A list of Working Papers on the last pages

No. 62, 1982

Missing Variables and Two-stage Least-squares Estimation from More than One Data Set

by

N. Anders Klevmarken*

April, 1982

* N. Anders Klevmarken is professor at the Department of Statistics, University of Gothenburg, Viktoriagatan 13, S-411 25 Sweden. This work was done while he was visiting the Survey Research Center, ISR, University of Michigan.

The author has benefited from comments received at seminars at Oueen's University, University of Michigan, University of Western Ontario and Michigan State University. Several helpful suggestions, in particular from S. Augustyniak, P. Howrey, J. Kmenta and W. Rodgers are gratefully acknowledged.

July 1981

Abstract

In a situation when no single sample includes all the endogenous variables of a simultaneous equation model but there are two (or more) non-overlapping samples and each variable is included in at least one, then it is possible to pool the data and estimate the model consistently by a two-stage least-squares procedure. The assymptotic variances of the estimates are not always larger than those which would have been obtained with TSLS from one complete sample. It is also shown that under certain assumptions the same approach can be applied to an ordinary regression model.

Key words: Missing data, Pooling data, Statistical matching, TSLS estimation

1. INTRODUCTION AND BACKGROUND

Survey data is a frequently used input in social science research and their importance is increasing. This has been true for a long time in sociology, for instance, but also in other disciplines there is a shift in research interests. In economics there is an increased emphasis on micro economics using survey data and panel studies as compared to macro economic problems analysed with aggregate time-series data. Large surveys are, however, very expensive and the increased response burden and the awareness of the privacy issue on the part of legislators and the general public makes it increasingly difficult to get the co-operation of the households and business establishments. We will thus frequently have to rely on already existing data files designed for different purposes. Most likely we then find that no data set contains all the information we would need. Sometimes it is possible to combine information from several data sets by exact matching, but this is not possible when they do not overlapp, i.e. when the probability for an individual to participate in more than one survey is very small, or when identifying information like the social security number is not available or its use is prevented by protection of privacy.

We might thus have to face a situation when it is impossible to obtain a single sample including all the variables we would need to estimate a model or test a hypothesis, but it might be possible to obtain two or more data sets, each of which would not include all relevant variables, but each variable would be included in at least one data set. Can this type of data be used at all and if so how?

One suggestion to deal with this situation is to use synthetic or statistical matching. If two or more data sets have some variables in common, but do not include the same individuals, the common variables could be used to match "alike" individuals. In this way data for two (or more) individuals, one from each data set, is merged to a new set of synthetic individuals. Ideally the new data set would have the same distributional properties as a proper survey, but doubts have been raised about the possibilities to obtain this without unrealisticly strong assumptions about the universe. A survey of the literature on statistical matching and an extensive list of references are given in a report from US Department of Commerce (1980).

Although the advocates of statistical matching usually emphasize other uses of a synthetic file than estimation of multivariate models or tests of hypothesis about human behavior with control for confounding factors, this is from the social scientists point of view a likely reason to attempt a statistical match. The theoretical basis for the statistical matching techniques is, however, relatively weak and the approach suggested in this paper is not a statistical match, but the results obtained below invite to a few comparative remarks about statistical matching at the end of the paper.

The problem treated in this paper is the estimation of one of the relations in a simultaneous equation model when the relevant variables have to be obtained from different data sets which have no individuals in common. The solution is a two-stage least-squares proceedure which does not require matching. A more rigorous specification of the problem and the model is first given in section 2. Then follow the estimation method, an analysis of its properties and a discussion of the consequences of alternative assumptions about the model and the data configuration. One special case is the linear regression model.

2. THE PROBLEM

The problem is to estimate the following equation,

$$y = Y_1\beta + X_1\delta + u; \tag{1}$$

which is part of the interdependent system,

$$YB' + X\Gamma' = U; \tag{2a}$$

$$E(U) = O; \; E(U'U) = n\Sigma \tag{2b, c}$$

where $Y_{n \cdot G}$ is a matrix of n observations on G endogenous variables,

$y_{n \cdot 1}$ is a vector of the n observations on the endogenous variable explained by (1),

$Y_{1,n\cdot g}$ is a matrix of the n observations on the g explanatory endogenous variables in (1),

$X_{n\cdot K}$ is the observational matrix of all K exogenous variables,

$X_{1,n\cdot k}$ is a submatrix of X which includes the k exogenous variables in (1),

$U_{n\cdot G}$ is a matrix of stochastic disturbances,

$u_{n\cdot 1}$ the vector of stochastic disturbances of (1), one of the columns of U.

