Statistical Matching Using File Concatenation With Adjusted Weights and Multiple Imputations

Donald B. Rubin
Department of Statistics, Harvard University, Cambridge, MA 02138

Statistically matched files are created in an attempt to solve the practical problem that exists when no single file has the full set of variables needed for drawing important inferences. Previous methods of file matching are reviewed, and the method of file concatenation with adjusted weights and multiple imputations is described and illustrated on an artificial example. A major benefit of this approach is the ability to display sensitivity of inference to untestable assumptions being made when creating the matched file.

KEY WORDS: File matching; Incomplete data; Missing data; Sensitivity analysis.

1. INTRODUCTION

In many situations it is necessary to create one micro data base for research purposes from several separate micro data bases, where the files (i.e., the data bases) have some variable in common (i.e., background variables recorded for all units from all files) as well as some that are unique (i.e., that are recorded only for units in one file). In the important and common case of two data bases, say A and B, one set of variables, say X, is observed in both A and B; another set of variables, Y, is observed in file A but not file B; and another set of variables, say Z, is observed in file B but not file A. This situation is depicted in Figure 1. For many evaluation, research, and planning activities, such as those using standard complete-data procedures based on complicated micro analytic simulation models, one file is required with X, Y, and Z available for all units.

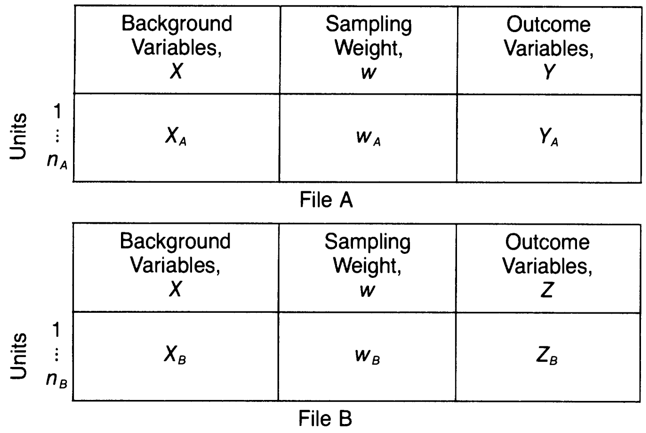
Consequently, an intermediate objective in the context of such activities is the creation of one data base, all of whose units have X, Y, and Z observed, which users can treat as a completely observed data set from a single source. Note that this objective is not the common statistical survey objective of estimating particular finite population quantities, nor is it the common statistical modeling objective of estimating the parameters of one particular hypothetical underlying model generating the variables X, Y, and Z. The created file may be eventually used for such survey or modeling purposes, but all of these purposes cannot be anticipated at the time the file is created.

This is the so-called "statistical file matching" problem. It and its variants have been addressed by Okner (1972, 1974), Sims (1972, 1974), Ruggles and Ruggles (1974), Alter (1974), Wolff (1977), Kadane (1978), and Turner and Gilliam (1978) and more recently by Klevmarken (1981), Paass (1982), Radner (1983), Rubin (1983), Woodbury (1983), and Rodgers (1984). Since no statistical file matching technique can actually recreate the true Y data for file B or the true Z data for file A, the created file has uncertainty that leads to uncertainty in resultant inferences based on the created file. In particular, measures of partial association between Y and Z given X (e.g., partial correlations between Y variables and Z variables given the X variables) are inestimable by any procedure yet are usually assumed to be zero. Typically and incorrectly, such uncertainty is ignored when drawing inferences from statistically matched files.

Several years ago I proposed (Rubin 1980b) an approach to file matching, called *file concatenation with adjusted weights and multiple imputations,* that can be used to create statistically matched files that allow the direct assessment of uncertainty due to both sampling variance and implicit but untestable assumptions regarding relationships between variables in the different files. This method concatenates the files A and B and then multiply imputes values for each missing Y and Z value to reflect uncertainty in the correct value to impute. Letting n_A and n_B be the number of units in the respective files, the result is a file of $n_A + n_B$ units with sampling weights and complete X data for all units, complete Y data for the A units, multiple versions (say, two) of the Z data for the A units, complete Z data for the B units, and multiple versions of the Y data for the B units. The multiple imputations allow an assessment of the effect of uncertainty through the examination of the variation in answers as each set of imputed values is treated as real. Figure 2 depicts the resultant multiply imputed data set with two imputations per missing value, and the two sets of analyses it generates to expose sensitivity. The purpose here is to report and illustrate the suggestions in Rubin (1980b) in the context of more recent work.

