# AN EVALUATION OF STATISTICAL MATCHING

Willard L. Rodgers and Edward B. DeVol, University of Michigan

## INTRODUCTION

Analyses of micro-data often require data that are not available from a single source, but that are available from a set of sources. For example, suppose that one is interested in the relationships among variables in two sets: perhaps one set consists of information about health care expenses incurred by individuals, and the other set consists of information about receipt of various types of welfare benefits. Suppose further that no existing data set contains all of the needed variables, but that two surveys have recently been conducted that, between them, contain all of the necessary variables for large samples of the target population. If mounting a new survey to obtain all of the needed variables from a single sample is not feasible, the only alternative to forgoing the analysis (and therefore, perhaps, developing legislation based on a poor understanding of the empirical relationships) is to utilize any information about relationships that is contained in the data files from the two separate data files. One technique that has been developed to permit such analyses is known as statistical matching.

In a statistical match, individual records from two or more sources are linked on the basis of their similarity on a set of characteristics that are measured in each source.[1] Suppose, for instance, that the set of data sources comprises two files, file A and file B. In order to carry out a statistical match of these files, it is necessary that a common core of information be available about the units in each file. The notation we shall use is as follows: let $\underline{X}_A$ be a set of measured variables on file A, and $\underline{X}_B$ be a set of measured variables on file B. It is assumed that either the components of $\underline{X}_A$ correspond directly to the components of $\underline{X}_B$, or these two sets of variables can be transformed to a common set of characteristics. We shall refer to this set of characteristics that are measured for individuals in both samples as the vector $\underline{X}=(X_1,\ldots,X_P)$. It is on a subset of these P variables that the similarity of units is measured in the process of matching two records. The remaining information or variables in each of the files, of which there is no overlap, shall be referred to as $\underline{Y}=(Y_1,\ldots,Y_Q)$ variables in file A and as $\underline{Z}=(Z_1,\ldots,Z_R)$ variables in file B.

The objective of statistical matching is to create a file (called a matched or synthetic file) in which each record comprises all the $\underline{X}$, $\underline{Y}$, and $\underline{Z}$ variables. For each unit in the A file, a similar unit in the B file is found, with similarity assessed in terms of a function of the $\underline{X}$ variables. The $\underline{Z}$ variables in the B file are then imputed to the matching record in the A file, thus giving rise to a record of complete ($\underline{X}$, $\underline{Y}$, and $\underline{Z}$) data representing not any individual unit, but rather what is referred to as a synthetic unit.

In this paper, we begin by considering what information about the relationships among the $\underline{Y}$ and $\underline{Z}$ variables is implicit in the observed relationships among the $\underline{X}$ and $\underline{Y}$ variables in file A, and among the $\underline{X}$ and $\underline{Z}$ variables in file B. We will observe that in practice this information would almost always be quite limited, and that the process of statistical matching necessarily involves underlying assumptions. After this mathematical exercise, we describe an empirical investigation of the potential usefulness of statistical matching. In this investigation, statistical matches are simulated using data from the Income Survey Development Program (ISDP) 1979 Research Panel, a prototype survey for the proposed Survey of Income and Program Participation (SIPP).[2] Preliminary findings from one such simulated match are presented, and the quality of the statistically matched file is evaluated.

## ASSUMPTIONS INHERENT IN STATISTICAL MATCHING

Let us consider a general framework for statistical matching and examine the assumptions inherent in such a procedure. Suppose that we have two independent samples from a population, say sample A with $n_A$ units and sample B with $n_B$ units. In sample A (henceforth referred to as file A) only $\underline{X}$ and $\underline{Y}$ variables are measured. In sample B (file B) only $\underline{X}$ and $\underline{Z}$ variables are measured. Creating a statistically matched file of synthetic units on which $\underline{X}$, $\underline{Y}$ and $\underline{Z}$ are available is accomplished by imputing to each A unit the $\underline{Z}$ variables from a B unit which, with respect to the common variables $\underline{X}$, is similar. Specifically, a distance function $D(\underline{X}_a,\underline{X}_b)$ is minimized for each unit a ($a=1,\ldots,n_A$) in file A, across all b units ($b=1,\ldots,n_B$) in file B. The inherent assumption in this process is that the conditional distribution of $\underline{Y}$ given $\underline{X}$ is independent of the conditional distribution of $\underline{Z}$ given $\underline{X}$. This assumption is made in the creation of the matched file regardless of the purpose for its creation or the might be used (cf. Sims, 1972).