$B_{G\cdot G}, \Gamma_{K\cdot K}$ are parameter matrices,

$\beta_{G\cdot 1}, \gamma_{k\cdot 1}$ are vectors of the non-zero parameters in (1)

$\Sigma_{G\cdot G}$ is an unknown positive definite moment matrix.

It is assumed that (1) is identified.

The reduced form of the complete system is,

$$Y = X\pi' + V; \tag{3a}$$

$$\text{where } \pi = -B^{-1}\Gamma; \tag{3b}$$

$$\text{and } V = U(B')^{-1}. \tag{3c}$$

The part of the reduced form corresponding to the endogenous variables to the right of the equality in (1) is,

$$Y_1 = X\pi_1' + V_1; \tag{4}$$

where π_1 and V_1 are the corresponding $g\cdot K$ and $n\cdot g$ submatrices of π and V respectively. For later use it is also convenient to introduce a $n\cdot(K-k)$ matrix X_2 defined by,

$$X = \{X_1 \mid X_2\}. \tag{5}$$

Suppose now that data are not available in the form of one complete sample but there are two samples, A and B, none of which contains all variables. Assume that data come in the following form,

$$\text{Sample A: } y^A_{(n_A\cdot 1)};\ X^A_{(n_A\cdot K)}$$

$$\text{Sample B: } Y^B_{1,(n_B\cdot g)};\ X^B_{(n_B\cdot K)}$$

n_A and n_B are the two sample sizes. They are not necessarily equal. Since (2c) implies that there is no residual correlation between observational units, the two samples can be treated as independent random samples.[1]

An example for which this problem specification might be applicable is the joint estimation of demand functions for consumer goods and household time-use functions, both derived from a household production type of model. Consumer expenditure data could be obtained from a household budget study, while time-use data would have to be taken from a separate time-use survey. There are presently no surveys which include both kinds of data. Both surveys would, however, give income data and other characteristics of the household.

3. ESTIMATION

3.1 The estimation proceedure.

Eq. (1) cannot be estimated from sample A alone since the Y_1-variables are missing, but the two samples can be combined in the following two-stage proceedure,

I. Estimate the reduced form equations (4) from sample B by OLS which gives the estimates $\hat{\pi}_1^B$. Use these estimates to predict Y_1 in sample A, i.e.

$$\hat{Y}_1^A = X^A \hat{\pi}_1'^B; \tag{6}$$

II. Estimate by OLS from sample A

$$y^A = \hat{Y}_1^A \beta + X_1^A \gamma + (u^A + \tilde{V}_1^A \beta); \tag{7}$$

$$\text{where } \tilde{V}_1^A = Y_1^A - \hat{Y}_1^A.$$

Note that $\tilde{V}_1^A$ is not the vector of least-squares prediction errors from sample A and thus not necessarily orthogonal to X^A.

With the following notation,

$$\delta' = \{\beta' \mid \gamma'\}_{1\cdot(g+k)};$$

$$Z = \{\hat{Y}_1^A \mid X_1^A\}_{n_A \cdot (g+k)};$$

then (7) becomes

$$y^A = Z\delta + (u^A + \tilde{V}_1^A \beta); \tag{8}$$

and the estimator of δ is,

$$\hat{\delta} = (Z'Z)^{-1} Z' y^A \tag{9}$$

If the two samples would coincide, $\hat{\delta}$ would be the usual TSLS estimator.

3.2 Properties

3.2.1 Small sample bias

The expected value of $\hat{\delta}$ over the whole sample space defined by both samples can be obtained in the following stepwise way:

$$E(\hat{\delta}\mid X^A,X^B) = E\{E(\hat{\delta}\mid \hat{\pi}_1{}^B,X^A)\mid X^A,X^B\}. \tag{10}$$

$$\begin{aligned} E(\hat{\delta}\mid \hat{\pi}_1{}^B,X^A) &= E\{(Z'Z)^{-1}Z'y^A\} = \\ & E\{\delta+(Z'Z)^{-1}Z'(u^A+\tilde{V}_1{}^A\beta)\} = \\ \delta+E\{(Z'Z)^{-1}Z'\tilde{V}_1{}^A\beta\} &= \delta+E\{(Z'Z)^{-1}Z'(Y_1{}^A-\hat{Y}_1{}^A)\beta\} = \\ & \delta+(Z'Z)^{-1}Z'X(\pi_1{}'-\hat{\pi}_1{}'^B)\beta. \end{aligned} \tag{11}$$

$$\therefore\ E(\hat{\delta}\mid X^A,X^B) = \delta+E\{(Z'Z)^{-1}Z'X(\pi_1{}'-\hat{\pi}_1{}'^B)\beta\mid X^A,X^B\}. \tag{12}$$