2. BACKGROUND

A recent and excellent review of practical issues and literature regarding statistical matching is given in Rodgers (1984). Included is a useful but highly artificial example



Units	Background Variables, X	Sampling Weight, w	Outcome Variables, Y
1 $\vdots$ n_A	X_A	w_A	Y_A

File A

Units	Background Variables, X	Sampling Weight, w	Outcome Variables, Z
1 $\vdots$ n_B	X_B	w_B	Z_B

File B

Figure 1. Two Files From the Same Population. File A is drawn according to sampling scheme A. File B is drawn according to sampling scheme B.

that is used to illustrate two standard procedures: the unconstrained approach (e.g., Okner 1972) and the constrained approach (e.g., Barr and Turner 1980). This example is given here as Tables 1a and 1b.

2.1 Unconstrained Statistical Matching

The unconstrained method in this example created a matched file A with X, Y, and Z data by finding for each A unit the same-sex B unit closest in age (e.g., $A1$ is male and 42, and the best matching B unit is $B5$, who is male and 41; $B5$ is also the closest match to $A2$). The result is displayed in Table 1c. Of course, the analogous procedure could be applied to obtain matches for B units from file A to create a matched B file. More generally, such unconstrained matching methods can be viewed as methods of single imputation for nonresponse.

2.2 Imputation Methods

Specific examples of imputation procedures include the hot-deck as used by the Census Bureau (e.g., Ford 1983) and statistical matching (e.g., Radner 1983; Wolff 1977). Other imputation techniques include regression methods and log-linear methods based on explicit models. Discussion and specific examples of imputation methods are given in the three volumes produced by the National Academy of Sciences Committee on Incomplete Data (Madow and Olkin 1983; Madow, Olkin, and Nisselson 1983; Madow, Olkin, and Rubin 1983).

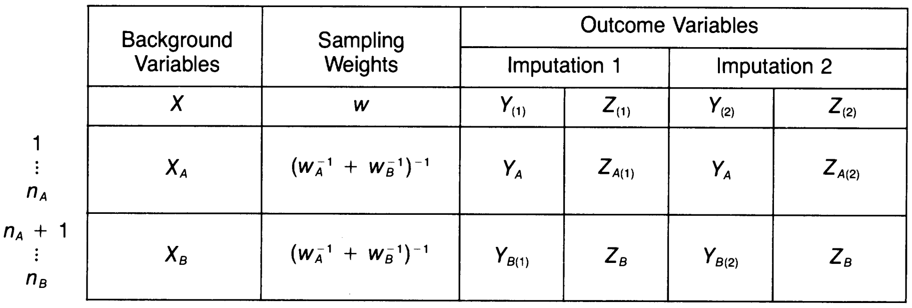
Discussion of matching methods in the different but related statistical context of observational studies is extensive. Some relevant matching references are Althauser and Rubin (1970), Cochran (1968), Cochran and Rubin (1973), Rubin (1973a,b, 1976a,b, 1979b, 1980a), and Rosenbaum and Rubin (1983, 1985). Rubin (1976a) presents a variety of metrics that can be used to define matching rules, and Rosenbaum and Rubin (1985) develop propensity matching, which is discussed in the survey context by Little (1986).

Likelihood-based methods for handling missing data lead to explicit imputation models such as the ones based on linear regression and logistic regression techniques, as illustrated, for example, by Herzog and Rubin (1983). Discussion of maximum likelihood techniques with missing data is now extensive and includes Rubin (1974), Dempster, Laird, and Rubin (1977), Little (1982), and Little and Rubin (1986), which provides many other references. An important advantage of the explicit modeling approach is that it forces the imputer to state what otherwise might be hidden assumptions, such as the conditional independence of Y and Z given X in the file matching context. Once stated, it is natural to ask whether the assumptions are important, that is, whether answers change in substantively important ways as the assumptions are varied.