The linear[3] interrelationships of the $\underline{X}$, $\underline{Y}$, and $\underline{Z}$ variables are given by the variance-covariance matrix:

$$C = \begin{bmatrix} \mathrm{Cov}(\underline{X},\underline{X}) & \mathrm{Cov}(\underline{X},\underline{Y}) & \mathrm{Cov}(\underline{X},\underline{Z}) \\ \mathrm{Cov}(\underline{Y},\underline{X}) & \mathrm{Cov}(\underline{Y},\underline{Y}) & \mathrm{Cov}(\underline{Y},\underline{Z}) \\ \mathrm{Cov}(\underline{Z},\underline{X}) & \mathrm{Cov}(\underline{Z},\underline{Y}) & \mathrm{Cov}(\underline{Z},\underline{Z}) \end{bmatrix} .$$

Of the various components of C, only $\mathrm{Cov}(\underline{Y},\underline{Z})$ cannot be estimated from the two separate samples. Thus, it is often assumed that the covariance between $\underline{Y}$ and $\underline{Z}$ given $\underline{X}$ is known. Typically, it is taken to be zero, which is equivalent to assuming that conditional on $\underline{X}$, the variables $\underline{Y}$ and $\underline{Z}$ are independent, given the assumption of linear relationships. With the conditional independence assumption, statistical matching becomes a valid method of estimating $\mathrm{Cov}(\underline{Y},\underline{Z})$, essentially by creating synthetic units on which both $\underline{Y}$ and $\underline{Z}$ are present.

In general, then, an assumption that

underlies most statistical matching operations is that the distinct variables from two separate datasets are conditionally independent, after controlling on the information common to both files. This is a strong assumption, for which little justification (other than one of practicality) has generally been offered. It implies that $\underline{Y}$'s relationships to $\underline{Z}$ can be totally inferred by $\underline{Y}$'s relationships to $\underline{X}$ and $\underline{Z}$'s relationship to $\underline{X}$. That is,

$$E[\underline{Y}|\underline{Z}] = E[\underline{Y}|\underline{X},\underline{Z}] = E[\underline{Y}|\underline{X}].$$

Occasionally, information about the relationship of a Y-Z pair is available from another source, and the assumption of conditional independence is not made. It should be clear, however, that this does not change the basic assertion: namely, that the statistical matching procedure does not generate new information about the conditional relationship of the Y-Z pair, but only reflects the assumption used in creating the matched file.

To illustrate the strength of the conditional independence assumption, it is useful to consider the total range of values that the correlation between a single observed Y variable and a single observed Z variable could have, given the constraints implied by the observed correlations of those variables with the $\underline{X}$ variables (see Wolfe, 1974). A wide range of possible values indicates that there is at least the possibility of drawing highly misleading conclusions about the covariance of statistically matched variables. A narrow range, on the other hand, indicates that the bivariate relationships among the matched variables can be accurately estimated through the statistical matching procedure. (It should be noted that this discussion is in terms of the possible range of observed correlations of the Y and Z variables. Since all such observed correlations have sampling variability, the range of the population correlations would be wider than the range of observed correlations.)

Table 1 gives the range of possible values for the correlation between a Y and a Z variable, given particular values of the observed multiple correlations of the Y and Z variables on the set of X variables. Only if the multiple correlation of either a Y or a Z variable on the set of available X variables is very high--say above .90--can a reasonably narrow range for the unobserved correlation of the Y and Z variables be specified (see Rodgers and DeVol, 1980, for derivation of ranges of possible values; see also Wolfe, 1974). Multiple correlations as high as .90 are rare at the microdata level, except when one has multiple measures of essentially the same concept.

The situation with respect to statistical matching becomes even more tenuous when the objective is to carry out multivariate analyses involving variables from all three sets--the $\underline{X}$, $\underline{Y}$, and $\underline{Z}$ variables. For example, the last two columns of Table 1 show that if one wishes to estimate the parameters in a causal model which specifies a Z variable as the dependent variable and includes one or more Y variables along with all of the X variables as predictors, the range of possible values for the regression coefficients of the Y variables is extremely wide and centered at 0.

## AN EMPIRICAL TEST OF STATISTICAL MATCHING

The major conclusion of this paper can already be stated, since it rests on the nature of statistical matching rather than any empirical analysis: statistically matched files are a risky basis for any analyses that involve the relationship between a Y and a Z variable.