The last term of (12) is in general not zero and the estimator is thus biased, a property it shares with the usual TSLS estimator. The following simple example might clarify this point further. All variables and parameters are scalars.

Structural form:

$$y_1 = \beta_{12}y_2+\gamma_{10} \qquad +u_1$$

$$y_2 = \beta_{21}y_1+\gamma_{20}+\gamma_{21}x+u_2$$

Reduced form:

$$y_1 = \pi_{10}+\pi_{11}x+v_1$$

$$y_2 = \pi_{20}+\pi_{21}x+v_2$$

$$\text{where } \pi_{11} = \beta_{12}\gamma_{21}/(1-\beta_{12}\beta_{21})$$

$$\text{and } \pi_{21} = \gamma_{21}/(1-\beta_{12}\beta_{21})$$

The first equation of the structural form is estimated with the following two samples,

Sample A: $y_1{}^A, x^A$.

Sample B: $y_2{}^B, x^B$.

The first step of the estimation proceedure gives,

$$\hat{y}_2{}^A = \hat{\pi}_{20}{}^B+\hat{\pi}_{21}{}^B x^A;$$

and the estimator of β_{12} then becomes,

$$\hat{\beta}_{12} = \frac{\Sigma(\hat{y}_2{}^A - \bar{\hat{y}}_2{}^A)y_1{}^A}{\Sigma(\hat{y}_2{}^A - \bar{\hat{y}}_2{}^A)^2} = \frac{\Sigma\hat{\pi}_{21}{}^B(x^A - \bar{x}^A)y_1{}^A}{\Sigma(\hat{\pi}_{21}{}^B)^2(x^A - \bar{x}^A)^2} = \frac{\hat{\pi}_{11}{}^A}{\hat{\pi}_{21}{}^B}.$$

The first equation is exactly identified which explains why $\hat{\beta}_{12}$ in this case is a simple ratio of the estimates of two reduced form parameters.

We now find that,

$$E(\hat{\beta}_{12} \mid x^A, x^B) = E_B\{E_A(\hat{\pi}_{11}{}^A \mid \hat{\pi}_{21}{}^B)\} = E_B(\pi_{11} \mid \hat{\pi}_{21}{}^B) \neq \pi_{11}/\pi_{21} = \beta_{12}.$$

3.2.2 Consistency

We will first look at the case when n_A is finite and fixed while n_B tends towards infinity. Assume that the matrix $\{\frac{1}{n_B}(x^B)'X^B\}$ tends towards a finite non-singular matrix when n_B tends towards infinity. It then follows that

$$\underset{n_B \to \infty}{\text{plim}}\ \hat{\pi}_1{}^B = \pi_1; \tag{13a}$$

$$\hat{Y}_1{}^A \to E(Y_1{}^A) \text{ when } n_B \to \infty; \tag{13b}$$

$$\tilde{V}_1{}^A \to V_1{}^A \text{ when } n_B \to \infty; \tag{13c}$$

$$Z \to \{X^A \pi_1{}' \mid X_1{}^A\} \text{ when } n_B \to \infty. \tag{13d}$$

Set $Z_0 = \{X^A \pi_1{}' \mid X_1{}^A\}$, then

$$\underset{n_B \to \infty}{\text{plim}}\ \hat{\delta} = \delta + (Z_0{}'Z_0)^{-1}Z_0{}'(u^A + V_1{}^A\beta) \tag{14}$$

The expected value of this limit for the sample space defined by sample A is,

$$E_A(\underset{n_B \to \infty}{\text{plim}}\ \hat{\delta}) = \delta. \tag{15}$$

Thus, if sample B is "very large" the estimation proceedure is almost equivalent to replacing $Y_1{}^A$ by its expected value and estimating the following relation by OLS.[2]

$$y_1{}^A = E(Y_1{}^A)\beta + X_1{}^A\gamma + (u^A + V_1{}^A\beta) \tag{16}$$

For very large n_B the estimates of β and γ are thus almost unbiased.

