Chapter3 Estimation of Distributed Weights Cross-referencing Commodity Classifications: Towards the Formulation of SITC-R1 Three-digit level Classification Codes

Chapter 3

Estimation of Distributed Weights for Cross-referencing Commodity Classifications :

Towards the Formulation of SITC-R1 Three-digit level Classification Codes

NODA Yosuke

The purpose of this chapter is to use OECD trade statistics data for Japan export to formulate long-term time series data employing SITC-R1 three-digit level commodity classification codes converted utilizing distributed weights calculated on the basis of cross-referencing with the SITC series. Conversion of trade statistics for Japan and Korea by means of distributed weights has already been carried out using neural networks as a calculation method, as discussed by Noda in *Conversion of Trade Statistics with Reversion to Commodity Classification: Case Study of Japan and Korea* and Shirosaka in *Estimation of Distributed Weight to Conversion for 3digits Level of SITC: Application of Neural Network Method* in No. 83 of the IDE Statistical Papers Series.

The neural network method will presumably be effective in the future, but at present it has certain drawbacks, such as the fact that its high cost makes it inconvenient to use when needed. For the purposes of this chapter we have therefore utilized the least squares method with constraints to calculate weights.

This chapter takes up conversion that goes in the direction from A to B. These correspondences and the distribution structure of the distributed weight are presented in organized, where A_i represents the individual classification code from A and x_i represents the statistical value related to the individual classification code A_i for i=1···m. Similarly, B_j represents the individual classification code from B and y_i represents the statistical value related to the individual classification code B_j. for j=1···n. If k pieces of data can be obtained each from A_i and B_j as observed values, then the statistical value X from A and the statistical value Y from B will be represented as:

$$Y = \begin{pmatrix} y_1' \\ \vdots \\ y_n' \end{pmatrix} = \begin{pmatrix} y_{11} & \cdots & y_{1k} \\ & & \\ y_{n1} & \cdots & y_{nk} \end{pmatrix}$$

and

$$X = \begin{pmatrix} x_1' \\ \vdots \\ x_m' \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ & & \\ x_{m1} & \cdots & x_{mk} \end{pmatrix}$$

The weight for distribution in a direction going from A_i to B_j is defined as:

$$\omega_{ij} \quad i = 1 \cdots m, j = 1 \cdots n$$

The distribution method involves taking the statistical values x_i corresponding to the individual classification code A and splitting it into n pieces of y_j. The sum of these for y_j is shown in the following formula. This can be expressed as:

$$y_{j1} = x_{11}\omega_{1j} + \cdots + x_{m1}\omega_{mj} + u_{j1}$$

Classification B is made up of n elements, $j = 1 \cdots n$. Showing these as a matrix, they are:

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{n1} \end{pmatrix} = \begin{pmatrix} \omega_{11} & \cdots & \omega_{m1} \\ \vdots & & \vdots \\ \omega_{1n} & \cdots & \omega_{mn} \end{pmatrix} \begin{pmatrix} x_{11} \\ \vdots \\ x_{m1} \end{pmatrix} + \begin{pmatrix} u_{11} \\ \vdots \\ u_{n1} \end{pmatrix}$$

Taking the weight distribution matrix as W and then saying $\omega_i' = (\omega_{1i} \cdots \omega_{mi})$, we have:

$$W = \begin{pmatrix} \omega_{11} & \cdots & \omega_{m1} \\ \vdots & & \vdots \\ \omega_{1n} & \cdots & \omega_{mn} \end{pmatrix} = \begin{pmatrix} \omega_1' \\ \vdots \\ \omega_n' \end{pmatrix}$$

The relationship between A, B, and the weight W can be expressed as:

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

where a matrix U is similarly defined, taking k pieces of observed values. If the m-dimensioned vector composed of elements that are all 1 is defined as l_m, then, from the conditions for weight distribution in formula (1), they satisfy the identity:

(2) $$W' l_n = l_m$$

Formula (1) and (2) are forms of expression used to model by using neural networks.

To express the foumula of (1) and (2) in an ordinary linear regression model, the formulas are transposed as, $Y' = X'W' + U'$ and arranged to

$$y_i = X'\omega_i + u_i \quad i = 1 \cdots n .$$

Expressing them in a matrix form makes

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

and let y be a vector $y' = (y_1' \cdots y_n')$, X^* be

$$X^* = \begin{pmatrix} X' & & \\ & \ddots & \\ & & X' \end{pmatrix}$$

ω be a vector $\omega = (\omega_1' \cdots \omega_m')'$, u be $u' = (u_1' \cdots u_n')$, we have

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

Besides, let I_m be a m-th identify matrix, matrxix C be $C = (I_m \cdots I_m)$. As a vector ω is a weight,

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

is satisfied。

In general commodity groups have some zero weights within the elements of cross referencing among classification code. The assumption for previous model of (3) and (4) is made without zero weigths and we need to change he model with zero weights. To remove elements of zero weights from the distributed weight ω_i, let D_i be a matrix removed from j-th line from I_m. An adjusted vector ω_i^D which does not include zero elements of weights is obtained by the linear transformation as $\omega_i^D = D_i \omega_i, \quad i = 1 \cdots m$. From both $(\omega^D)' = ((\omega_1^D)' \cdots (\omega_m^D)')$ and

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

we have $\omega^D = D\omega$.

