International Microsimulation Association

RESEARCH ARTICLE

Pooling Incomplete Data Sets

Anders Klevmarken[1]*

[1]University of Gothenburg, Gothenburg, Sweden

Abstract Data needed in micro studies are not always available from just a single source. In such cases it might be possible to combine two or more independent samples. The problem studied in this paper is to estimate a linear function between *y* and *x* from one sample of y-observations and another of x-observations. This is feasible if there are common variables *z* which can be used to predict *x*. A two-stage least squares estimator is propounded, which, for the model considered, is also an ML estimator. Simulation experiments show that it has a good relative efficiency and virtually no small sample bias.

DOI: https://doi.org/10.34196/ijm.00248

1. Introduction

Microdata, i.e. data on individuals, households or firms, have long been used in social science research, and their importance is increasing. Since human behavior sometimes calls for very complex explanations and since social experimentation is rarely a feasible approach, social scientists frequently work with complex models involving many variables in efforts to control for confounding effects. To estimate these models we do not only need large samples of microdata, but ideally we would also need to observe many aspects of behavior for *each* individual, household or firm.

Surveys are the primary source of microdata, but survey research is very expensive, and few researchers can afford new surveys. Contributing to the high costs are the increasing difficulties in many countries to gain the cooperation of the respondents. In particular, when many questions are asked and there is a heavy respondent burden, the respondents tend to economize their time. They have also become aware of the privacy issues. The public debate about the use of computers and the risks for invasion of privacy, which in some countries have resulted in data legislation, have made it even more difficult to make the respondents cooperate.

As pointed out in ***Dalenius (1982)***, matrix sampling is a class of sampling schemes which may cope with these problems. Since with such a design only a sample of all variables is observed for each selected unit, the respondent burden and the risk for invasion of privacy are reduced. It is, however, obvious that multivariate analysis from such a sample might meet with difficulties.

High costs and nonresponse problems also make us look for alternative data sources, i.e. already existing data files and administrative records. Frequently, however, we are unable to find all the variables needed in those files and records. If feasible, matching of two or more data sets might give us what we need. Matching is also much less expensive than a new survey. However, in most cases exact matching is not possible, either because there is no overlap, i.e. no individual, household or firm can be found in more than one data set, or there is no unique identification term, or the use of this term is prohibited in order to protect personal privacy. In Sweden, for instance, matching data on individuals falls under the Data Act and requires a permit from the National Data Inspection Board.[1]

If neither a new complete data collection nor an exact matching of existing files are feasible solutions, to what extent is it then possible to use non-overlapping datasets? The answer to this question will in the general case depend on the intended analysis and in what sense the data are incomplete. This paper treats the problem of estimating a linear relation from two independent data sets, none

***For correspondence:**
anders@klevmarken.nu

1. For a discussion of the need for microdata in economics, see ***Klevmarken (1983a)***.

of which includes all the relevant variables. Suppose, for instance, that we want to estimate a model according to which *y* depends on *x*. *y* and *x* are, however, not observed for the same units, but *y*- and x-observations are found in two independent samples.

This is a hopeless situation unless something more is known about *y* and *x*, and there is some common information in both samples. Suppose a third variable, *z*, which could be used as a predictor of *x*, is observed in both samples. It is then possible to estimate the predictive relationship between *x* and *z* from one sample and use the estimates in the other sample to predict *x*. The relation between *y* and *x* is finally estimated on the basis of these predictions and observed y-values. This procedure requires model assumptions which make it possible to pool two (or more) samples. The properties of the estimates, of course, depend on the assumptions made, and the whole approach is useful only if these are not too constraining. As we shall see, it is possible to work within quite a general class of linear models.

When invasion of privacy is a real issue or nonresponse a severe problem, a complete data collection might not be permissible or feasible. However, if good predictors can be found, each individual, household or firm would only need to contribute partial information, which would be less sensitive, and an analysis would still be feasible with the methods suggested in this paper.

Missing variables is not a new problem. In most textbooks of regression analysis and econometrics it is discussed as a specification error. In the applied literature various *ad hoc* approaches can be found. For instance, in the economic literature on compensating wage differentials, aggregate data on occupational characteristics such as accident rates and aggregate work environment data are matched with individuals on the basis of their occupation and industry. One example is given in ***Brown (1980)***. As we shall find, this is a special application of the prediction approach discussed below.

An alternative approach is statistical matching (***U.S. Department of Commerce, 1980***). The theoretical basis for statistical matching is still undeveloped, and it is also a very expensive approach (***Barr et al. (1982)***, ***Paass (1982)***, ***Rodgers and DeVol, 1982***). The approach discussed in this paper has the advantages of being based on conventional statistical theory and of requiring no expensive matching of microdata.