Now, let both n_A and n_B tend to infinity. Assume that $\{\frac{1}{n_A}(X^A)'(X^A)\}$ and $\{\frac{1}{n_B}(X^B)'(X^B)\}$ both tend to finite non-singular matrices as n_A and n_B respectively tend to infinity. Then,

$$\operatorname*{plim}_{\substack{n_A\to\infty\\ n_B\to\infty}} \hat{\delta} = \operatorname*{plim}_{n_A\to\infty} (\operatorname*{plim}_{n_B\to\infty} \hat{\delta}) =$$

$$\delta + (\operatorname*{plim}_{n_A\to\infty} \frac{1}{n_A}(Z_0'Z_0))^{-1}(\operatorname*{plim}_{n_A\to\infty} (\frac{1}{n_A} Z_0'u^A) + \operatorname*{plim}_{n_A\to\infty} (\frac{1}{n_A} Z_0'V_1{}^A\beta)) = \delta. \tag{17}$$

The second equality follows from (14) and the third equality from the by definition zero correlation between the stochastic residuals and the exogenous variables. $\hat{\delta}$ is thus a consistent estimator.

3.2.3 Asymptotic distribution

Assume that $n_B = kn_A$, where $k>0$ is an arbitrary finite constant, and that $(1/n_A)(X^{A\prime}X^A)$ and $(1/n_B)(X^{B\prime}X^B)$ both tend to finite non-singular limits when n_A and n_B tend to infinity. Assume also that the rows of U are not only uncorrelated but also independent.

Since $\hat{\pi}_1^B$ is a consistent estimator it follows that $(1/n_A)(Z'Z)$ tends in probability to a finite non-singular matrix, say Q. It also follows that $\tilde{V}_1^A$ tends in distribution to V_1, the submatrix of reduced form errors. Thus,

$$\sqrt{n_A}(\hat{\delta}-\delta) = (n_A^{-1}Z'Z)^{-1}(1/\sqrt{n_A})Z'(u^A+\tilde{V}_1^A\beta) \tag{18}$$

tends in distribution to $Q^{-1}(1/\sqrt{n_A})Z_0'(u^A+V_1\beta)$. It will be proved below that $(u^A+V_1\beta)$ has a scalar moment matrix, say $\sigma^2 I$. It then follows from the Lindeberg-Levy theorem that $(1/\sqrt{n_A})Z_0'(u^A+V_1\beta)$ is asymptoticly normal with zero mean vector and covariance matrix $\sigma^2 Q$. (For a proof se Theil (1971) p. 380). The asymptotic distribution of $Q^{-1}(1/\sqrt{n_A})Z_0'(u^A+V_1\beta)$ is thus normal with zero mean vector but with the covariance matrix $\sigma^2 Q^{-1}$.

To prove that the covariance matrix of $(u^A+V_1\beta)$ is scalar, assume without loss of generality that (1) is the first structural equation of the system (2) and that the endogenous variables of that equation are the first g+1 variables of Y. If we partition the inverse of the parameter matrix B in the following way,

$$(B')^{-1}_{G\times G} = (B^{*}_{G\times g} \mid B^{**}_{G\times(G-g)}); \tag{19}$$

it follows that,

$$V=(V_{\underset{n\times g}{1}} \mid V_{\underset{n\times(G-g)}{2}}) = U(B')^{-1} = U(B^{*} \mid B^{**}); \tag{20}$$

and thus,

$$V_1=UB^{*}. \tag{21}$$

The covariance matrix of $(u^A+V_1\beta)$ now becomes,

$$E(u^A+V_1\beta)(u^A+V_1\beta)' = E(u^A u^{A\prime}) + E(u^A\beta'B^{*\prime}U') + E(UB^{*}\beta u^{A\prime})+ E(UB^{*}\beta\beta'B^{*\prime}U'). \tag{22}$$

