ISSN: 2688-9323

The Theory and Application of Multivariate Linear Factor Analysis

Jianxin Zheng[1, a] ,Rui Niu[2, b]

[1]School of Mathematical Sciences, TIANGONG University, Tianjin, China

[2] Tianjin Zhengneng Culture Media Co., Ltd. Tianjin, China

[a]747158067@qq.com, [b]m13920392038@163.com

Abstract

This article is mainly based on multiple linear regression to study factor analysis and principal component analysis. First introduce the mathematical model of factor analysis and principal component analysis. In the process of introducing factor analysis and principal component analysis, practical applications are made based on local GDP, per capita GDP, and per capita disposable income. Finally, make some reasonable suggestions for the results obtained in the application. Try to provide more application methods for factor analysis and principal component analysis.

Keywords

Factor analysis, Multiple linear, Correlation analysis.

1. Introduction

A common problem in research is to study the influence of several variables or factors on some corresponding factors Y [1]. For a food production company preparing to produce water-baked cakes, it must study the precise amount of flour, fat, sugar, milk or water, eggs, baking powder and other materials, and also strictly control the baking temperature and time. Therefore, there are many factors that make people pay attention to, and any one of them will affect the taste and quality of the cake. Similarly, a research project aimed at understanding how a country can increase the production of major grains might try to measure the impact of different fertilizers on the production of nitrogen, phosphorus and potassium. This type of problem often occurs in the industry. For complex chemical processes, the final product may have as high as 10 to 20 influencing factors.

In the early years, it was often the research of one factor. It was necessary to experiment with each factor separately. Later, Fisher pointed out that there is a great advantage in studying the combination of several factors in the same factor experiment, because each observation provides all the information contained in the factor for the experiment. In addition, factor experiment is a systematic method to study the relationship between different factors.

Multiple linear regression is very complicated because of the problems involving multiple indicators. [2] The purpose of adding observations is to make the research process more complete and the factors considered in the research more comprehensive, but from another perspective, in order to make the observation results clear, adding observation indicators will make people confused. In fact, indicators are often related to a certain degree, so people want to use fewer indicators instead of the original more indicators, but the premise is that they can still fully reflect all the information. This idea is "dimensionality reduction." Analysis methods such as factor analysis have come into being. [3]

At the beginning of the 20th century, Carl Pearson and Charles Spearman proposed factor analysis in their statistical analysis experiments on intelligence testing. [4] Finding the basic

ISSN: 2688-9323

structure of variables to simplify the observation system is to reduce the dimension of variables. For complex problems, a few variables can be used to solve the purpose of factor analysis. Now, factor analysis has been widely used in the fields of psychology, medicine, meteorology and economics. Principal component analysis and factor analysis are basically similar in thinking. They also use "dimensionality reduction" to study the basic structure of observation data by studying the internal dependencies between many variables, and use several abstract variables to express its basic data structure. Factor analysis is a statistical analysis method that evaluates important variables of latent variables by using specific indicators to evaluate abstract factors. Extract some comprehensive indicators to reflect the overall situation [5].

2. Factor Analysis Model

2.1. Factor

The premise of factor analysis is the least amount of information loss, and the factor is the integration of many original variables into a few comprehensive variables. This idea can be expressed by a mathematical model, that is, there are p original variables x1, x2, ..., xp, and each variable or after normalization has a mean value of 0 and a standard deviation of 1. Now put Each original variable is represented by a linear combination of k (k<p) factors f1, f2, ..., fk, there is:

$$\begin{cases} x_1 = a_{11}f_1 + a_{12}f_2 + a_{13}f_3 + \cdots + a_{1k}f_k + \varepsilon_1 \\ x_2 = a_{21}f_1 + a_{22}f_2 + a_{23}f_3 + \cdots + a_{2k}f_k + \varepsilon_2 \\ \vdots \\ x_p = a_{p1}f_1 + a_{p2}f_2 + a_{p3}f_3 + \cdots + a_{pk}f_k + \varepsilon_p \end{cases} \quad (1)$$

The above formula is called the output model of factor analysis, and its matrix expression is:

$$X=AF^T+\varepsilon \quad (2)$$

Among them, F is a factor, because the linear expression in each original variable has them, so it is also called a common factor.

2.2. Factor Characteristics

In the factor analysis of multiple linear regression, the factors have the following characteristics: (1) The number of factors is much smaller than the number of original variables. Because factor analysis and principal component analysis both use the idea of "dimensionality reduction", which simplifies complex data and reduces data dimensions. Therefore, the number must be much smaller than the number of original variables, otherwise factor analysis and principal component analysis will lose their significance. After the original variables are combined into several factors, the factor will replace the original variables to participate in data modeling, thus effectively overcoming the problem of too many variables caused by defects in the analysis.

