Cough = x6, idleness = x10, fever = x7, headache = x8, Palpebral = x5, Difficulty breathing = x9) and they Explained (17.545%) of the variance of the measurement which is mean the factor most important.

Factor 2 has strong relationships with 4 variables (appetite = x13, not smelling = x14, not to taste = x15, Diarrhea = x11) and they Explained (17.459%) of the variance of the measurement.

Factor 3 has strong relationships with 4 variables (sex = x2, work = x3, smoking = x17, drinking alcohol = x18) and they Explained (12.046%) of the variance of the measurement.
Factor 4 has strong relationships with 3 variables (diabetes = x20, Age = x1, pressure = x19) and they Explained (9.544%) of the variance of the measurement.

Factor 5 has strong relationships with two variables (vomiting = x16, heartburn = x12), and they Explained (7.335%) of the variance of the measurement

Factor 6 has a strong relationship with two variables (blood type = x4, drinking alcohol=x18). They Explained (5.438%) of the variance of the measurement.

Suggested preventive actions involve hand cleaning, social distancing, wearing a facemask, sterilizing surfaces, ventilation and air filtering, and monitoring and self-isolation if exposed or symptomatic.

References

[1] G. L. Canivez, M. W. Watkins and R. J. McGill., "Construct validity of the Wechsler Intelligence Scale for Children–fifth UK edition: Exploratory and confirmatory factor analyses of the 16 primary and secondary subtests", British *Journal of Educational Psychology*, vol.89, no.2, pp.195-224, 2019.
[2] N. Chen, M. Zhou, Dong X, J.Q. F. Gong, Y. Han, et al., "Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study", *Lancet*, vol.395, no.10223, pp.507–513, 2020.
[3] J. L. Coker., D. Catlin, S. Ray-Griffith., B. Knight and Z. N Stowe, "Buprenorphine medication-assisted treatment during pregnancy: An exploratory factor analysis associated with adherence," *Drug and Alcohol Dependence*, vol.192, pp.146-149, 2018.
[4] E. Grimaccia and A, Naccarato, "Confirmatory *and Quality& Quantity," International Journal of Methodology*, vol. 54, no.4, pp. 1211-1232, 2020.
[5] L. R. Fabrigar and D. T.Wegener, Exploratory factor analysis, Oxford University Press, 2011.
[6] A. Filed, "Discovering statistics using spss," ISM introducing statistical methods," vol.2, 2005.
[7] S. Kavit. E. Srividhya and S. Muthuselvan., "Prediction of Diabetics using factor analysis," *International Journal of Recent Technology and Engineering*, vol. 7, no.6, 2019.
[8] R. Gomez, V. Stavropoulos, M.D. Griffiths, "Confirmatory factor analysis and exploratory. Structural equation modeling of the factor structure of the Depression Anxiety and Stress Scales" PLoS *ONE*, vol.15, no.6, e0233998, 2020.
[9] J. Murray, BFA: Bayesian factor analysis, R package version 0.4, 2016.
[10] M. Can and O. Gürsoy, "Clustering 16S rRNA for OTU prediction: A similarity based method", Heritage and Sustainable Development, vol. 1, no. 2, pp. 78-83, Dec. 2019.
[11] J. W. Osborne, Best Practices in Exploratory Factor Analysis. Louiseville, NY: Create Space Independent Publishing Platform, 2014.