

Statistical Note on a Problem Arising in the Economic Data Compilation

by

Kazuhiko MATSUNO

Abstract

When an econometric model to be estimated is specified in terms of a stock variable and necessary data corresponding to the stock variable is not available, we usually compile the stock data by accumulating available flow data and then carry the estimation using the compiled stock data. We examine the statistical properties of this method and propose a new method.

1. Motivation .

In estimating economic relationships, we are not free from the paucity of economic statistics. Some kind of data compilation from various sources is inevitable to prepare the data required for the estimation.

For instance, when a model to be estimated is specified in terms of stock variable and necessary data corresponding to the stock variable is not available, we usually compile stock data by accumulating available flow data and then carry the estimation using the compiled stock data.

In econometric analyses, we commonly presume that there is no harm in data collection or compilation prior to the estimation, and then we proceed to apply sophisticated statistical methods to the somewhat arbitrary data.

A model for household saving behavior demands a statistical solution for the optimal procedure of data compilation combined with estimation problem, since the existing data is not exactly the type that the model requires for the estimation. In this article we try to clarify the nature of the underlying problem of data compilation in relation to the tentative method used in this analysis of saving behavior. A statistical solution is given within a simplified framework.

2. Example

The following specific problem arose in the analysis of a model of household saving behavior. This is worth mentioning for its generality in the sense that similar problems appear in other fields of econometrics.

The model specifies that the amount of saving (stock variable F_t) at the end of the year t is a function of income (exogenous explanatory variable I_t). The function involves unknown parameters (α 's) to be estimated. The estimation requires time series data of the stock variable F_t and the explanatory variable I_t . Yearly data is appropriate. But in fact the survey for collecting the data of F_t and I_t is not conducted every year. Therefore the sample size of the data is not large enough to estimate the parameters. In other similar cases, we find the situation where the size of the available data is less than required.

Actually another independent survey is conducted every year for collecting data concerning household saving and income. This survey, however, provides statistics of saving in terms of increase or decrease of saving (flow variable S_t) and income as well. For simplicity, we assume in the following that the stock data survey is conducted every odd years so that the stock data of even year does not exist.

One intuitive possibility of estimating the parameters is the following: Theoretically the definitional equation,

$$S_t = F_t - F_{t-1},$$

holds. By this equation the original stock function (the function for the stock variable) can be transformed into the flow function (the function for the flow variable), where the same explanatory variable and the parameters appear. Then the flow function can be used for estimating the parameters based on the data of the flow variable S_t and income with sufficient sample size.

As another possibility, we compile stock data for F_t of even years t by adding the flow data S_t to the available stock data F_{t-1} of the preceding year $t-1$. Denoting the stock data compiled in this way by \hat{F}_t we have

$$\hat{F}_t = F_{t-1} + S_t.$$

The directly available stock data of odd years and the compiled stock data of even years constitute a set of sample with sufficient sample size for estimating the parameters based on the original stock function.

Similar sorts of problems arising elsewhere in econometrics have been discussed from a point of view of economic theory. We will present a statistical understanding of the problem.

3. Economic relationship and observational relationship

An event generates from an economic relationship, which is supposed to be an object of our econometric analysis. We have many different ways of observing the event and of recording the observations. Therefore many different kinds of economic statistics result concerning the event.

One aspect of the event at time t , denoted by E_t , is expressed quantitatively by a variable ξ_t . We build a model to explain the event or underlying economic relationship by specifying that the theoretical variable ξ_t is a function of other variables ζ_t 's,

$$(3.1) \quad \xi_t = f(\zeta_t, \alpha),$$

where unknown parameters α 's are included and to be estimated.

For the purpose of estimating the parameters α , we conduct a series of observations on empirical counterparts of the theoretical variables, ξ_t and ζ_t 's, and record the observations of ξ_t and ζ_t as x_t and z_t respectively. We call the series of observations the survey g_1 , and (x_t, z_t) is said to be economic statistics by the survey g_1 .

Another aspect of the event E_t is expressed by a variable η_t . The theoretical variable η_t can be the object of another series of observations. Economic theory may specify that the two variables ξ_t and η_t are related with each other by the equation

$$(3.2) \quad \eta_t = h(\xi_t, \beta),$$

where other parameters β 's may be included. Accordingly the relationship (3.1) is transformed into

$$(3.3) \quad \eta_t = h(f(\zeta_t, \alpha), \beta) \equiv F(\zeta_t, \alpha, \beta).$$

Another series of observations (the survey denoted by g_2) is conducted for the theoretical variables η_t and ζ_t . The survey g_2 provides the economic statistics y_t and z_t' , observational counterparts of η_t and ζ_t respectively.