Dealing with same process as ω^D, X^* and C are needed to transform $X^{*D} = X^* D'$ and $C^D = CD'$. The adjusted ω^D, X^{*D} and C^D for the relation of cross-refercing gives

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

for the previous formula of (3), and

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

for the previous formula of (4).

To incoporate a information of (3') and (4'), the method of restricted least squares is proposed. We seek the ω^D that minimize the sum of squared residuals $(u^D)' u^D$ of (3') subject to the restriction $C^D \omega^D = l_m$ of (4'). Therefore we minimize s with respect to ω and λ,

(5) $$s = (u^D)' u^D + \lambda' (C^D \omega^D - l_m)$$

where λ is a vector of Lagrange multipliers. Setting the derivative of s with respect to ω and λ equal to 0 gives for the minimizing value $\tilde{\omega}^D$.

$$\partial s / \partial \omega^D = -2(X^{*D})'(y - X^{*D}\omega^D) + (C^D)'\lambda$$

If $(X^{*D})' X^{*D}$ is a nonsingular matrix then $\tilde{\omega}^D$ is the unique solution of $\partial s / \partial \omega^D = 0$.

(6) $$\tilde{\omega}^D = \hat{\omega}^D - [(X^{*D})' X^{*D}]^{-1} (C^D)' \lambda$$

where $\hat{\omega}^D$ is the unrestricted least squares estimator,

(7) $$\hat{\omega}^D = [(X^{*D})' X^{*D}]^{-1} (X^{*D})' y$$

and satisfy the condition,

(8) $$C^D \tilde{\omega}^D - l_m = 0$$

which is the restricted condition obtained from $\partial s / \partial \lambda = 0$. Premultiplying both sides of (6) by C^D gives,

$$C^D \tilde{\omega}^D = C^D \hat{\omega}^D - C^D [(X^{*D})' X^{*D}]^{-1} (C^D)' \lambda$$
$$= l_m$$

whence

(9) $$\lambda = \{C^D [(X^{*D})' X^{*D}]^{-1} (C^D)'\}^{-1} \bullet (C^D \hat{\omega}^D - l_m)$$

and let M be

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

Inserting (9) and (10)back into (6) gives

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

The restricted least squares estimator (11) is clearly unbiased.

Because weights were calculated on the basis of cross-referencing among classification codes for each commodity group, when the groups were large, the fact that a greater number of weights were generated than the amount of data obtainable from trade statistics made calculation impossible. In this chapter the bootstrap method is used to ensure a sufficient amount of data: the data obtained from trade statistics is used to calculate a sampling distribution, from which the sample is randomly drawn.

The trade statistics data used in this chapter come from the IDE world trade database system, the AID-XT (Ajiken Indicators on Developing Economies: Extended for Trade Statistics); the data for Japan are import and export trade statistics formulated by the OECD, and cover the period from 1962 to 2000. Trade data obtained from international organizations as UN and OECD is subjected to a sum check of commodity classifications and partner nations for each reporting nation and importing and exporting sector to test consistency. Data from nations which do not maintain consistency of commodity classifications are corrected to achieve the highest level of consistency possible in the sum check. Commodities which do not have a lower level classification code among their commodity classification codes are termed sub-category codes; when the transaction value of the sub-category codes is totaled the collection of classification codes agreeing with the total value of commodities are termed "consistent sub-category codes". The IDE uses trade statistics data based on commodity classifications employing consistent sub-category codes as the base data for the AID-XT.

Because the SITC-R1 original series of Japan was used from 1962 to 1977, it was employed when formulating SITC-R1 data. The SITC-R1 series was not used from 1978 to 1987; SITC-R2 was employed as the original series, and SITC-R1 categories were therefore formulated by calculating distributed weights with a direction from SITC-R2 to SITC-R1. Because the original series from 1988 to 1993 was SITC-R3, calculation has similarly been performed from SITC-R3 to SITC-R2 and the SITC-R2 data obtained reconverted to SITC-R1. It is the purpose of this chapter to perform these two conversions using distributed weights.

Since 1994 the HS system has been used for commodity classifications, and the method described above was therefore applied after conversion to SITC-R3 for years from 1994 onwards. HS-O (HS 1988 version)commodity classifications have been converted to SITC-R3 using an HS-O – SITC-R3 correspondence table. The fact that the cross-referencing is basically of type 1 or type 3 enables combined conversion to be performed.

The time-series data for the SITC-R1 series in the 1962–1999 range can be obtained by putting together these converted series. SITC-R2 and SITC-R3 series can also be derived by conversion in a similar manner.