(2) Factors can reflect most of the information of the original variables. If the factors are not the original variables, the result is not a simple choice. It will not cause a large loss of the original variables, and most of the information of the original variables can be extracted. And whether it is factor analysis or principal component analysis, the least information loss is a prerequisite.

(3) The linear relationship between factors is not significant. When there are more variables, it will increase the complexity of the analysis problem, because there may be a certain correlation between the variables, resulting in overlap of information among multiple variables. In order to avoid the phenomenon of information overlap, we must use principal component analysis

ISSN: 2688-9323

and factor analysis. Therefore, there will not be a strong linear relationship between the recombined factors.

(4) Factors have naming explanatory properties. Generally speaking, the factors generated by factor analysis and principal component analysis can obtain the final naming explanatory character in various ways. The explanatory nature of factor naming can help evaluate the results of factor analysis and is of great significance for the further application of the factor.

2.3. Factor Loading

For factor models:

$$X_1 = a_{11}f_1 + a_{12}f_2 + \ldots + a_{1k}f_k + \varepsilon_i \tag{3}$$

Among them, a_{ij} is the factor load, which is the load of the i-th variable on the j-th factor. After finishing the above formula, the covariance of $x_i$and f $_i$is as follows:

$$\begin{aligned} \mathrm{Cov}(x_i, f_j) &= \mathrm{Cov}\left(\sum_{k=1}^{p} a_{ik}f_k + \varepsilon_i, f_j\right) \\ &= \mathrm{Cov}\left(\sum_{k=1}^{p} a_{ik}f_k + f_j\right) + \mathrm{Cov}(\varepsilon_i + f_j) \\ &= a_{11} \end{aligned} \tag{4}$$

If x_i is standardized, the standard deviation of x_i is 1, and the standard deviation of f_j is 1, so:

$$\frac{\mathrm{Cov}(x_i, f_j)}{\sqrt{D(X_i)}\sqrt{D(f_i)}} = \mathrm{Cov}(x_1, f_j) = a_{1j} \tag{5}$$

It can be seen from the above analysis that the correlation coefficient between x_i and f_j is standardizedx_i a_{ij}, that is to say, the correlation coefficient between variable x_i and factorf_j is the factor load, which reflects the fact that the variable xi and The degree of correlation of the factor f_j. If the factor load in the conclusion is larger, it proves that the relationship between the i-th variable and the j-th factor is closer; the smaller the factor load is, it proves that the relationship between the i-th variable and the j-th factor is more distant. Moreover, the factor loading is also a reflection of the important influence and degree of the factor f_j on the variable x_i

2.4. Factor Contribution to Variance

Suppose the factor loading matrix is A, we call the sum of squares of the elements in the jth column:

$$g_j^2 = \sum_{i=1}^{k} a_{ij}^2 \tag{6}$$

(j=1, 2, 3. . . , k) It is the contribution of factor f_j to variable x, that is to say, g_j^2 represents the sum of the contrast contributions provided by the same factor $f_j$to each variable, and also reflects the ability of factorf_j to explain the total variance of the original variable.

ISSN: 2688-9323

The higher the variance contribution value, the higher the importance of the corresponding factor. So the scale to measure the relative importance of each factor is the factor variance contribution and the variance contribution rate.

Principal component analysis decomposes the total variance of p original variables $x_1, x_2, x_3, \ldots, x_p$ into the sum of variances of p independent variables $y_1, y_2, y_3, \ldots y_p$ then: $\varphi_k = \lambda_k / \sum_{k=1}^{p} \lambda_k k$ is called the variance contribution rate of the k-th principal component yk. The greater the contribution rate of the first principal component, the stronger the ability of y1 to integrate the original variables, while the overall ability of $y_1, y_2, y_3, \ldots, y_p$ decreases in turn.

If only m principal components are taken (m<p), then:

$$\psi_m = \sum_{k=1}^{m} \lambda_k / \sum_{k=1}^{p} \lambda_k$$

It is the cumulative contribution rate of the m principal components. The cumulative contribution rate represents the y1, y2, y3, ..., ym integrated x1, x2, x3, ..., xp capabilities. In general, the cumulative contribution rate of m is taken to reach a higher percentage.

3. Factor Analysis Results

KMO and spherical Bartlett test are used as the applicability test of factor analysis.

