Factor Analysis: Some Basic Principles and an Application

By Frederick V. Waugh

Economists have paid very little attention to the statistical methods known as "factor analysis" or "component analysis." These methods were developed primarily by psychologists. Sociologists have also used them to some extent. These methods may eventually help economists to solve some of their problems. With this in mind, the author has tried to set forth the basic principles of factor analysis, and to indicate how the U.S. Department of Agriculture has used them to establish county level-of-living indexes for farm families in the United States. Those readers who are interested in a full, detailed, theoretical treatment of the subject might well read Harman's recent book which also includes an extensive bibliography.[1] Hagood and Price [2] discuss applications to sociological research. Tintner [3] discusses uses in economic research. The author thanks Ralph Champion, Farm Population Branch, ERS, for his help.

COUNTY INDEXES of levels of living are weighted averages of certain county census data, such as the percentages of farms with telephones, automobiles, home freezers, average value of farm products sold, and average value of land and buildings. Factor analysis helps determine which census elements to use and how to weight each element.

We shall show here that factor analysis maximizes the variance of the index, and that it also maximizes the sum of the squared correlations between the index and the several elements. Maximum variation in the index enables it to discriminate effectively between high, medium, and low levels.

To avoid excessive length, this problem is illustrated by considering an index based upon only three elements, say, X_1, X_2, and X_3. If I is the index, it can be written

$$I=K+b_1X_1+b_2X_2+b_3X_3, \tag{1}$$

where K, b_1, b_2, and b_3 are constants to be determined in the analysis.

If lower-case letters are used to indicate deviations from the national averages, equation (1) can be written

$$i=b_1x_1+b_2x_2+b_3x_3 \tag{2}$$

If the x's are standardized by dividing each by its standard deviation, equation (2) can be written

$$i=w_1z_1+w_2z_2+w_3z_3 \tag{3}$$

where w_1, w_2, w_3 are the weights to z_1, z_2, z_3.

The weights should be assigned in such a way as to provide a great deal of variation in the index so that it will discriminate most effectively between counties that have high, medium, and low levels of living.

The variance of the index is

$$\text{var } i=\frac{1}{n}\sum i^2=w_1^2+2r_{12}w_1w_2 +2r_{13}w_1w_3+w_2^2+2r_{23}w_2w_3+w_3^2, \tag{4}$$

where r_{12}, r_{13}, and r_{23} are zero-order correlation coefficients.

As it stands, this variance has no maximum; that is, it could be increased indefinitely by multiplying all of the w's by a constant greater than 1. For instance, if each w were multiplied by 2, the variance of i would be multiplied by 4. To avoid this arbitrary result, we propose to maximize the variance of i, subject to the condition (or restraint) that the sum of the squared weights is unity; that is, so that

$$w_1^2+w_2^2+w_3^2=1 \tag{5}$$

To do this, we shall use a Lagrange multiplier. Such multipliers are discussed in modern texts on

1. Harry H. Harman, *Modern Factor Analysis.* Univ. of Chicago Press, 1960.

2. Margaret Jarman Hagood and Daniel O. Price, *Statistics for Sociologists.* Henry Holt & Co., New York, rev. ed., 1952. Ch. 26.

3. Gerhard Tintner, *Econometrics.* Wiley and Sons, New York, 1952. Ch. 6.

643432—62——1

advanced calculus [4] and in some books on economic theory.[5] In order to maximize (4) subject to condition (5) we can maximize

$$F=\text{var } i-g(w_1^2+w_2^2+w_3^2-1), \tag{6}$$

where g is known as a Lagrange multiplier. We can maximize F in equation (6) by differentiating it with respect to w_1, w_2, w_3, and g, and setting all the derivatives equal to zero. This gives us

$$\begin{aligned}(1-g)w_1+r_{12}w_2+r_{13}w_3&=0\\ r_{12}w_1+(1-g)w_2+r_{23}w_3&=0\\ r_{13}w_1+r_{23}w_2+(1-g)w_3&=0\\ w_1^2+w_2^2+w_3^2&=1\end{aligned} \tag{7}$$

While these equations are necessary for a maximum solution, they are not sufficient. In other words, equations (7) may be satisfied with some values of g that result in a minimum F and some may result in neither a minimum nor a maximum. In order to have a true maximum, it is necessary that equations (7) be met, and also that the Hessian matrix

$$H=\begin{bmatrix}1-g & r_{12} & r_{13}\\ r_{12} & 1-g & r_{23}\\ r_{13} & r_{23} & 1-g\end{bmatrix} \tag{8}$$

be negative definite; that is, all diagonal elements of H must be negative, all principal 2-row determinants must be positive, all 3-row principal determinants must be negative, and so on. Thus, it is clear that in order to have a maximum solution, g must be greater than 1. In general, there will always be several values of g that will satisfy equations (7). In the case of three variables, there will be three such values of g. In the case of n variables, there will be n different values of g that will satisfy the equations. Any value of g that satisfies the equation is known as the root of the matrix shown in (8). We are concerned here with the maximum positive root, which is often known as the "dominant" root. In order to understand the meaning of the dominant root and also to give the basis of an iterative method of solving equations (7) the equations can be written in the form

$$\begin{aligned}w_1+r_{12}w_2+r_{13}w_3&=gw_1\\ r_{12}w_1+w_2+r_{23}w_3&=gw_2\\ r_{13}w_1+r_{23}w_2+w_3&=gw_3\end{aligned} \tag{9}$$

If the first equation is multiplied by w_1, the second by w_2, and the third by w_3, and three equations are added, the result is

$$\text{var } i=g(w_1^2+w_2^2+w_3^2) \tag{10}$$

Since the squares of the weights are subject to condition (5), equation (10) indicates that the dominant root g equals the variance of the index. Thus, as is often the case, the Lagrange multiplier g turns out to have a real statistical meaning.