As for the example of Section 2, the survey g_1 is not conducted every year but every odd year for observing the stock variable F_t , and the survey g_2 is conducted every year for the flow variable S_t .

Formally, we write the observational relationships as

$$(3.4) \quad \begin{aligned} [x_t, z_t] &= g_1([\xi_t, \zeta_t]), & t &= 1, 3, 5, \dots, \\ [y_t, z_t'] &= g_2([\eta_t, \zeta_t]), & t &= 1, 2, 3, \dots \end{aligned}$$

The economic statistics x_t, z_t by the survey g_1 are assumed to be subject to observational errors, and similarly the statistics y_t, z_t' are assumed to accompany observational errors. This requires us to estimate parameters characterizing the distribution of the errors. In the present discussion we confine ourselves to the case where z_t and z_t' do not involve observational errors and they are identical to each other and to the theoretical variable ζ_t .

4. Regression model

We assume a linear regression model for making the discussion explicit. That is, the event E_t at time t is expressed by the linear relationship between the theoretical variables ξ_t and ζ_t 's.

$$(4.1) \quad \xi_t = \sum_{k=1}^K \alpha_k \zeta_{kt}.$$

For instance, the stock of household saving ξ_t is a linear function of the household income at time t , ζ_{1t} , and other exogenous variables ζ_{kt} .

Let η_t be the flow of saving, then by definition

$$(4.2) \quad \eta_t = \xi_t - \xi_{t-1}.$$

The relationship (4.1) is written in terms of the flow variable as

$$(4.3) \quad \eta_t = \sum_{k=1}^K \alpha_k (\zeta_{kt} - \zeta_{kt-1}).$$

The survey g_1 is conducted every odd year for observing ξ_t and ζ_{kt} , and the observations of ξ_t and ζ_{kt} are recorded as x_t and z_{kt} respectively. The observation x_t is assumed to be subject to an additive observational error u_t ,

$$(4.4) \quad x_t = \xi_t + u_t, \quad t = 1, 3, 5, \dots, T-1,$$

where T is an even number. Then the regression model follows

$$(4.5) \quad x_t = \sum_{k=1}^K \alpha_k z_{kt} + u_t.$$

The other survey g_2 is conducted every year for observing η_t and ζ_{kt} , and the observations of η_t and ζ_{kt} are recorded as y_t and z_{kt}' in accordance with the equation (4.3). The statistic y_t involves an additive observational error v_t , i.e.,

$$(4.6) \quad y_t = \eta_t + v_t.$$

Hence the regression model follows in terms of different variables,

$$(4.7) \quad y_t = \sum \alpha_k (z_{kt} - z_{kt-1}) + v_t, \quad t = 1, 2, 3, \dots, T,$$

where the variable z_{kt} appears on the right side since we assume no observational error in the explanatory variables.

An observational counterpart of the definition (4.2) is given as

$$(4.8) \quad y_t = x_t - x_{t-1}.$$

Using this equation, the stock data of even year is compiled according to the equation

$$(4.9) \quad \hat{x}_t = x_{t-1} + y_t, \quad t = 2, 4, 6, \dots.$$

The regression model for the variable \hat{x}_t is

$$(4.10) \quad \hat{x}_t = \sum \alpha_k z_{kt} + (u_{t-1} + v_t).$$

The discussion above yields several regression models for estimating the parameters α , in accordance with the choice of data to be used for the estimation.

Model A; using the directly available stock data x_I and the data Z_I of the explanatory variables by the survey g_1 , where

$$(4.11) \quad \begin{aligned} x_I' &= [x_1 x_3 x_5 \dots x_{T-1}], \\ Z_I &= \begin{pmatrix} z_1' \\ z_3' \\ \vdots \\ z_{T-1}' \end{pmatrix}, \\ z_i' &= [z_{1i} z_{2i} \dots z_{Ki}]. \end{aligned}$$

The regression equation for Model A is

$$(4.12) \quad x_I = Z_I \alpha + u_I,$$

where

$$(4.13) \quad u_I' = [u_1 u_3 u_5 \cdots u_{T-1}].$$

Model B; using the flow data $y' = [y_I', y_{II}']$ and the observations for $z_{kt} - z_{kt-1}$ provided by the survey g_2 , where

$$(4.14) \quad \begin{aligned} y_I' &= [y_1 y_3 y_5 \cdots y_{T-1}], \\ y_{II}' &= [y_2 y_4 y_6 \cdots y_T]. \end{aligned}$$