In SPSS, the Bartlett sphere test can be used to judge whether the variables are correlated. If the correlation matrix is a unit matrix, then the variables are independent, which shows that the factor analysis method is not useful.

Table 1. KMO and Bartlett's test

Kaiser-Meyer-Olkin measure of sampling adequacy	**.637**
Bartlett's sphericity test approximate chi-square	**246.692**
Df	**15**
Sig.	**.000**

According to the Bartlett test in Table 1, it can be concluded that the hypothesis that each variable is independent should be rejected, which means that the variables have a strong correlation. The KMO statistic is 0.637<0.7, reflecting that the degree of overlap of information between various variables may not be very high. It is possible that the factor analysis model obtained is not very perfect, but it is still worth trying.

Table 2. Common factor variance

classification	initial	extract
Area GDP	**1.000**	**.929**
Area per capita GDP	**1.000**	**.921**
Disposable income per capita	**1.000**	**.944**
Resident RMB savings deposits	**1.000**	**.922**
Resident consumption level	**1.000**	**.969**
Illiterate population	**1.000**	**.825**

The common factor variance is used to express the extent to which the common factor of each variable can extract the information in the original variable. From the extraction results shown

ISSN: 2688-9323

in Table 2, it can be seen that the commonality of all variables is more than 80%, so these extracted factors the explanatory ability of a common factor for each variable is very strong.

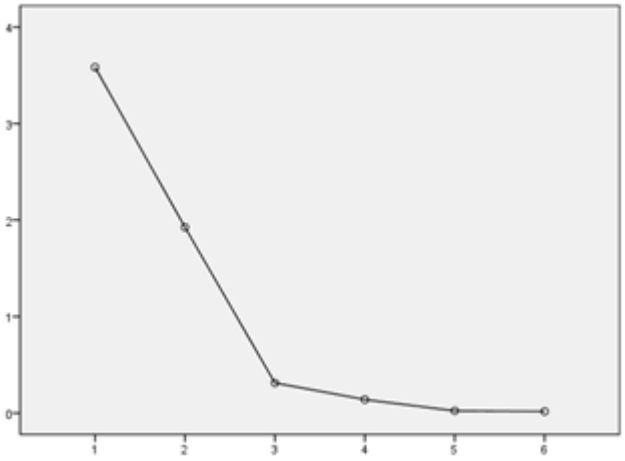

Figure 1: Gravel map

Figure 1 is a lithotripsy diagram, which is used to express the importance of each factor. The horizontal axis of the lithotripsy diagram is the factor number, and the vertical axis is the characteristic root size. He will sort according to the characteristic roots from small to large, so that you can intuitively see which factor is the most important. The steep part corresponds to the larger characteristic root, and the effect is obvious. The flat part corresponds to the smaller characteristic root, and its influence is not significant. In Figure 1, the scatter points of the first two factors form a steep slope, and the scatter points of the last four factors are located on a flat platform, and the characteristic roots are all less than 1, so it is enough to consider the first two common factors at most.

Table 3. Explained total variance

ingredient	Initial eigenvalue			Extract the sum of squares and load			Rotate the sum of squares loading		
	Total	Variance %	Grand total%	Total	Variance %	Grand total%	Total	Variance %	Grand total%
1	3.584	59.729	59.729						
2	1.925	32.091	91.820						
3	.312	5.194	97.014	3.584	59.729	59.729	3.199	53.309	53.309
4	.140	2.327	99.341	1.925	32.091	91.820	2.311	38.511	91.820
5	.023	.386	99.727						
6	.016	.273	100.000						

The eigenvalues, variance contribution rate and cumulative contribution rate calculated from the correlation coefficient matrix R are shown in Table 3. It can be seen that only the first two eigenvalues are greater than 1, so SPSS only extracts the first two common factors. The cumulative contribution rate of the variance of the latter two common factors has changed after the rotation, and the variance contribution rate of the first two factors is still 91.82%, which is exactly the same as before the rotation, so the selection of the first two factors is sufficient to describe the income influencing factors of the original premium.

