Chapter 2

Technique for Estimation of Distributed Weights in Cross-referencing Commodity Classification

NODA Yosuke

Because definitions of commodities and the specific coverage of these definitions will not necessarily be identical before and after the year in which a revision of commodity classifications is put into effect, when trade statistics are used as long-term time series data, claims cannot be made for the continuity and consistency of transaction values and quantities before and after the year of revision of the classification. When conducting a time series analysis, it is therefore necessary to achieve consistency between the data before and after the revision using a common classification system. This is enabled by a process of conversion involving the calculation of distributed weights on the basis of cross referencing of classifications before and after the revision year, and the use of these weights to redistribute statistical values corresponding to the respective classification codes.

The purpose of this chapter is to discuss several methods used by Institute of Developing Economies : IDE to calculate distributed weights based on cross referencing of commodity classifications before and after revision using trade statistics as data. Conversion on the basis of the results of these calculations enables formulation of long-term time series data employing a single unified commodity classification system.

Where a cross reference table exists for individual mutually related classification codes in the different commodity classifications *A* and *B*, cross references between *A* and *B* for all individual classification codes can be expressed in the form of a table. A number of closed cross references can be formulated by appropriate rearrangement of *A* and *B* on the cross reference table. These closed relationships between *A* and *B* are termed commodity groups. *n* elements of *A* within the commodity group, $A_1 \cdots A_n$, correspond to the statistical values $x_1 \cdots x_n$. x_i is represented by a vector formulated from *k* samples, and $x_j = (x_{j1} \cdots x_{jk})'$ for $j = 1 \cdots n$. The statistical values corresponding to *m* elements of *B* in the same commodity group, $B_1 \cdots B_m$, are termed $y_1 \cdots y_m$. y_j is represented by a vector formulated from *k* samples, and $y_i = (y_{i1} \cdots y_{ik})'$ for $i = 1 \cdots m$. $x_\bullet$, the total of the statistical values of x_i for *A* in the commodity group, is preserved without alteration in *B* after conversion, and is therefore identical to $y_\bullet$, the total of statistical values of y_j .

Looking at conversion from *A* to *B*, the distributed weight for conversion from classification code A_j to B_i is termed ω_{ij} . Here we assume that all cross references exist, and therefore $\omega_{ij} \neq 0$. y_i , for $i = 1 \cdots m$, can be expressed as

$$y_i = x_1\omega_{i1} + \cdots + x_n\omega_{in} + u_i$$

where the distributed weight is $\omega_{1j} + \cdots + \omega_{mj} = 1$ for $j = 1 \cdots n$ and u_i is a disturbance term for a vector having the same structure as y_i. Expressing this as a matrix gives us

$$\begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = \begin{pmatrix} \omega_{11} & \cdots & \omega_{1n} \\ \vdots & & \vdots \\ \omega_{m1} & \cdots & \omega_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} u_1 \\ \vdots \\ u_m \end{pmatrix}$$

Terming the distributed weight matrix of the m×n matrix $W = (\omega_1 \cdots \omega_m)'$. The vector of the m dimension, all the elements of which are composed by 1, is termed l_m. Given the weight conditions, W satisfies

(1) $$l_m{}'W = l_n{}'$$

Given the existence of k statistical values, the statistical value matrix for B is an m×k matrix which we term Y. The statistical value matrix for A is an n×k matrix we term X. In the same way, the m×k matrix for the disturbance terms is designated as U. When cross references exist between all codes, the structure of the distributed weights for the transaction values can be expressed as

(2) $$Y = WX + U$$

based on the weight conditions in equation (1.

It is normal that weights for which there is no cross reference in a commodity group exist among the commodity classification cross references. Weights for which 0 elements exist are termed W_g. Like the weight conditions in equation (2), W_g satisfies

(3) $$l_m{}'W_g = l_n{}'$$

and can be expressed as

(4) $$Y = W_g X + U$$

Given this, the purpose of this chapter is to use equations (3) and (4) in calculating the distributed weight matrix W_g when the cross references between A and B in the commodity groups and the statistical values X and Y corresponding to the respective classifications have been obtained. When the distributed weight matrix W_g from A to B has been obtained, it is a simple matter to calculate the distributed weight matrix in the opposite direction, from B to A.