From (2c) it follows that $E(u^A u^{A\prime}) = \sigma_{11}I$, where σ_{11} is the top left element of the moment matrix Σ. In order to evaluate the last three terms of (22) partition $\Sigma$by its columns,

$$\Sigma=(\sigma_1 \mid \sigma_2 \mid \ldots \mid \sigma_G) \tag{23}$$

We then obtain,

$$E(u^A\beta'B^{*\prime}U') = (I_{n\times n} \otimes \beta'B^{*\prime})E(u^A \otimes U') = (I \otimes \beta'B^{*\prime})(I_{n\times n} \otimes \sigma_1)$$

$$= \beta'B^{*\prime}\ \sigma_1 I; \tag{24}$$

and,

$$E(UB^{*}\beta u^{A\prime}) = \sigma_1' B^{*}\beta I; \tag{25}$$

and,

$$E(UB^*\beta\beta'B^{*\prime}U') = E\{(UB^*\beta)\otimes(\beta'B^{*\prime}U')\} = E\{(I_{n\times n} \otimes \beta'B^{*\prime})(U\otimes U')(B^*\beta\otimes I)\}$$

$$=(I\otimes\beta'B^{*\prime})(I\otimes\sigma_1 \mid I\otimes\sigma_2 \mid \ldots\ldots \mid I\otimes\sigma_G)(B^*\beta\otimes I)$$

$$=(\beta'B^{*\prime}\sigma_1 I \mid \beta'B^{*\prime}\sigma_2 I \mid \ldots\ldots \mid \beta'B^{*\prime}\sigma_G I)(B^*\beta\otimes I)$$

$$=\{(\beta'B^{*\prime}\sigma_1 \mid \beta'B^{*\prime}\sigma_2 \mid \ldots\ldots \mid \beta'B^{*\prime}\sigma_G)\otimes I\}\{B^*\beta\otimes I\}$$

$$=(\beta'B^*\Sigma\otimes I)(B^*\beta\otimes I) = \beta'B^{*\prime}\Sigma B^*\beta I. \tag{26}$$

(22), (24)-(26) now give,

$$E(u^A+V_1\beta)(u^A+V_1\beta)' = (\sigma_{11}+2\sigma_1'B^*\beta+\beta'B^{*\prime}\Sigma B^*\beta)I. \tag{27}$$

The expression within brackets is thus the scalar σ^2 refered to above.

To conclude, if $n_B = kn_A$, then $\sqrt{n_A}(\hat{\delta}-\delta)$ asymptoticly follows a normal distribution with zero mean vector and covariance matrix $(\sigma_{11}+2\sigma_1'B^*\Sigma B^*\beta)Q^{-1}$.

In this case the variance of the two-stage least-squares estimates based on two samples thus differs from the variance of the ordinary TSLS estimator based on a complete sample A by the second and third terms inside the parenthesis of (27). Since these terms do not only depend on all variances and covariances but also on all the elements of B, the relative magnitude of the assymptotic variance of $\hat{\delta}$ is difficult to evaluate without knowing at least the structure of B and the signs of the non-zero parameters. One might believe that the two incomplete samples would be less informative than one complete sample. but this is not necessarily true because a large sample B might compensate for the missing variables in sample A. Also assymptoticly the variance of $\hat{\delta}$ can be exceeded by the variance of the TSLS estimator based only on sample A which, for instance, can be shown with the two-equation model used in the example above. In this model

$$B = \begin{Bmatrix} 1 & -\beta_{12} \\ -\beta_{21} & 1 \end{Bmatrix} ;$$

and thus,

$$\text{Asy. var}(\hat{\delta}) = n_A^{-1}\{\sigma_{11}+2(\sigma_{11}\beta_{12}^2/(1-\beta_{12}\beta_{21})+\sigma_{12}\ \beta_{12}(1-\beta_{12}\beta_{21}))+$$
$$(\beta_{12}/(1-\beta_{12}\beta_{21}))^2(\sigma_{11}\beta_{12}^2+2\sigma_{12}\beta_{12}+\sigma_{22})\}\times \text{plim}(n_A^{-1}Z'Z).^{-1}$$

With, for instance, $\beta_{12}=0.05$, $\beta_{21}=1$, $\sigma_{11}=\sigma_{22}=1$ and $\sigma_{12}=-0.9$ the scalar expression within braces is less than σ_{11}, but if the sign of σ_{12} is reversed it exceeds σ_{11}.[3]
The fact that the two-sample estimator may have a smaller variance than the ordinary TSLS estimator might at first be a surprise. The explanation is that since the last term of (18), $Z'V_1\beta$, does not vanish, unlike the corresponding term of the ordinary TSLS estimator, its limit in distribution may be negatively correlated with the first term, $Z'u^A$, and if this correlation is sufficiently strong the variance of the total error will become less than the variance of the first error term.