Table 1: Range of Possible values of Correlation of Y and Z Variables, and their Regression Coefficients, Given Selected Values of their Multiple Correlations with a Set of X Variables

| Observed Multiple Correlations | | Range of Possible Correlations, $r_{YZ}$* | | Range of Standardized Regression Coefficients | |
|---|---|---|---|---|---|
| $R_{Y.X}$ | $R_{Z.X}$ | Lower Bound | Upper Bound | $B_{Z.Y(X)}$ | $B_{Y.Z(X)}$ |
| .99 | .80 | .311 | .481 | $\pm$4.25 | $\pm$.235 |
| | .50 | .125 | .370 | $\pm$6.15 | $\pm$.163 |
| | .30 | .014 | .283 | $\pm$6.76 | $\pm$.148 |
| .80 | .80 | -.040 | .680 | +1.000 | +1.000 |
| | .50 | -.320 | .720 | $\pm$1.443 | $\pm$.693 |
| | .30 | -.452 | .692 | $\pm$1.590 | $\pm$.629 |
| .50 | .50 | -.625 | .875 | +1.000 | +1.000 |
| | .30 | -.751 | .901 | $\pm$1.102 | $\pm$.908 |

* These values all assume a value for the correlation between the predicted values of the Y and Z variables, based on the X variables, of .50. This assumption does not affect the width of the range of possible $r_{YZ}$ values, only its midpoint.

The separate files contain no information about the conditional relationships among the Y and Z variables, and statistical matching adds no information, but only reflects the implicit or explicit assumptions made in the match procedure. An important question, then, is how much confidence can be placed in the assumption of conditional independence. This question cannot be answered in general, of course, but we can explore how often, and how well, such an assumption is met for a particular set of variables.

For this reason, it is useful to carry out empirical tests of statistically matched files in order to provide guidelines for the use of this procedure. Such empirical tests may serve to indicate how often analyses based on statistically matched files lead to erroneous conclusions, and the magnitude of the errors introduced. They can also serve to demonstrate the importance of such factors as: the strength of associations among the X variables and the Y and Z variables; the number of cases in the component files; the nature of the distance function used to match cases and the choice among possible alternative matching procedures.

With respect to the two basic dimensions of a dataset, variables and cases, the features of the ISDP microdata file seem to be typical of files that have been used in statistical matching. The conclusions from our simulation study should therefore be of wide applicability.

## METHODOLOGY

Our basic strategy was to treat the variables measured in a single survey as if they came from two distinct surveys, and to go through various matching procedures with these two files. The variables on the original file are divided into three sets: the X, Y, and Z variables described earlier. Subfile A, then, consists of the X (matching) variables and the Y variables for each of the original cases; and subfile B consists of the X and the Z variables for those same cases. These two files can then be matched as if they came from different sources (with only slight modifications to the matching procedures to prevent, for example, a case on file A from being matched with itself on file B). Analyses of a matched file created in this fashion can be compared with analyses of the original file to provide criteria for evaluating the statistical matching procedures.

The sample: Data from the first interview the ISDP 1979 Research Panel were used as the data source for this study. The dataset is a structured one with 8975 households comprising 24,789 individuals and an extensive range of variables. We focus on the individual level as the matching unit. Since income variables (usually missing for children) would generally be very important in a match data from the proposed SIPP, we confined the simulations to adult cases. We eliminated 3000 adult cases because of inconsistencies between variables in the data file now available to us, leaving us with a sample size of 15,675 adults.

Selection of X, Y, and Z variable sets: The objective was to define three sets of variables that would be typical of the variables that might be encountered in an actual match. That is, we wanted the $\underline{X}$ variables which formed the basis for matching cases to be typical of variables that have been used as match variables in previous statistical matches and to be available for this purpose in matching data from the proposed SIPP with other sample survey data. We also wanted the sets of $\underline{Y}$ and $\underline{Z}$ variables to be sets which might reasonably be expected in two separate surveys.

The 22 X variables that we used include characteristics of the household (type of family, number of adults and children, home ownership, and a total income estimate); and characteristics of the individual (age, sex, race, marital status, education, work status, etc.).

The set of 24 variables that we designated as Y variables are related to welfare and transfer payments: whether or not the respondent received each of seven types of benefit, and if so, the dollar amount received in the proceeding quarter. The specific types of payment were: Social Security benefits; Federal Supplemental Security Income (SSI); Medicare; Medicaid; Worker's, Veteran's and Unemployment Compensation; Food Stamps; and Aid to Families with Dependent Children. For the set of 25 Z variables, we used variables which indicate other income sources (and respective amounts), e.g. earnings, property income, interests and dividends, and so forth.