ISSN: 2688-9323

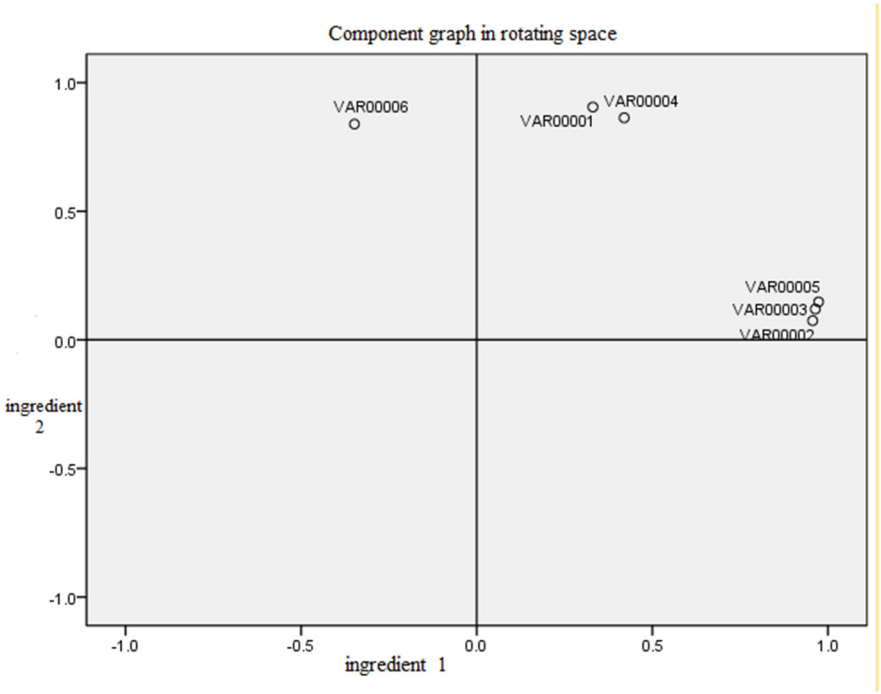

Figure 2: Component diagram in rotating space

Table 4. Rotation component matrix

	Ingredient	
	1	2
Resident consumption level (yuan)	.973	.148
Per capita disposable income (yuan)	.964	.120
GDP per capita (yuan)	.957	.074
Gross Regional Product (100 million yuan)	.331	.905
Resident RMB savings deposits (100 million yuan)	.420	.863
Illiterate population (number)	-.348	.839

The following formula can be obtained from the rotation component matrix in Table 4:

$$\begin{cases} X_1^* = 0.973\mathrm{F}_1 + 0.148\mathrm{F}_2 \\ X_2^* = 0.964\mathrm{F}_1 + 0.120\mathrm{F}_2 \\ X_3^* = 0.957\mathrm{F}_1 + 0.074\mathrm{F}_2 \\ X_4^* = 0.331\mathrm{F}_1 + 0.905\mathrm{F}_2 \\ X_5^* = 0.420\mathrm{F}_1 + 0.863\mathrm{F}_2 \\ X_6^* = -0.348\mathrm{F}_1 + 0.839\mathrm{F}_2 \end{cases}$$

It can be seen that the first common factor has a large load onX_2, X_3, and X_5, which mainly reflects the influencing factors of the original premium income from the consumption level, per capita disposable income, and per capita gross regional product. It can be named as the resident consumption factor. The second common factor has a large load onX_1, X_4, andX_6. It mainly reflects the influencing factors of the original premium income from the gross regional product, the renminbi savings deposits of residents, and the number of illiterate people, which can be named the quality of life factor. Comparing with before rotation, it can be seen that the meaning of the common factors after rotation is obviously more clear and reasonable.

ISSN: 2688-9323

4. Conclusion

From the final regression model, we can see that the consumption and quality of life of residents in various regions have affected the income of premiums. Although the six variables I chose have correlations, they are all positively correlated from the regression model, so we propose the following Suggest:

Efforts to increase insurance awareness. With the emergence of new forms of materials and commodities, people will also enjoy new risks. This requires people to establish and develop risk awareness in the new situation. The government can increase the scope and intensity of risk awareness education to make it a universal education, and it can also cultivate professional talents. Through training and publicity, the effect of popularization is an important issue.

References

[1] Zhang Wentong, Dong Wei. Advanced Course of SPSS Statistical Analysis [M]. Beijing: Higher Education Press, 2000.

[2] Liu Dahai, Li Ning, Chao Yang. SPSS15.0 statistical analysis from entry to proficiency[M]. Beijing: Tsinghua University Press, 2008.

[3] Xiao Qian. Research on China's Insurance Statistics [D]. Wuhan University, 2014.

[4] Lu Yuanhong. Methods of Mathematical Statistics [M]. Shanghai: East China University of Science and Technology Press, 2005.

[5] Huang Jindan. The application of mathematical statistics in analysis and testing [J]. Fujian Analysis and Testing, 2016, 25: 32-34.