In Section 2 of the paper we review some results for a model with stochastically dependent equations. Since interdependent systems have mostly been used in connection with aggregate time-series, while microdata applications are less frequent, it might erroneously be believed that the approach suggested in this paper is only of minor interest in microstudies. However, the importance of this type of model is likely to increase in microstudies as well, when more longitudinal data become available and when ample data make it feasible to analyze joint decisions of microunits, e.g. joint household decisions about work, leisure, consumption and savings. Also, and more important, the estimation approach suggested is not only applicable to interdependent systems but to a much wider class of models. In Section 3 this approach will be used to estimate a linear regression model. One reason to discuss its application to an interdependent system first is that the "predictors" of the unobserved variables are in a natural way given by theory, which is not obviously the case with a regression model.

Section 4 presents some findings from a simulation study of the small sample properties of the estimator, and Section 5 gives some concluding remarks.

2. Estimation of an interdependent linear model

The estimation of one equation in an interdependent system of equations from two independent samples with missing variables was discussed in ***Klevmarken (1982)***. This equation was specified as

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

and it was part of the interdependent system,

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

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

where *Yn·G* is a matrix of *n* observations on G endogenous variables,
$y_{n.1}$ is a vector of the *n* observations on the endogenous variable explained by (1),
$Y_{1,n.g}$ is a matrix of the *n* observations on the g explanatory endogenous variables in (1),
$X_{n.K}$ is a matrix which includes all K exogenous variables, ,

$X_{1,n.k}$ is a submatrix of X which includes the k exogenous variables in (1),
$U_{n.G}$ is a matrix of stochastic disturbances,
$u_{n.1}$ is the vector of stochastic disturbances of (1), one of the columns of *U*,
$B_{G.G}$ and $\Gamma_{G.K}$ are parameter matrices,
$\beta_{g.1}$ and $\gamma_{k.1}$ are vectors of the nonzero parameters in (1),
$\Sigma_{G.G}$ 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 in equation (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 an $n\cdot(K\text{-}k)$ matrix X_2 defined by,

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

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

$$\text{Sample A}: y^A_{n_{A.1}}; X^A_{n_{A.K}}$$

$$\text{Sample B}: Y^B_{1,(n_{B.K})}; X^B_{n_{B.K}}$$

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.

An example to 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 expenditure study, while time-use data would have to be taken from a separate time-use survey. There are at present practically no surveys which include both kinds of data. Both kinds of surveys would, however, give income data and other characteristics of the household.

Eq. (1) cannot be estimated from sample *A* alone, since the Y_1-variables are missing, but, if $g \leq K\text{-}k$, the two samples can be combined in the following two-stage procedure:

I. Estimate the reduced form equations (4) from sample *B* by OLS, which gives the estimates $\widehat{\pi_1^B}$. 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*

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

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

$$\partial' = \{\beta' | \gamma'\}_{1.(g+k)}$$

$$Z = \left\{ \hat{Y}_1^A | X_1^A \right\}_{n_A.(g+k)};$$

(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, δ would be the usual TSLS estimator. In ***Klevmarken (1982)*** it was shown that the estimator δ is biased but consistent. The following asymptotic properties can also be proved.[2]

If $n_B = cn_A$, where c>0 is an arbitrary finite constant, and if $(1/n_A)$ $(X^{A\prime}X^A)$ 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, and if the rows of the error matrix U are stochastically independent, then $\sqrt{n_A}\,(\hat{\delta} - \delta)$ asymptotically follows a normal distribution with zero mean vector and co-variance matrix.

$$(\sigma_{11} + 2\sigma_1' B^* \beta + 2\beta' B^{*\prime} \Sigma B^* \beta) Q^{-1}$$

$\sigma_1$is the first column of the matrix Σ. The first element of this vector is σ_{11} and the other elements are the covariances between the error term in the first equation and those in the other equations. B^* is a matrix of g columns from $(B')^{-1}$ such that

$$V_1 = UB^*.$$

Q, a finite non-singular matrix, is the limit to which $(1/n_A)(Z'Z)$ tends in probability when n_A tends to infinity.

The asymptotic moment matrix of the ordinary TSLS estimator based on a complete A- sample is $\sigma_{11} Q^{-1}$. As shown in ***Klevmarken (1982)***, it is possible to find cases for which the asymptotic variance of δ is smaller than the asymptotic variance of the ordinary TSLS estimator.