The regression equation for Model B is (4.7), which is rewritten as

$$(4.15) \quad \begin{bmatrix} y_I \\ y_{II} \end{bmatrix} = \begin{bmatrix} Z_I - Z_0 \\ Z_{II} - Z_I \end{bmatrix} \alpha + \begin{bmatrix} v_I \\ v_{II} \end{bmatrix},$$

where

$$(4.16) \quad \begin{aligned} Z_0 &= \begin{bmatrix} z_0' \\ z_2' \\ \vdots \\ z_{T-2}' \end{bmatrix}, & Z_{II} &= \begin{bmatrix} z_2' \\ z_4' \\ \vdots \\ z_T' \end{bmatrix}, \\ v_I' &= [v_1 v_3 v_5 \cdots v_{T-1}], \\ v_{II}' &= [v_2 v_4 v_6 \cdots v_T]. \end{aligned}$$

Model C; using the stock data x_I and g_1 and the compiled stock data \hat{x}_{II} and the observations of the explanatory variable z_{kt} , where

$$(4.17) \quad \hat{x}_{II}' = [\hat{x}_2 \hat{x}_4 \hat{x}_6 \cdots \hat{x}_T].$$

The regression equation for Model C is a set of (4.5) and (4.10), in matrix form,

$$(4.18) \quad \begin{bmatrix} x_I \\ \hat{x}_{II} \end{bmatrix} = \begin{bmatrix} Z_I \\ Z_{II} \end{bmatrix} \alpha + \begin{bmatrix} u_I \\ u_I + v_{II} \end{bmatrix}.$$

Model D is possible; using overall data directly available, the stock data x_I by g_1 and the flow data y_I, y_{II} by g_2 , but without the data compilation. The regression equation is a combination of the stock function (4.5) and the flow function (74.),

$$(4.19) \quad \begin{bmatrix} x_I \\ y_I \\ y_{II} \end{bmatrix} = \begin{bmatrix} Z_I \\ Z_I - Z_0 \\ Z_{II} - Z_I \end{bmatrix} \alpha + \begin{bmatrix} u_I \\ v_I \\ v_{II} \end{bmatrix}.$$

Model E is formally set; using x_I, \hat{x}_{II} and y_I . The regression equation is

$$(4.20) \quad \begin{bmatrix} x_I \\ \hat{x}_{II} \\ y_I \end{bmatrix} = \begin{bmatrix} Z_I \\ Z_{II} \\ Z_I - Z_0 \end{bmatrix} \alpha + \begin{bmatrix} u_I \\ u_I + v_{II} \\ v_I \end{bmatrix}.$$

It is noted (i) that from a standpoint of Model D, Model A ignores the available flow data y_I, y_{II} , and (ii) that similarly Model B ignores the stock data x_I . Model E is a nonsingular transformation of Model D so that they are observationally equivalent. It is seen (iii) that Model C, compared with Model E or with equivalent Model D, ignores the flow data y_I .

5. Least squares solution

The observational errors are, by assumption, distributed as

$$(5.1) \quad \begin{bmatrix} u_I \\ v_I \\ v_{II} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 I & 0 \\ 0 & \omega^2 I \end{bmatrix} \right),$$

and are independent of z_{kit} . We do not consider systematic observational error which might be represented by non-zero mean. The variances σ^2 and ω^2 are measures of accuracy of the surveys g_1 and g_2 respectively.

The variance-covariance matrices of the error term of Model A through Model E are

$$(5.2) \quad \begin{aligned} \Sigma_A &= \sigma^2 I, \\ \Sigma_B &= \omega^2 I, \\ \Sigma_C &= \begin{bmatrix} \sigma^2 I & \sigma^2 I \\ \sigma^2 I & (\sigma^2 + \omega^2) I \end{bmatrix}, \\ \Sigma_D &= \begin{bmatrix} \sigma^2 I & 0 \\ 0 & \omega^2 I \end{bmatrix}, \\ \Sigma_E &= \begin{bmatrix} \sigma^2 I & \sigma^2 I & 0 \\ \sigma^2 I & (\sigma^2 + \omega^2) I & 0 \\ 0 & 0 & \omega^2 I \end{bmatrix}. \end{aligned}$$