1. Equally distributed weight

Equally distributed weights are estimated using only the number of commodities in the commodity group cross-referenced on the basis of the lowest level of the commodity code used in the classification system, without consideration of transaction value. Estimation in this method is based on whether or not correspondence exists, and therefore for corresponding commodities, the spread of the distributed weights is uniform, i.e. they are equally distributed.

The function $a(W_g)$ is 1 when cross references exist between the respective elements and 0 when cross references do not exist. Equal distribution can be assumed for the distributed weights in a distributed weight matrix showing only existing cross references, and can therefore be formulated by making $a(W_g)$ satisfy the weight conditions in equation (3). That is, if the distributed weight matrix using equal distribution is designated $W(e)$,

(5) $$W(e) = a(W_g) \cdot D(l_m{}'a(W_g))^{-1}$$

2. Estimation of weights with equal pattern

Estimation of weights with equal pattern is a method in which transaction value is considered. First, we assume that $x_\bullet = y_\bullet$ for the total transaction value of the commodity group, based on the assumption that all the cross references exist in the group. The distributed values for the cross references from A_j to B_i is $x_j\omega_{ij}$, the total value of A_j is x_j, and the total value of B_i is y_i. If A and B are assumed to be independent, the distributed value is given by $x_j\omega_{ij} = x_\bullet(y_i / x_\bullet)(x_j / x_\bullet)$ Therefore, the distributed weight is $\omega_{ij} = y_i / x_\bullet$, and is entirely dependent on *i* and independent of *j*.

If $\omega_{ij}=\omega_i$ for $j=1\cdots n$, the distributed weight matrix in which all correspondences are established is $W=(\omega_1\cdots\omega_m)' l_n'$. From this equation, we derive

$$(\omega_1\cdots\omega_m)' = Yl_k/(l_n' Xl_k) .$$

Because 0 elements exist in the normal distributed weight matrix, areas with no cross references are adjusted using $a(W_g)$. The matrix is reformulated to satisfy the weight conditions, assuming

$$W_2 = D(\omega_1\cdots\omega_m)a(W_g) .$$

A distributed weight matrix with equal patterns,

(6) $$W(p)=W_2\cdot D(l_m' W_2)^{-1}$$

is given.

3. Ordinary least squares method

The ordinary least squares method is a method of estimating the quantity of *W*, in which all correspondences are assumed to be established, using the least squares of *U* in equation (2). We define the least squares method as a method of minimizing total variation. That is, given that

(7) $$s=tr(UU')$$

for *U* in equation (2), $\hat{W}$ is made the solution when equation (7) is partially differentiated using the elements of *W* with substitution of 0 as

$$\hat{W}=YX'(XX')^{-1}.$$

0 is substituted for the negative elements of $\hat{W}$ and we obtain W_2.

Further, considering cross references $W_3=W_2\bullet a(W_g)$, where the operator $\bullet$ represents the product of each of the elements corresponding to the matrix., $A\bullet B=(a_{ij}b_{ij})$. If W_3 is reformulated to satisfy the weight conditions, the distributed weight matrix with no restrictions applied,

$$W(l)=W_3\cdot D(l_n' W_3)^{-1}$$

is given.

4. Restricted least squares of weights method

The restricted least squares of weights method is a method of estimating the quantity of *W*, in which all correspondences are assumed to be established, using the least squares of *U* in weight condition equations (1- and (2). Total variation is the foundation of the least squares method. That is, given that

(8) $$s_2=tr(UU')+(l_m' W-l_n')2\lambda$$

the solution when equation (8) is partially differentiated using the elements of *W* with substitution of 0 is termed $\widetilde{W}=(I_m-m^{-1}L_{mn})\hat{W}-m^{-1}L_{nn}$. 0 is substituted for the negative elements of $\widetilde{W}$ to obtain W_2. Taking cross references into consideration, gives us $W_3=W_2\bullet a(W_g)$. Reformulating to satisfy weight conditions, the distributed weight matrix with restrictions, $W(rl)=W_3\cdot D(l_n' W_3)^{-1}$ is given.