As a preliminary to the next section, suppose that sample A does not only lack the Y_1-observations but also all observations on one or more of the X_1-variables. Could we then use the information in sample B to predict X_1 in A? Since X_1 by definition is exogenous, there is no theoretical justification for predicting X_1 within the present model. Additional assumptions about X_1 are needed. This problem and the nature of the new assumptions are, however, not particular to a model of interdependent equations, and could more conveniently be discussed within the framework of an ordinary regression model.

3. Estimation of a linear regression model

As above, we assume that there are two independently drawn samples, A and B. These now include the following variables and observations,

$$\text{Sample A}: y^A_{(n_A.1)} \;\; X^A_{2(n_A.k_2)}$$

$$\text{Sample B}: X^B_{1(n_B.k_1)} \; X^B_{2(n_B.k_2)}.$$

The two samples will be used to estimate the model,

$$y = X_1\gamma + u;\; u \sim N(0, \sigma_u^2 I); \tag{3.1a,b}$$

2. This corrects results given in ***Klevmarken (1982)***. A proof is parallel to that given in ***Klevmarken (1983b)***.

where γ is an unknown parameter vector, and X_1 is assumed to be a matrix of k_1 stochastic variables, which depends linearly on X_2.

$$X_1 = X_2 R + \epsilon; \tag{3.2a}$$

$$\epsilon = \{\epsilon_1, ..., \epsilon_{k_1}\} \tag{3.2b}$$

$$\epsilon_j \sim N(0, \sigma_j^2 I); \tag{3.2c}$$

$$E(\epsilon_i \epsilon_j) = 0 \; \forall \; i \neq j; \tag{3.2d}$$

R is a parameter matrix. X_2 is treated as a matrix of non-stochastic variables. It is also assumed that there are at least as many predictors as predictands, i.e. $k_2 \geq k_1$. For the two samples we thus obtain the following three relations,

$$y^A = X_1^A \gamma + u^A \tag{3.3a}$$

$$X_1^A = X_2^A R + \epsilon^A \tag{3.3b}$$

$$X_1^B = X_2^B R + \epsilon^B \tag{3.3c}$$

Since X_1^A is not observed, (3.3 b) is inserted into (3.3 a) to give the following reduced system of observed variables

$$y^A = X_2^A R\gamma + \epsilon^A \gamma + u^A \tag{3.4a}$$

$$X_1^B = X_2^B R + \epsilon^B; \tag{3.4b}$$

This model shows great similarities with an errors-in-variables model analysed in ***Goldberger (1972)***. Goldberger's starting point was a regression model containing a single explanatory variable observed with a random error. This model also assumed that the true unobserved variable was a non-stochastic linear function of a number of independent variables. ***Zellner (1970)*** has previously considered this model and developed a generalized least-squares estimator and also presented a Bayesian analysis of the model. Goldberger developed the corresponding maximum likelihood theory and extended it to a model in which the true unobserved variable is a stochastic function of the independent variables rather than an exact function. The observable equivalent of Goldberger's model has the same form as eq:s (3.4 a, b). The only difference is that the two equations refer to two different samples, while Goldberger's observations were assumed to come from a single sample.

It is convenient to rewrite eq:s (3.4 a, b) in the following way,

$$y^A = X_2^A \pi + \nu, \tag{3.5a}$$

$$x_{1j}^B = X_2^B r_j + \epsilon_j^B; \; j = 1, ..., k_1, \tag{3.5b}$$

where X_{1j}^B is the j:th column of X_1^B , r_j is the j:th column of R, and

$$\pi = R\gamma; \tag{3.5c}$$

$$v = \epsilon\gamma + u; \tag{3.5d}$$

It is assumed that u and ϵ_j are uncorrelated for all j. It follows that,

$$E(VV') = (\gamma' \Sigma \gamma + \sigma_u^2) I; \tag{3.6a}$$

where

$$\Sigma = diag\ \left\{\sigma_1^2, ..., \sigma_{k_1}^2\right\}; \tag{3.6b}$$

Since the two samples are independent, the joint distribution of the observable variables is,

$$F(y^A,\ X_1^B) = F_y(y^A)F_{X1}(X_1^B); \tag{3.7}$$

The likelihood function becomes (disregarding irrelevant constants),

$$L(\gamma,\ R,\ \sigma_u^2,\ \Sigma)\ =\ -n_A log(\gamma'\Sigma\gamma + \sigma_u^2) - (\gamma'\Sigma\gamma + \sigma_u^2)^{-1}(y - X_2^A Ry)'(y - X_2^A Ry) - n_B\Sigma_j log(\sigma_j^2) - \Sigma_j(\sigma_j^2)^{-1}(X_{1j}^B - X_2^B r_j)'(X_{1j}^B - X_2^B r_j); \tag{3.8}$$