4. ALTERNATIVE ASSUMPTIONS ABOUT DATA CONFIGURATION AND MODEL

Note that <u>all</u> exogenous variables are included in both samples. In the ordinary case with only one sample, consistent estimates of the parameters of (1) can be obtained by the instrumental variables method if there are at least g X_2-variables included in the sample to serve as instruments. However, for the case discussed in this paper it is not possible to obtain consistent estimates with one or more of the exogenous variables missing from either sample. If we would attempt to estimate a reduced form with some of the X_2-variables missing or replaced by other variables the estimates of π_1 would in general be biased and inconsistent, (13c) would no longer hold and $\hat{\delta}$ would not be a consistent estimator. To see this note that $\underset{n_A\to\infty}{\text{plim}}\,(n_A^{-1}Z_0'V_1^A\beta)$ in (17) has to be replaced by $\underset{\substack{n_A\to\infty\\ n_B\to\infty}}{\text{plim}}\,(n_A^{-1}Z'\tilde{V}_1^A\beta)$ and that the critical part of this expression is,

$$\underset{\substack{n_A\to\infty\\ n_B\to\infty}}{\text{plim}}\,(n_A^{-1}(X_1^A)'\tilde{V}_1^A\beta) = \{\text{plim}(n_A^{-1}(X_1^A)'Y_1^A) - \text{plim}(n_A^{-1}(X_1^A)'\hat{Y}_1^A)\}\beta =$$

$$\underset{n_A\to\infty}{\text{plim}}\,(n_A^{-1}(X_1^A)'X)\{\pi_1' - \underset{n_B\to\infty}{\text{plim}}\,(\hat{\pi}_1^B)'\}\beta. \tag{29}$$

Thus, when $\hat{\pi}_1^B$ is not a consistent estimator (29) does not vanish and $\hat{\delta}$ becomes inconsistent.

One may also note that even if sample A would include data on all endogenous variables in (1) but there would be less than g X_2-variables included in the sample, the information in sample B cannot be utilized to obtain consistent estimates.

If sample A would include all the endogenous variables of (1) but not all X_1-variables, could we then use the information in sample B to estimate X_1? It is not obvious that such a proceedure can be justified within the present model. The problem is that there is no theoretical basis for predicting X_1 since these variables are exogenous. However, if it, for instance, would be realistic to add to the model the assumption that all exogenous variables are multivariate normal then one could proceed to use both samples to estimate the model.[4]

A special case of (1), with β=0, is the common model,

$$y = X_1\gamma + u;\ \ E(u|X_1) = 0;\ \ E(uu'|X_1) = \sigma_{11}I. \tag{30}$$

Assume that,

$$\{X_1 \mid X_2\} \sim N(\{\mu_1 \mid \mu_2\};\ \Omega). \tag{31}$$

Since the regression surfaces in a multivariate normal distribution are linear, we can write,

$$X_1 = X_2R + \varepsilon;\ \ E(\varepsilon|X_2) = 0; \tag{32}$$

where R is a matrix function of μ_1, μ_2 and Ω. (32) inserted into (30) gives,

$$y = X_2R\gamma + (u + \varepsilon\gamma). \tag{33}$$

R can be estimated from sample B and provided $K\text{-}k \geq k$, $X_2^A\hat{R}^B$ gives k linearly independent predictions of X_1^A which inserted into (33) give,

$$y^A = X_2^A\hat{R}^B\gamma + (u + \varepsilon\gamma + X_2^AD\gamma); \tag{34}$$

where $D=R-\hat{R}$. (34) is an errors-in-variable model and the OLS estimates of γ will have a small sample bias. They will, however, be consistent and assymptotically unbiased since $D\to 0$ when $n_B\to\infty$.

If the assumption of no explanatory endogenous variables, $\beta=0$, is relaxed again, the two-stage least-squares proceedure taking into account both the simultaneity of the model and the need to estimate X_1 would require that $K\text{-}k \geq g+k$.

5. A BRIEF COMPARISON WITH STATISTICAL MATCHING

The two-stage least-squares proceedure described above can be compared with statistical matching. Suppose we want to estimate (1) using the two samples A and B as given on page 3. If statistical matching is defined as a random drawing of a vector of Y_1-values, say $Y_{1_i}^*$, among those observations of sample B with a given vector $\{X_{1_i}^A \mid X_{2_i}^A\}$, to replace the unknown Y_{1_i}-vector in sample A, then the equation to estimate becomes,

$$y^A = Y_1^*\beta + X_1^A\gamma + (G\beta + u^A); \tag{35}$$

where $G = Y_1 - Y_1^*$. It is assumed that n_B is so much larger than n_A that a match can always be found.