Unconstrained matching procedures: The data that are described in this paper are from an unconstrained match, a procedure which has the objective of finding the case on the supplemental file that is most similar to each case on the base file, where similarity is necessarily defined in terms of what is known about cases on both files. That is, in terms that we have been using, it is necessary to define a distance function for each pair of cases in terms of the observed values on the X variables. The form of the distance function which we used was defined by Radner et al. (1980, p. 42):

$$D_{ij} = \sum_{p=1}^{p} [W_p \times g_p(X_{ip} - X_{jp})],$$

where $W_p$ is a predefined weight reflecting the importance attached to the $p^{th}$ X variable (Radner's notation has been changed to correspond to that used here). The $g_p$ functions used here were of three types: the absolute difference in values of the X variable; the square of the difference; or an indicator variable indicating agreement or disagreement.

## RESULTS

Through the matching procedure just described, we generated a data file consisting of a record for each case with the original X, Y, and Z variables, and also values from the matched case for each of the Y variables. We present here some preliminary findings in our evaluation of the match procedure.

In the first place we can assess the appropriateness of the assumption of conditional

independence of the matched variables. This aspect of the evaluation does not depend on the simulations; initial analyses of the data revealed the extent to which Y and Z variables are related after controlling on the X variables. Each of the 600 Y-Z pairs is an example of a pair of variables that might be matched in a real match, and it is appropriate to consider the degree to which the conditional independence assumption is violated for each such pair. Most of the partial correlations are small: 89% have an absolute magnitude less then .05. However, about a fifth of the partial correlations would be judged significantly different from zero at the = .001 level, if individual tests had been prespecified for them. Moreover, it is important to assess these partial correlations relative to the zero-order correlations, 68% of which have absolute magnitudes less than .05. In comparison to the zero-order correlations the partial correlations are substantial:

$$[\sum_{Y}\sum_{Z}(r_{YZ.X})^2 / \sum_{Y}\sum_{Z}(r_{YZ})^2]^{1/2} = 0.347.$$

The second part of the evaluation is to assess the robustness of findings obtained through statistical matching in the face of violation of the conditional independence assumption. We plan to examine a large number of bivariate and multivariate relationships, comparing the estimates of various statistics (e.g., correlations, regression coefficients, and proportions of explained variance) that are obtained using the original (measured) variables and the matched variables.

At this point in our analysis, we can report on a limited number of comparisons of actual and matched data. We first examined the univariate and bivariate distributions of the matched variables. Although there were some deviations, the distributions of the matched Y variables were generally quite similar to those of the original Y variables. The means and standard deviations of the matched variables, and their correlations with one another, were almost all close to those of the actual variables.

A more important test of the match is a comparison of the pairwise correlations of the Y and Z variables. If we consider the absolute magnitudes of the differences in correlations observed between a matched Y variable and a Z variable vs. an actual Y variable and that Z variable, these differences are mostly small in absolute terms. Of the 600 correlations of the matched Y variables with Z variables, 89% differed by less than .05 from the corresponding correlations of the actual Y variables. Only eleven of the differences exceeded .10 in absolute value. If we consider the size of differences relative to the actual Y-Z correlations, however, we learn that the differences are rather substantial:

$$[\sum_{YZ}\sum (r_{YZ}^{A}-r_{YZ}^{M})^2 / \sum_{YZ}\sum (r_{YZ}^{A})^2]^{1/2}=.305.$$

We suspect that the small absolute values of most of the discrepancies between the observed and matched correlations should not be interpreted as meaning that statistical matching is a satisfactory technique, necessarily, but that it may only reflect the generally small magnitudes of the observed correlations. One piece of evidence that supports this caution is that the larger the observed correlation between a Y and a Z variable, the larger the difference tends to be between the correlation for the matched vs. the observed Y variable. Table 2 is a cross-classification of the magnitude of the actual correlation and the magnitude of the difference between the actual correlation and the correlation using the matched variable. We conclude from these analyses of the simulated match that there is a distinct possibility of finding misleading relationships if analyses are done with statistically matched files.