Differentiating with respect to the unknown parameters and putting the derivatives equal to zero gives,

$$\frac{\delta L}{\partial\sigma_u^2} = -n_A(\gamma'\Sigma\gamma + \sigma_u^2)^{-1} + (\gamma'\Sigma\gamma + \sigma_u^2)^{-2}(y - X_2^A Ry)'(y - X_2^A Ry) = 0; \tag{3.9}$$

If we define $\xi = (\gamma^{'}\Sigma\gamma + \sigma_u^2)$ it thus follows that,

$$\hat{\xi} = n_A^{-1}(y - X_2^A Ry)'(y - X_2^A Ry); \tag{3.10}$$

$$\frac{\delta L}{\partial\sigma_j^2} = -n_A(\gamma'\Sigma\gamma + \sigma_u^2)^{-1}\gamma_j^2 + \gamma_j^2(\gamma'\Sigma\gamma + \sigma_u^2)^{-2}(y - X_2^A R\gamma)'(y - X_2^A R\gamma) - n_B(\sigma_j^2)^{-1} + (\sigma_j^2)^{-2}(x_{1j}^B - X_2^B r_j)'(x_{1j}^B - X_2^B r_j) = 0; \tag{3.11}$$

From eq:s (3.9) and (3.11) it follows that,

$$\widehat{\sigma}_j^2 = n_B^{-1}(X_{1j}^B - X_2^B r_j)'(X_{1j}^B - X_2^B r_j); \tag{3.12}$$

The concentrated likelihood function then becomes,

$$L^* = -n_A log\left\{(y^A - X_2^A R\gamma)'(y^A - X_2^A R\gamma)\right\} - n_B\Sigma_j log\left\{(x_{1j}^B - X_2^B r_j)'(x_{1j}^B - X_2^B r_j)\right\}; \tag{3.13}$$

$$\frac{\partial L*}{\partial\gamma} = -n_A\left\{(y^A - X_2^A R\gamma)'(y^A - X_2^A R\gamma)\right\}^{-1} R'X_2^{A\prime}(y^A - X_2^A R\gamma) = 0 \tag{3.14}$$

$$\because \hat{\gamma} = (R'X_2^{A\prime}X_2^A R)^{-1}R'X_2^{A\prime}y^A; \tag{3.15}$$

If eq. (3.15) is premultiplied by $R^{'}X_2^{A\prime}X_2^A R$ one easily sees that,

$$\hat{\pi} = R\hat{\gamma} = (X_2^{A\prime}X_2^A)^{-1}X_2^{A\prime}y^A; \tag{3.16}$$

Eq. (3.16) inserted into eq. (3.13) gives a new concentrated likelihood function,

$$L^{**} = -n_B\Sigma_j log\left\{(X_{1j}^B - X_2^B r_j)'(X_{1j}^B - X_2^B r_j)\right\}; \tag{3.17}$$

If the derivative of this function is put equal to zero, we obtain the ordinary least-squares solution,

$$\hat{r_j} = (X_2^{B\prime}X_2^B)^{-1}X_2^{B\prime}x_{1j}; \tag{3.18a}$$

or in matrix form,

$$\hat{R} = (X_2^{B\prime}X_2^B)^{-1}X_2^{B\prime}X_1^B; \tag{3.18b}$$

From (3.15) and (3.18b) we thus find that the maximum-likelihood solution is equivalent to the TSLS procedure analogous to the estimation method suggested for the interdependent model in the previous section.

This result no longer holds if the X1-variables are correlated, i.e. if Σ is not diagonal, or if the errors are heteroscedastic. With such changes in the model specification we would have to find the maximum likelihood estimates by numerical maximization of the likelihood function.