Since Y_1^* comes from sample B and u^A from sample A and there is by assumption no correlation between the residuals of the two samples, Y_1^* is uncorrelated with U^A. Statistical matching thus takes care of the simultaneity problem, but at the same time it introduces an errors in variables problem because the matching error G is now part of the residual. OLS estimates of (35) will thus neither be unbiased nor consistent, but this problem can be overcome if (35) is estimated by a method which takes the matching error into account, for instance, and instumental variables method. TSLS applied to (35) would be assymptotically equivalent to the two-stage proceedure suggested above, but in small samples it might be less efficient since some of the information in the larger B-sample is ignored.[5]

If it is not always possible to find a match with identically the same vector $\{X_{1_i} \mid X_{2_i}\}$ but instead a match is defined by some distance function on the exogenous variables, a systematic error is introduced which presumably makes also the TSLS estimates of (35) inconsistent. Matching constrained to a one-to-one correspondence between the X-values of the two samples is almost equivalent to simulating the reduced form (4), but matching of observations with only approximately the same X-values can be compared to a simulation based on the wrong vector $\{X_{1_i} \mid X_{2_i}\}$.

The same results seem to carry over to the regression model with multivariate normal X-variables. The replacement of the unobserved X_1^A by a match from sample B, say X_1^*, will introduce a matching or measurement error. Estimation of γ would then require a method which takes these errors into account.

The simple form of statistical matching assumed here does not do justice to the variety of techniques used in practice, but one general conclusion is that the estimation method used after a statistical match should take into account the random - and if possible any non-random - matching error. Ordinary least-squares will in general not do this.

6. CONCLUDING REMARKS

Future research based on micro data might have to rely more and more on the kind of incomplete data discussed in this paper. To be able to do this we will need some vehicle, a model, which links the variables of the different data sets. In our case the simultaneous-equation model and the multivariate normal distribution both served this purpose. Such a theoretical basis is necessary for solving the missing data problem whether this is done by the two-stage least-squares proceedure or statistical matching and it seems unlikely that a model free and purely design-based proceedure could be developed. If this is true the kind of general purpose argument, sometimes given for statistical matching, would have little validity for these two approaches. It also raised the issue of how robust these methods are for model specification errors.

In the first part of the paper no particular family of distributions was assumed which in a natural way leads to least-squares theory. In section 4 a multivariate normal distribution was introduced. The particular family is not principally important - although the normal distribution is very convenient - but if it is realistic to assume a distribution there might be more efficient methods based on maximum likelihood theory. With large micro data samples efficiency might, however, be of secondary importance.

Footnotes

1 Note that this model specification is within the econometric "superpopulation" tradition and there is no mentioning of a sample design. Although this is a controversal issue, it can be argued that the estimation proceedure will not depend on the sampling design as long as the selection probabilities are independent of the residuals U.

2 Note the similarity with Wold's generalized interdependent systems, GEID, (Mosbaek & Wold, 1970).

3 For $\sigma_{12} = -0.9$ asy. var $(\hat{\delta}) \simeq 0.9131\, n_A^{-1} \text{plim}(n_A^{-1} Z'Z)^{-1}$. For $\sigma_{12} = +0.9$ asy. var $(\hat{\delta}) \simeq 1.1030\, n_A^{-1}\, \text{plim}(n_A^{-1} Z'Z)^{-1}$.

4 This assumption does not imply any causal relation between the exogenous variables.

5 In practice the smaller sample A would probably be matched into sample B, which implies that each y_i-value might be used repeatedly and more of sample B would be used. There is, however, no guarantee that the whole of sample B can be used since there may be vectors $\{X_{1_i}{}^B \mid X_{2_i}{}^B\}$ which have no correspondence in sample A.

REFERENCES

Mosbaek, E.J. and Wold, H.O. (ed.), (1970), Interdependent systems: Structure and Estimation, Amersterdam, North-Holland.

U.S. Department of Commerce (1980), Report on Exact and Statistical Matching Techniques, Statistical Policy Working Paper 5, Office of Federal Statistical Policy and Standards, U.S. Government Printing Office.

Theil, H., (1971), Principles of Econometrics, John Wiley & Sons, New york