The correlation between the z's and the index can be found by multiplying equation (3) successively by z_1, by z_2, and by z_3, and dividing each product by the standard deviation of the index (i.e., by $\sqrt{g}$). Then using (9), we find that

$$\begin{aligned}r_{z_1 i}&=\frac{w_1+r_{12}w_2+r_{13}w_3}{\sqrt{g}}=\sqrt{g}\cdot w_1\\ r_{z_2 i}&=\frac{r_{12}w_1+w_2+r_{23}w_3}{\sqrt{g}}=\sqrt{g}\cdot w_2\\ r_{z_3 i}&=\frac{r_{13}w_1+r_{23}w_2+w_3}{\sqrt{g}}=\sqrt{g}\cdot w_3\end{aligned} \tag{11}$$

The sum of the squared correlations is

$$r_{z_1 i}^2+r_{z_2 i}^2+r_{z_3 i}^2=g(w_1^2+w_2^2+w_3^2) \tag{12}$$

Since the sum of the squared weights is 1, the Lagrange multiplier also equals the sum of the squared correlations between the index and its elements. In the process of maximizing the variance of the index, we also maximize the sum of squared correlations.

Three Problems of Computation

There are three main practical problems in computing the index: (1) solving equations (7) to determine the weights associated with any particular set of census items, (2) comparing the results obtained by forming indexes from dif-

[4] Angus E. Taylor, *Advanced Calculus*. Ginn & Co., Boston. 1955. pp. 198–204.

[5] Paul A. Samuelson, *Foundations of Economic Analysis*. Harvard Univ. Press. 1958. pp. 362–365.

ferent groups of census items and determining which group to select, and (3) transforming the index into the most usable form. Each of these will be discussed briefly here. Then a numerical problem will be analyzed to illustrate the procedures.

Solving Equations (7)

Equations (7) can be solved by either of two processes: First, one could set the determinant of the coefficients of the first three equations equal to zero, solve for g, and then solve for the w's. In a problem of only three variables, this is a reasonably simple operation. However, when dealing with a large number of variables, such a direct solution is very time-consuming and laborious. In any case, it is possible to solve the equivalent equations (9) by an iterative process. One simply starts with any assumed initial values of the w's, inserts them into the left side of equation (9), and computes the second approximations of the weights. This process will be illustrated in the solution of the numerical example.

Comparing Sets of Census Items

Indexes can be computed for various sets of 3, 4, 5 or more census items. In general, we want an index that has a large variance. In comparing the results of several indexes, each of which is based on n items, we should ordinarily choose the one that gives the greatest variance. Sometimes, however, the addition of an item to an index may not increase the variance "significantly." Here there are perhaps no hard and fast rules for choosing between an index based on n variables and an index based on $n+1$ variables. Some statisticians have used as a general guide the variance divided by the number of items in the index. This is the same as the average of the squared correlations between the index and the several elements. I do not think that this sort of guide should be followed rigidly nor uncritically. If it were, one would always choose an index based upon a single element. Then the squared correlation would be 1, since the correlation of any variable with itself is 1. No index based upon more than one element could give an average squared correlation higher than 1.

The variance (or sum of squared correlations) in the 3-variable problem to be discussed next is 2.456—an average of 0.815. The final index used in this report is based upon 5 variables. Its variance (or sum of squared correlations) is 3.115—an average squared correlation of 0.623. Yet the additions of two items not only seems desirable from general considerations, but also raises the variance of the index by almost 27 percent.

Scaling the Index

The index so far described is in terms of the z's. Specifically, it can be written

$$I=w_1z_1+w_2z_2+w_3z_3 \tag{13}$$

But each $z_i=x_i/s_i$, where s_i is the standard deviation of X_i. So the index can be written

$$w_1\frac{X_1}{s_1}+w_2\frac{X_2}{s_2}+w_3\frac{X_3}{s_3} \tag{14}$$

Commonly we want to scale the index in such a way that it meets two criteria: (1) when all the items are 0, the index is to be 0; and, (2) when the value of each item in the index is at the national average, the index should be 100. We can scale the index by determining a constant k such that

$$k\left(w_1\frac{\overline{X}_1}{s_1}+w_2\frac{\overline{X}_2}{s_2}+w_3\frac{\overline{X}_3}{s_3}\right)=100, \tag{15}$$

where $\overline{X}_1$, $\overline{X}_2$, $\overline{X}_3$ are the national averages of X_1, X_2, X_3.