If we allow both for contemporaneously correlated errors and for heteroscedasticity, the model can be written in the following way,

$$y^A = X_2^A R\gamma + \epsilon_\gamma^A + u^A; \tag{3.19}$$

$$X^B = (I \otimes X_2^B)r + \epsilon^B; \tag{3.20}$$

where $X^{B\prime} = \{x_{11}{}', x_{12}{}', ..., x_{1_{k1}}{}'\}$,

$$r' = \{r_1{}', r_2{}', ..., r_{k_l}{}'\},$$

$$\epsilon^{B\prime} = \left\{\epsilon_1^{'}, \epsilon_2^{'}, ..., \epsilon_{k_l}^{'}\right\},$$

$$\in_j{}' = \{\in_{j1}, \in_{j2}, ..., \in_{jn_B}\}$$

Now let

$$\mathrm{E}(\in_{is}\in_{jt}) = \left\{\begin{matrix} \sigma_{ij(t)} \; if \; s = t \\ 0 \; if \; s \neq t \end{matrix}\right\} \tag{3.21}$$

We define the contemporaneous moment matrices for observation t,

$$E(\in_{1t}, \in_{2t}, ... \in_{k_1 t})(\in_{1t}, \in_{2t}, ... \in_{k_1 t})' = \Sigma(t) = \left\{\sigma_{ij(t)}\right\}_{k_1.k_1}; \tag{3.22}$$

and the individual specific moment matrices,

$$E(\in_i \in_j') = \Omega_{ij} = diag\left\{\sigma_{ij(1)}, \sigma_{ij(2)} ... \sigma_{ij(n_B)}\right\}; \tag{3.23}$$

These two types of matrices are thus functions of the same parameters. We also allow u^A to be heteroscedastic, i.e.

$$E(uu') = \Omega_u = diag\left\{\sigma_{u(1)}^2 \sigma_{u(2)}^2 ... \sigma_{u(n_A)}^2\right\}; \tag{3.24}$$

With these assumptions it follows that,

$$\mathrm{E}(\epsilon\gamma + u)(\epsilon\gamma + u)' = E(\epsilon\gamma\gamma'\epsilon + uu') = \begin{bmatrix} \gamma' \sum_{(1)} \gamma & 0 & & 0 \\ 0 & & \gamma' \sum_{(2)} \gamma & 0 \\ & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \gamma' \sum_{(n_A)} \gamma \end{bmatrix} + \Omega_u; \tag{3.25}$$

and,

$$\mathrm{E}(\in^B \in^{B\prime}) = \begin{bmatrix} \Omega_{11} & \Omega_{12} & \dots & \Omega_{1K_1} \\ \Omega_{21} & \Omega_{22} & & \Omega_{2K_1} \\ & \vdots & \ddots & \vdots \\ \Omega_{K_1 1} & \Omega_{K_1 2} & \dots & \Omega_{K_1 K_1} \end{bmatrix}; \tag{3.26}$$

The matrix (3.25) is denoted V and the matrix (3.26) Ω. Disregarding irrelevant constants the likelihood function then becomes,

$$L = -log\ det(V) + tr\left\{V^{-1}(y^A - X_2^A R\gamma)(y^A - X_2^A R\gamma)'\right\} - log\ det(\Omega) + tr\left\{\Omega^{-1}(x - (I \otimes X_2^B)r)(x - (I \otimes X_2^B)r)'\right\}; \quad (3.27)$$

The model will not be identified unless additional assumptions are made about the nature of the heteroscedasticity.

The TSLS estimate of γ is consistent. This result follows, because it is a function of consistent estimates of the parameters R of the auxiliary relations and it is in itself consistent conditional on R. This is true also for non-normal errors.

γ, however, is not unbiased. This is easily seen from the simple model with $k_1 = k_2 = 1$ and only one γ-parameter:

$$\hat{\gamma} = \left(\frac{\Sigma x_1^B x_2^B}{\Sigma (x_2^B)^2}\right)^{-1} \left(\frac{\Sigma x_2^A y}{\Sigma (x_2^A)^2}\right)$$

$$E(\hat{\gamma}) = E_{X_1}(E(\hat{\gamma}|x_1)) = E_{X_1}\left\{\left(\frac{\Sigma x_1^B x_2^B}{\Sigma (x_2^B)^2}\right)^{-1} \gamma r\right\} \neq r^{-1}\gamma r;$$

The critical assumptions of the present model and of the whole approach are those about the auxiliary relations (3.3b) and (3.3c). It might at first seem very constraining to assume that there exist linear relations which explain the unobserved variables and that these relations are the same for both samples. However, we do not necessarily have to give them a causal interpretation but could rather look upon them as predictive relations. In practical applications of this approach, we will the face the problem of finding good predictors. Linearity is not a binding restriction. At the minor cost of discontinuities, non-linear relations can be transformed into linear relations with dummy variables.

If the samples are sufficiently large, we could use dummy variables only and still obtain a good precision of the estimates. That would be equivalent to grouping the two samples by the same groups, merging them at the group level and then estimating the relation between y and X_1 from group means.