5. Restricted least squares of weights method in regression

In this method, the least squares method is reexpressed in the form of a regression model. Summarizing each element of this equation, we have $y_i=X'\omega_i+u_i$ for $i=1\cdots n$. Expressed as a matrix, the equation becomes

$$\begin{pmatrix} y_1\\ \vdots\\ y_n\end{pmatrix}=\begin{pmatrix} X' & & \\ & \ddots & \\ & & X'\end{pmatrix}\begin{pmatrix}\omega_1\\ \vdots\\ \omega_m\end{pmatrix}+\begin{pmatrix}u_1\\ \vdots\\ u_n\end{pmatrix}$$

Summarizing this gives us

(9) $$y=X^*\omega+u$$

If the unit matrix of the *m* dimension is designated I_m, the matrix formulated from *n* units of I_m arranged horizontally is given as $C=(I_m\cdots I_m)$. Because vector ω expresses the weights,

(10) $$C\omega=\omega_1+\cdots+\omega_n=l_m$$

In commodity classification cross references, it is normal for weights for which there is no cross reference to exist in the commodity group, and 0 elements are included in the elements of ω_i. An adjusted vector is a vector which does not have 0 as an element. Conversion of the adjusted vector enables us to obtain $\omega_i^D = D_i\omega_i$, $i = 1\cdots m$. Because $D_i = I_m$ for ω_i^D, if $(\omega^D)' = ((\omega_1^D)' \cdots (\omega_m^D)')$ and

$$D = \begin{pmatrix} D_1 & & \\ & \ddots & \\ & & D_m \end{pmatrix}$$

then $\omega^D = D\omega$. Because the same operations are required for the observation matrix corresponding to vector ω, X^*, and C, $X^{*D} = X^*D'$ and $C^D = CD'$.

(9') $$y = X^{*D}\omega^D + u^D$$

corresponds to equation (9) for ω^D, X^{*D} and C^D after adjustment of cross references, and

(10') $$C^D\omega^D = l_m$$

corresponds to equation (10). The Lagrange multiplier method is used to obtain $\tilde{\omega}^D$ for ω^D, the minimum of $(u^D)'u^D$, the residual sum of squares in equation (9'), based on the restriction conditions satisfying equation (10'). Assuming

$$M = [(X^{*D})'X^{*D}]^{-1}(C^D)'\bullet$$
$$\{C^D[(X^{*D})'X^{*D}]^{-1}(C^D)'\}^{-1}$$

estimation by restricted least squares method gives

(11) $$\begin{aligned}\tilde{\omega}^D &= \hat{\omega}^D - M(C^D\hat{\omega}^D - l_m)\\ &= (I - MC^D)\hat{\omega}^D + Ml_m\end{aligned}$$

6. Iterative scaling procedure

The $m \times n$ matrix, the elements of which are positive real numbers, is termed W_g. The values of its elements, ω_{ij} are unknown. Given vectors y and x, the values of which are known, and the $m \times n$ matrix $W_g^{(0)}$ of which all the elements are known, the marginal sum of W_g is $W_g l_n = y$, $l_m' W_g = x$, $x_\bullet = y_\bullet$. The iterative scaling procedure, or ISP, is a statistical method of calculating all the elements of matrix W_g. To ensure that

(12) $$G^{(2k-1)} = G^{(2k-2)}D(l_m' G^{(2k-2)})^{-1}D(x)$$

when the number of iterations is an odd number, and

(13) $$G^{(2k)} = D(y)D(G^{(2k-1)}l_n)^{-1}G^{(2k-1)}$$

when the number of iterations is an even number for $k = 1\cdots n$, iteration is conducted with equations (12) and (13) combined. Therefore, when k is increased without restriction for all integers and $G^{(k)} \to W_g$ uniquely.

To provide actual examples of the estimation of distributed weights, we will present the results of estimations for the cross referenced commodity group 212 from 4-digit level classification codes of SITC-R2 to ones of SITC-R1, obtained using a variety of different methods. The distributed weights were estimated for data for exports from Japan, drawn from AID-XT base data. AID-XT data is UN COMTRADE data available online, adjusted for use in the Ajiken trade statistics data system.