We consider a linear model in general,

$$(5.3) \quad \begin{bmatrix} x \\ y \end{bmatrix} \sim N \left(\begin{bmatrix} X \\ Y \end{bmatrix} \theta, \begin{bmatrix} \Sigma & 0 \\ 0 & \Omega \end{bmatrix} \right),$$

where Σ and Ω are known. The generalized least squares estimator for the parameter θ using data x and y is given by

$$(5.4) \quad \hat{\theta}(x, y) = (A + B)^{-1} (a + b),$$

where

$$(5.5) \quad \begin{aligned} A &= X' \Sigma^{-1} X, \\ B &= Y' \Omega^{-1} Y, \\ a &= X' \Sigma^{-1} x, \\ b &= Y' \Omega^{-1} y. \end{aligned}$$

The variance-covariance matrix of $\hat{\theta}(x, y)$ is

$$(5.6) \quad \Phi_{\hat{\theta}(x, y)} = (A + B)^{-1}$$

The generalized least squares estimator using only y is

$$(5.7) \quad \hat{\theta}(y) = B^{-1} b,$$

the variance-covariance matrix of which is

$$(5.8) \quad \Phi_{\hat{\theta}(y)} = B^{-1}.$$

The difference $\Phi_{\hat{\theta}(y)} - \Phi_{\hat{\theta}(x, y)}$ is at least positive semidefinite, and therefore $\hat{\theta}(x, y)$ is more efficient relative to $\hat{\theta}(y)$ in this sense.

We apply the above discussion to the models of Section 4. Given σ^2 and ω^2 , the generalized least squares estimator of α , using whole data in accordance with Model D, is

$$(5.9) \quad \hat{\alpha}(x_I, y) = \left(\frac{1}{\sigma^2} Z_I' Z_I + \frac{1}{\omega^2} W' W \right)^{-1} \left(\frac{1}{\sigma^2} Z_I' x_I + \frac{1}{\omega^2} W' y \right),$$

where $W = \begin{bmatrix} Z_I - Z_0 \\ Z_{II} - Z_I \end{bmatrix}$. The generalized least squares for Model A using the stock data x_I only is

$$(5.10) \quad \hat{\alpha}(x_I) = (Z_I' Z_I)^{-1} (Z_I' x_I),$$

and for Model B using the flow data y is

$$(5.11) \quad \hat{\alpha}(y) = (W' W)^{-1} (W' y).$$

We conclude from the above discussion that $\hat{\alpha}(x_I, y)$ is more efficient relative to $\hat{\alpha}(x_I)$ and $\hat{\alpha}(y)$.

We consider in general a linear model,

$$(5.12) \quad x \sim N(X\theta, \Sigma),$$

and a model obtained by a nonsingular matrix Ψ transformation

$$(5.13) \quad \Psi x \sim N(\Psi X\theta, \Psi \Sigma \Psi').$$

The least squares estimators of (5.12) and (5.13) are identical.

Model E is obtained from Model D by multiplying the nonsingular matrix of appropriate order,

$$(5.14) \quad \Psi = \begin{bmatrix} I & 0 & 0 \\ I & 0 & I \\ 0 & I & 0 \end{bmatrix}.$$

Therefore we see that $\hat{\alpha}(x_I, y_I, y_{II})$ based on Model D is more efficient relative to $\hat{\alpha}(x_I, x_{II})$ based on Model C.

Thus it is suggested that we use Model D, that is, we estimate the parameters using the directly available stock and flow data without compiling data. The combination of sets of data provided by the survey g_1 and the survey g_2 results in the gain of efficiency. It should, of course, be noted that the result is valid if the two sets of data are generated from the identical economic relationship.

In the case where σ^2 and ω^2 are unknown, the feasible estimator for Model D is

$$(5.15) \quad \hat{\alpha}(x_I, y) = \left(\frac{1}{\hat{\sigma}^2} Z_I' Z_I + \frac{1}{\hat{\omega}^2} W' W \right)^{-1} \left(\frac{1}{\hat{\sigma}^2} Z_I' x_I + \frac{1}{\hat{\omega}^2} W' y \right),$$

where $\hat{\sigma}^2$ and $\hat{\omega}^2$ are consistent estimators of σ^2 and ω^2 ,

$$(5.16) \quad \begin{aligned} \hat{\sigma}^2 &= \frac{1}{T/2} (x_I - Z_I \hat{\alpha} (x_I))' (x_I - Z_I \hat{\alpha} (x_I)) , \\ \hat{\omega}^2 &= \frac{1}{T} (y - W \hat{\alpha} (y))' (y - W \hat{\alpha} (y)) . \end{aligned}$$