The auxiliary relations must be formulated in such a way that no parameters are needed to predict X_1 in sample A, which cannot be estimated from sample B. This means that the two samples cannot differ to much with respect to observational units, variable definitions etc. However, the two samples do not have to be of the same size, and the number of observations with a particular configuration of X_2-values can differ between the samples. Other differences might also be acceptable. For instance, if the model explains household behavior and sample A includes household data, it might be possible to use data on individual persons in sample B, viz. if household behavior (data) can be predicted from individual behavior (data).

4. A simulation experiment

A simulation experiment was made to illustrate the estimation procedure numerically and to get an idea about its small sample properties. The experiment was based on a demand equation, which relates expenditures on food, beverages and other every day commodities to household disposable income, age of household head and household size. The parameters of this function were estimated from two samples of household data. The first sample included all variables except disposable income. There was, however, in this sample information about tax assessed income for each household member. The second sample included both household disposable income and assessed income, as well as additional variables which could be used to predict disposable income, the most important one being the ownership of owner-occupied houses.[3] The model was specified as follows,

$$\mathrm{LEXP} = 3.594 + .4\mathrm{LDISP} + .04\mathrm{AGE} - .00043\mathrm{AGE}^2 + .644\mathrm{HS} - .0723\mathrm{HS}^2 + \mathrm{u};$$

$$u \sim N(0, \sqrt{.45});$$

3. Deduction of interest payments reduces the assessed income for owners of owner-occupied houses, as compared with non-owners

$$\text{LDISP} = 2.665 + .7445\text{LTAXINC} + .5626\text{DH} - .00000176\text{DHTAXINC} + \in;$$

$$\in \sim N(0, \sqrt{.053});$$

$$E(u \in) = 0;$$

where

LEXP = The logarithm of expenditures on food, beverages and other everyday commodities,
LDISP= The logarithm of disposable income,
AGE= The age of the household head,
HS = Household size,
LTAXINC = The logarithm of the total of the assessed incomes of all household members,
DH = A dummy variable for owners of owner-occupied houses, and DHTAXINC = Assessed income for owners of owner-occupied houses; zero for non-owners.

With minor adjustments the parameter values were obtained as the least squares estimates of the model from a small sample of 144 households, which included survey and register information of all variables. The observations on the exogenous variables of this sample were then used in the simulations. First, the sample was randomly divided into two samples, *A* and *B*. Then for each observation, LDISP was simulated either four times or ten times. In sample *A* the simulated LDISP values were used to simulate LEXP values. In this way we obtained a simulated sample four or ten times larger than the original sample but with the same covariance structure of the exogenous variables. Finally, the parameters of the demand function were estimated by the TSLS procedure using the two subsamples. The whole simulation and estimation procedure was then replicated.

There are altogether seven simulations. In all but the last two of these, the two subsamples were of the same size, 288 in simulation 1 and 2 and 788 in 3 and 5. In simulations 7 and 8 sample A included 360 observations, while sample B was three times as large.

The results are shown in ***Table 1***. They indicate that, when the true predictive relation is used, there is virtually no bias. The bias estimates given in the table are so small that they are dominated by the random fluctuations of the simulation experiment. In simulation 5 the predictive relation used in the estimation was misspecified. The house-owner variable and the interaction variable were both deleted, and the logarithm of assessed income was thus the only predictor.[4] The result is a small bias. The income elasticity is underestimated by 5 per cent, and there is also an 8-9 per cent bias in the estimates of the household size parameters.

Since the LDISP variable is predicted with an error, it might be expected that the efficiency of the TSLS estimator would be less than for the LS estimator based on a complete sample. In ***Table 1*** we can compare the relative root mean-square errors for these two cases.[5] In both cases this measure includes the variability caused by the drawing of a new LDISP vector for each sample. We find that the estimated relative efficiency, defined as the ratio of the two root mean-square errors, is approximately 80 per cent for the income elasticity, i.e. for parameter γ_2. It is between 80 and 90 per cent for the intercept and above 90 per cent for the other parameters. The number of replications is not large enough to justify any conclusions about the dependence of the efficiency on the sample size. When the predictor relation is incorrectly specified there is an additional, but in this particular case, modest loss in efficiency. This is at least partly caused by the bias component.

Finally, ***Table 1*** also shows that the variance formula for the ordinary least-squares estimates applied to the second step of the TSLS procedure, i.e. the diagonal elements of $S^2(\hat{X_1}'\hat{X_1})^{-1}$, gives good estimates of the true variance of $\hat{\gamma}$. S^2 is the residual variance in the second estimation step. The last row of each panel of ***Table 1*** shows the square-root of the mean of the replications of each diagonal element relative to the true parameter value. They differ very little from their corresponding relative root mean-square errors.