To illustrate the potential dangers of analyses based on matched files, consider the relationship between a Y variable which is the answer to a question about whether the individual is receiving Social Security benefits, and a Z variable which concerns whether he or she is receiving benefits from a private pension. The observed correlation between these variables is r = .276, while the correlation based on the matched file is r = .191. While this difference may not appear great, it translates in percentage terms to a

Table 2: Cross-classification of magnitude of observed correlations with magnitudes of difference between observed and matched correlations

| Absolute Value of Actual Correlations | Absolute Value of Difference Between Actual and Matched Correlations: 0.00-0.04 | 0.04-0.08 | 0.08-0.12 | 0.12-0.16 | Total |
|---|---|---|---|---|---|
| 0.00-0.10 | 458 | 37 | 1 | 1 | 497 |
| 0.10-0.20 | 31 | 12 | 3 | 3 | 49 |
| 0.20-0.30 | 16 | 11 | 9 | 0 | 36 |
| 0.30+ | 5 | 9 | 3 | 1 | 18 |

difference between 72%, according to the observed data, and 56%, according to the mateched data, of those who are receiving private pensions who are also receiving Social Security benefits.

## FURTHER RESEARCH

It is too early at this point in our simulation exercise to draw firm conclusions, other than the general conclusion stated earlier: that by the nature of matching, analyses based on a matched filed must be considered risky. As we continue our study, we will be making further assessments of the quality of the data obtained through the unconstrained matching procedure. We also will be comparing matched files obtained by this procedure with files obtained through constrained matching[4], and by a procedure in which predicted values from each case are added to residual values from a matched case (described in Rodgers and DeVol, 1981). We further plan to test the importance of the distance function used to match cases, and of the number of cases in the two files.

We are convinced that simulation studies of this type are important. Statistical matching has become a widely used technique, and offers obvious attractions. However, before we accept statistical matching as a legitimate basis for substantive analyses, we need to subject this technique to more scrutiny than it has received in the past. Simulation studies offer one means of evaluating the usefulness of statistical matching.

## REFERENCES


Barr, R.S. and Turner, J.S. (1978). A new, linear programming approach to microdata file merging. 1978 Compendium of Tax Research, Office of Tax Analysis, Department of the Treasury. U.S. Government Printing Office, Washington, D.C.

Gillo, M.W. and Shelly, M.W. (1974). Predictive modeling of multivariable and multivariate data. Journal of the American Statistical Association, 69, 646-653.

Lininger, C.A. (1980). The goals and objectives of the Survey of Income and Program Participation. 1980 Proceedings of the Section on Survey Research Methods, American Statistical Association, 480-485.

Radner, D.B., Allen, R., Gonzalez, M.E., Jabine, T.B. and Muller, H.J. (1980). Report on Exact and Statistical Matching Techniques. Statistical Policy Working Paper No. 5, U.S. Department of Commerce. U.S. Government Printing Office, Washington, D.C.

Rodgers, W.L. and DeVol, E.B. (1980). Statistical Matching: First Interim Report. Ann Arbor, Mich.: Institute for Social Research.

Rodgers, W.L. and DeVol, E.B. (1981). Statistical Matching: Second Interim Report. Ann Arbor, Mich.: Institute for Social Research.

Sims, C.A. (1972). Comments. Annals of Economic and Social Measurement, 1, 343-346.

Sonquist, J.A., Baker, E.L. and Morgan, J.N. (1971). Searching for Structure (Alias AID-III): An Approach to Analysis of Substantial Bodies of Micro-Data and Documentation of a Computer Program. Ann Arbor, Mich.: Institute for Social Research.

Wolfe, E.N. (1974). The goodness of match. National Bureau of Economic Research Working Paper No. 72 (December).

Ycas, M., and Lininger, C.A. (1980). The Income Survey Development Program: A review. 1980 Proceedings of the Section on Survey Research Methods, American Statistical Association, 486-490.


## FOOTNOTES

1. Further elaboration on these definitions and references to descriptions of statistical matches may be found in a working paper of the Office of Federal Statistical Policy and Standards (Radner et al., 1980).

2. The Income Survey Development Program is co-sponsored by the U.S. Department of Health and Human Services (HHS) and the Department of Commerce (Bureau of the Census). Ycas and Lininger (1980) and Lininger (1980) provide summaries of the history and objectives of the ISDP and the proposed SIPP.

3. Here and throughout this paper, it is assumed that non-linear and nonadditive relationships among the $\underline{X}$ and $\underline{Y}$ variables, and among the $\underline{X}$ and $\underline{Z}$ variables, are accounted for through transformations of the original variables, inclusion of pattern variables, and similar techniques.

4. The constrained match will be carried out by Richard Barr using a technique developed at the U.S. Department of Treasury (Barr and Turner, 1978).