6. Data compilation by economic theory

The variance of the compiled stock data is

$$(6.1) \quad \text{var} (\hat{x}_t) = \sigma^2 + \omega^2 .$$

If we iterate the method one step further like

$$(6.2) \quad \hat{x}_{t+1} = \hat{x}_t + y_{t+1} = \sum \alpha_k z_{kt-1} + u_{t-1} + v_t + v_{t-1} ,$$

then

$$(6.3) \quad \text{var} (\hat{x}_{t+1}) = \sigma^2 + 2\omega^2 .$$

For the s -th step iteration, we have

$$(6.4) \quad \text{var} (\hat{x}_{t+s}) = \sigma^2 + (s+1) \omega^2 ,$$

which shows that the variance accumulates as s increases. This method of compilation is based on the definitional equation (4.9)

We can consider another method of compilation based on the economic relationship (4.5). This is a prediction of the type

$$(5.5) \quad \tilde{x}_{t+s} = a' z_{t+s} ,$$

where a denotes an estimate vector for α , and $\hat{\alpha}(x_t, y)$ can substitute for it. The variance of \tilde{x}_{t+s} does not accumulate as s increases. The prediction error is due to the estimation error for α and is independent of s . It is, however, shown in the previous discussion that there is no need to compile the stock data for the estimation.

7. Likelihood solution

The problem is not restricted to linear models. A perspective for generalizing the analysis to non-linear models is given.

The economic relationship is described in terms of the theoretical variable ξ_t as

$$(7.1) \quad \xi_t = f(\zeta_t, \alpha) .$$

The relationship is also written in terms of $\eta_t = h(\xi_t, \beta)$ as

$$(7.2) \quad \eta_t = h(f(\zeta_t, \alpha), \beta) = F(\zeta_t, \alpha, \beta)$$

The survey g_1 is conducted at time s to give statistic x_s for ξ_s in the way formally described as

$$(7.3) \quad x_s = g_1(\xi_s, u_s) = g_1(f(\zeta_s, \alpha), u_s) \equiv G_1(\zeta_s, u_s, \alpha) ,$$

where u_s is an observational error distributed independently as

$$(7.4) \quad u_s \sim \phi(u_s, \gamma),$$

γ being the parameter. The survey g_2 is conducted at time s' to provide statistic $y_{s'}$ for $\eta_{s'}$ in the way described as

$$(7.5) \quad \begin{aligned} y_{s'} &= g_2(\eta_{s'}, v_{s'}) = g_2(F(\xi_{s'}, \alpha, \beta), v_{s'}) \\ &\equiv G_2(\xi_{s'}, v_{s'}, \alpha, \beta), \end{aligned}$$

where $v_{s'}$ is an observational error independently subject to the distribution, with parameter δ ,

$$(7.6) \quad v_{s'} \sim \psi(v_{s'}, \delta).$$

We assume that the surveys g_1 and g_2 record statistic z for ζ without errors.

With assumptions commonly employed, we can derive the likelihood function from the setting above. Thus the maximum likelihood estimates of the parameters α , β , γ and δ are obtained.

For the linear example of Section 3, the likelihood function is

$$(7.7) \quad L_\infty(\sigma^2)^{-T/4} (\omega^2)^{-T/2} \exp -\frac{1}{2} Q,$$

where

$$(7.8) \quad Q = \left(\begin{bmatrix} x_I \\ y \end{bmatrix} - \begin{bmatrix} Z_I \\ W \end{bmatrix} \alpha \right)' \begin{bmatrix} \sigma^2 I & 0 \\ 0 & \omega^2 I \end{bmatrix}^{-1} \left(\begin{bmatrix} x_I \\ y \end{bmatrix} - \begin{bmatrix} Z_I \\ W \end{bmatrix} \alpha \right).$$

And the corresponding likelihood equations are

$$(7.9) \quad \begin{aligned} \sigma^2 &= \frac{2}{T} (x_I - Z_I \alpha)' (x_I - Z_I \alpha), \\ \omega^2 &= \frac{1}{T} (y - W \alpha)' (y - W \alpha), \\ \left(\frac{1}{\sigma^2} Z_I' Z_I + \frac{1}{\omega^2} W' W \right) \alpha &= \left(\frac{1}{\sigma^2} Z_I' x_I + \frac{1}{\omega^2} W' y \right). \end{aligned}$$

8. Conclusion

Our analysis does not settle the problem completely. For we set simplifying assumptions in my respects. Among these the no-error assumption for the explanatory variables is the most important. Generalizing this point will lead us to the complications appearing in the discussion on the errors-in-variable model. Besides these reservations, it should be noted that the our discussion supplies a hypothesis under which we can combine various economic statistics, and the validity of the hypothesis is a matter of empirical verification.