4. The full model was, of course, used to simulate data.
5. The relative root mean-square error for the estimator $\hat{\gamma}_i$ is defined as $\sqrt{MSE_i}/\gamma_i$.

Table 1. Results from seven simulation experiments

		γ_1	γ_2	γ_3	γ_4	γ_5	γ_6
Parameter values		3.594	0.400	0.040	-0.00043	0.644	-0.0723
Simulation 1:	TSLS						
n_A	288						
n_B	288						
Replications	800						
Relative bias(%)		0.42	-0.82	2.42	1.72	0.79	1.36
Relative $\sqrt{\text{MSE}}$ (%)		32.82	29.55	57.50	53.49	27.00	36.66
Relative sq. root mean							
variance estimate (%)		33.46	30.30	56.00	53.49	27.83	38.59
Simulation 2:	LS, X_2=LDISP known						
n_A	288						
Replications	400						
Relative $\sqrt{\text{MSE}}$ (%)		29.54	24.60	56.25	53.49	24.78	34.16
Relative sq. root mean							
variance estimate (%)		28.54	24.75	54.75	51.16	26.89	37.48
Simulation 3:	TSLS						
n_A	720						
n_B	720						
Replications	600						
Relative bias (%)		-1.00	0.61	1.83	0.28	-0.14	-1.14
Relative $\sqrt{\text{MSE}}$ (%)		21.82	19.16	35.30	33.23	17.23	24.17
Relative sq. root mean							
variance estimate (%)		21.22	19.21	35.54	33.53	17.64	24.45
Simulation 4:	LS, X_2=LDISP known						
n_A	720						
Replications	1000						
Relative $\sqrt{\text{MSE}}$ (%)		17.57	15.05	34.83	32.94	16.62	23.49
Relative sq. root mean							
variance estimate (%)		17.99	15.61	34.57	32.66	16.99	23.69
Simulation 5:	TSLS, wrong predictors						
n_A	720						
n_B	720						
Replications	600						
Relative bias (%)		1.62	-4.86	2.18	-0.22	9.13	-7.70
Relative $\sqrt{\text{MSE}}$ (%)		24.16	22.57	37.55	35.52	19.66	25.84
Relative sq. root mean							
variance estimate (%)		24.02	21.56	36.22	34.28	17.38	24.51
Simulation 6:	TSLS						

Continued

Table 1. Continued

		γ_1	γ_2	γ_3	γ_4	γ_5	γ_6
n_A	360						
n_B	1080						
Replications	400						
Relative bias (%)		-1.37	2.24	-5.30	5.47	-1.65	3.02
Relative $\sqrt{\text{MSE}\ (\%)}$		31.11	31.31	59.01	56.01	28.78	42.99
Relative sq. root mean							
variance estimate (%)		29.74	29.85	60.06	57.20	27.48	40.72
Simulation 7:	LS, X_2=LDISP known						
n_A	360						
Replications	400						
Relative bias (%)							
Relative $\sqrt{\text{MSE}\ (\%)}$		25.37	23.90	54.13	51.74	27.62	41.92
Relative sq. root mean							
variance estimate (%)		24.91	23.56	56.28	53.79	26.59	39.77

5. Concluding remarks

It is not unusual in nonexperimental studies that variables are missing or replaced by proxies. Even when a new survey is to be made it might not be feasible to collect all items of information from every respondent, either because of the risk for invasion of privacy or because the respondent burden might become so high that the response rate would drop below an acceptable level. One approach to solve this problem has been suggested in this paper. When the statistical problem is to estimate a linear relation between *y* and *x* or an equation in an interdependent linear system, one sample including all variables is not necessarily needed, but two or more samples, each with missing variables, can be used instead. A sampling scheme like matrix sampling could thus be used in combination with the estimation method suggested. A condition is that it is possible to predict the missing variables.

If "true" predictors are used, the two-stage least-squares estimator suggested has good properties. It is consistent asymptotically normal. For a proof see ***Klevmarken (1983b)***. What has been shown here is that the estimator is a maximum-likelihood estimator, if the moment matrices of the regression model and the predictive relations are scalar. When they are not scalar matrices, the maximum-likelihood estimates can be obtained by numerical optimization of the likelihood function.

A sampling experiment has indicated that the two-stage least-squares estimator also has favorable small sample properties. We found virtually no bias, and the decrease in efficiency, relative to the case with one complete sample, was small.