Then the scaled index is

$$I_s=\frac{kw_1}{s_1}X_1+\frac{kw_2}{s_2}X_2+\frac{kw_3}{s_3}X_3 \tag{16}$$

A Numerical Example

To illustrate the methods described above, let us construct a level-of-living index based upon only three Census items in each county. These three items will be the percentages of farms with telephones, with homefreezers, and with automobiles. The zero-order correlation coefficients are

$$R=\begin{bmatrix}1.000 & 0.633 & 0.823\\ 0.633 & 1.000 & 0.683\\ 0.823 & 0.683 & 1.000\end{bmatrix} \tag{17}$$

where the order of the three variables is as indicated just above (17).

To get weights for the index we must solve a set of equations such as (7) or the equivalent (9). Equations corresponding to (9) are

$$
\begin{aligned}
w_1+0.633w_2+0.823w_3&=gw_1\\
0.633w_1+w_2+0.683w_3&=gw_2\\
0.823w_1+0.683w_2+w_3&=gw_3
\end{aligned}
\tag{18}
$$

This set of equations has the obvious trivial solution $w_1=w_2=w_3=0$. It also has three nontrivial solutions, each with a different value of g. We are interested here in a solution with the largest positive value of g.

One way of finding such a solution is an iterative process, starting with an arbitrary set of trial values, such as $w_1=w_2=w_3=1.0$. (Here we are concerned only with the proportions between the weights. We do not yet require that the sum of the squared weights equal 1.) Substituting $w_1=w_2=w_3=1.0$ into (18), we get as a second approximation, $gw_1=2.456$, $gw_2=2.316$, $gw_3=2.506$, These numbers are shown in the first column of the lower part of (19). We could use these numbers as second approximations of our weights, but to keep the numbers comparable we shall divide each of them by the first number (i.e., by 2.456). This gives us as a second approximation the weights 1.000, 0.923, 1.020. This process can be continued indefinitely, and will eventually converge to the correct set of weights. When that happens it will be apparent that further iterations will not significantly change the weights.

A simple worksheet for this process is as follows:

Trial values of the w's

(1)	(2)	(3)	(4)	(5)	(6)
1.000	1.000	1.000	1.000	1.000	1.000
1.000	0.943	0.933	0.931	0.930	0.930
1.000	1.020	1.021	1.021	1.021	1.021

(19) Estimates of gw from (18)

2.456	2.436	2.431	2.430	2.429
2.316	2.273	2.263	2.261	2.260
2.506	2.487	2.481	2.480	2.479

After four iterations we get the set of weights 1.000, 0.930, 1.021. These are apparently correct to three decimals. These numbers are shown in the last column at the top of (19).

The sum of the squares of these numbers is 2.9073, and the square root of this sum is 1.7051. If we want the sum of squares of the index weights to equal 1.0, we must divide each of these numb by 1.7051. This gives us the weights $w_1=0.58$, $w_2=0.545$, $w_3=0.599$. Thus, one form of the index would be $I=0.586z_1+0.545z_2+0.599z_3$. The variance of this index, in other words the value of g, is 2.429. Also, the sum of the squared correlations between the index and the three items is 2.429. This number is found at the top of the last column in the lower part of (19). It is easy to see that this number equals gw_1, for in this case w_1 is taken as 1.000. Thus the first equation in (18) shows that $2.429=g(1.000)$.

The correlations between each of the three items and the index can be found by multiplying the above weights by $\sqrt{g}$ (i.e. by 1.5586). This gives us the correlations

$$r_{z_1 I}=0.913,\ r_{z_2 I}=0.849,\ r_{z_3 I}=0.934$$

The standard deviations of X_1, X_2, X_3 were $s_1=23.92$, $s_2=14.66$, $s_3=15.18$. Remembering that each $z_i=x_i/s_i$, the equation $I=0.586z_1+0.545z_2+0.599z_3$ can be written.

$$I=0.586\,\frac{x_1}{23.92}+0.545\,\frac{x_2}{14.66}+0.599\,\frac{x_3}{15.18}$$

$$I=0.0245x_1+0.0372x_2+0.0395x_3 \tag{20}$$

This index would be zero for any county with no telephones, no freezers, and no automobiles. But we would like to scale the index so that a county with the average percentages of telephones, freezers, and automobiles would have an index of 100. The national averages (unweighted means of county means) were $\overline{X}_1=64.26$, $\overline{X}_2=55.57$, $\overline{X}_3=79.19$.

Inserting these averages into the above equation (20), we would have an index of 6.7696. Multiplying all the weights in the above equation by $100/6.7696=14.77$, we get the scaled index

$$I_s=0.362X_1+0.549X_2+0.583X_3 \tag{21}$$

Using this scaled index, a county with no telephones, no freezers, and no automobiles would get an index of zero. Also a county in which 64.26 percent of the farms had telephones, 55.57 percent had freezers, and 79.19 percent had automobiles would score

$$I_s=0.362\ (64.26)+0.549\ (55.57)+0.583\ (79.19)=1$$