In practice it might be difficult to know the "true" predictors. In the case of the interdependent model, the predictors are given by the model. If all of them cannot be used, e.g. because of a shortage of data, the estimates will no longer be consistent (see ***Klevmarken (1982)***). For the regression model the same conclusion does not necessarily follow. In this case the predictive relations do not necessarily have the same status of theory. To some extent we can choose these relations at our convenience. If x_1 is a stochastic variable, it is always possible to define distributions conditional on some x_2. If these distributions do not have the same mean, $E\left(x_1 | x_2\right)$ could, in principle, be used as a predictive relation. The efficiency of the resulting estimate would, however, depend on how well this relation predicts x_1. If the residual variance is high, the estimates are likely to have a high variance as well. In practical applications it would thus be desirable to use at least part of sample *B* to find good predictors. This search process is stochastic, and the variability introduced by the search should, in principle, be taken into account when the properties of the estimator are evaluated.

With X_1 unobserved and with no supplementary information, the regression model is unidentified. The identification is achieved by the predictive relation, which adds the necessary *a priori* information

and bridges the two samples. *A priori* information might also come in other forms. For instance, if the unobserved X_1 variables do not only explain one y-variable but two or more variables, we could look upon the y-variables as indicators of X_1 and arrive at a model which is similar to a factor analysis model. All these indicators do not necessarily have to come from the same sample. Although the details have to be worked out, it should be possible to combine two or more independent samples in this case as well.

There are common features of the approach suggested in this paper and the general principles of statistical matching. In statistical matching similar observations in the two samples are matched. Similarity is defined either by a grouping principle or by a distance measure defined on a set of variables. These variables basically serve the same purpose as the predictor variables in the TSLS approach. There are, however, also important differences. In most applications of statistical matching there is the implicit assumption of independence between y and X_1 conditional on X_2. The models discussed in this paper do not involve this assumption. The conditional covariance between y and X_1 in the regression model is a function of the structural parameters γ and the moment matrix of the errors in the predictor relations.

Finally, a remark on functional form. Linear models have been assumed throughout this paper. Extending the same approach to non-linear models would be quite conceivable, and there is no reason why this would not be feasible. The properties of the estimates would, however, be more difficult to derive, and this remains to be done.

Acknowledgements

This article was orginally published as "Statistical Review 1983:5, pp 67-79, Essays in Honour of Tore Dalenius, SCB (Statistics Sweden), Stockholm.
Paul Olovsson very efficiently helped with the programming, and Claes Cassel contributed useful comments on a previous draft.

Funding

A grant from The Bank of Sweden Tercentenary Foundation is greatfully acknowledged.

Conflict of Interest

No competing interests reported.

Data and code availability

Not applicable.

References

Barr RS, Stewart WH, Turner JS. 1982. An Empirical Evaluation of Statistical Matching Methodologies. Report Prepared for the Office of Assistant Secretary for Planning and Evaluation. U.S. Department of Health and Human Resources.

Brown C. 1980. Equalizing Differences in the Labor Market. *The Quarterly Journal of Economics* **94**: 113. DOI: https://doi.org/10.2307/1884607

Dalenius T. 1982. A Sample of Ideas for Research and Development in the Theory and Methods of Sample Surveys. *Utilitas Mathematica* **21 A**: 59–74.

Goldberger AS. 1972. Maximum-Likelihood Estimation of Regressions Containing Unobservable Independent Variables. *International Economic Review* **13**: 1. DOI: https://doi.org/10.2307/2525901

Klevmarken NA. 1982. Missing Variables and Two-Stage Least-Squares. Estimation from More than One Data Set. American Statistical Association, Proceedings of the Business and Economic Statistics Section. p. 156–161.

Klevmarken NA. 1983a. Micro Econometrics, the IUI Yearbook 1982/83, The Industrial Institute for Economics and Social Research. Stockholm, Sweden.

Klevmarken NA. 1983b. Asymptotic Properties of a Least-Squares Estimator Using Incomplete Data. Research Report 1983:3. Sweden: Department of Statistics, University of Gothenburg.

Paass G. 1982. Statistical Match with Additional Information. Report IPES.82.0204. Bonn: Gesellschaft fur Matematik und Datenverarbeitung MBH.

Rodgers WL, DeVol EB. 1982. An Evaluation of Statistical Matching. Report from ISR. University of Michigan.

U.S. Department of Commerce. 1980. Report on Exact and Statistical Matching Techniques. Statistical Policy Working Paper 5.

Zellner A. 1970. Estimation of Regression Relationships Containing Unobservable Variables. *International Economic Review* **11**: 441. DOI: https://doi.org/10.2307/2525323