2.3 Comments on the Unconstrained Approach

A serious problem with the unconstrained approach is that since only one value is imputed for each missing value, it cannot reflect uncertainty about the actual value due to finite samples or due to assumptions in the modeling structure itself, in particular the assumption that Y and Z are conditionally independent given X. Since n_A and n_B are usually very large, the uncertainty due to assumptions in the modeling structure is usually far more important than uncertainty due to sampling variability, although sampling variability is important if interval estimates and significance tests are to be used.

An additional feature that has been viewed as a problem



	Background Variables	Sampling Weights	Outcome Variables			
			Imputation 1		Imputation 2	
	X	w	$Y_{(1)}$	$Z_{(1)}$	$Y_{(2)}$	$Z_{(2)}$
1 $\vdots$ n_A	X_A	$(w_A^{-1} + w_B^{-1})^{-1}$	Y_A	$Z_{A(1)}$	Y_A	$Z_{A(2)}$
$n_A + 1$ $\vdots$ n_B	X_B	$(w_A^{-1} + w_B^{-1})^{-1}$	$Y_{B(1)}$	Z_B	$Y_{B(2)}$	Z_B

Figure 2. A Concatenated File With Adjusted Weights and Multiple (i.e., 2) Imputations. See Figure 1 for notation for original files. $Z_{A(1)}$, $Z_{A(2)}$, $Y_{B(1)}$, $Y_{B(2)}$ are imputed values. Analysis 1 uses X, w_{AB}, $Y_{(1)}$, and $Z_{(1)}$ and produces summary statistic T_1. Analysis 2 uses X, w_{AB}, $Y_{(2)}$, and $Z_{(2)}$ and produces summary statistic T_2.

Table 1a. Rodger's (1984) Example Illustrating Statistical Matching for a Population of N = 24 Units: File A

Unit	*W*	X_1	X_2	*Y*
*A*1	3	1	42	9.156
*A*2	3	1	35	9.149
*A*3	3	0	63	9.287
*A*4	3	1	55	9.512
*A*5	3	0	28	8.484
*A*6	3	0	53	8.891
*A*7	3	0	22	8.425
*A*8	3	1	25	8.867

NOTE: W = sampling weight. X_1 = sex: 1 = male; 0 = female. X_2 = age in years. Y = log(E), where E = personal earnings in dollars. Z = log(P), where P = property income in dollars.

Table 1c. Rodger's (1984) Example Illustrating Statistical Matching for a Population of N = 24 Units: Unconstrained Match for File A

Unit	*W*	X_1	X_2	*Matches'* X_2	*Y*	*Z*
*A*1	3	1	42	41	9.156	7.243
*A*2	3	1	35	41	9.149	7.243
*A*3	3	0	63	59	9.287	6.147
*A*4	3	1	55	52	9.512	5.524
*A*5	3	0	28	33	8.494	6.932
*A*6	3	0	53	59	8.891	6.147
*A*7	3	0	22	33	8.425	6.932
*A*8	3	1	25	28	8.867	4.223

NOTE: See Note to Table 1a.

is that the marginal distributions of Y and Z in the resultant file may not match their marginal distributions in the original files. One method for obtaining a matched file with margins that match the margins in the original files is the constrained approach.

2.4 The Optimal Constrained or Population Matching Approach

The optimal constrained or population matching approach has been used, for example, by the Department of the Treasury to match Statistics of Income and Current Population Survey files. My description of the method is as in Rubin (1980b) and differs from the description in Rodgers (1984) to emphasize that the constrained approach suffers from the same major problem as the unconstrained approach, namely, the illusion of certainty for all reported values.

First, each file (A and B) is "exploded," essentially as if the sample perfectly represented the population. That is, suppose that the ith unit in file A has sampling weight w_{Ai}; then the ith unit along with its X and Y values is duplicated w_{Ai} times. The results are that exploded file A has N units, where N is the number of units in the population, and the exploded file B also has N units (see Fig. 3). Of course the exploded files do not exactly represent the population because the w_{Ai} units in the population represented in file A by the ith unit in general have various values of X and Y, whereas in the exploded file A they all have the same values of X and Y. Because the samples are a small fraction of the population, most X, Y, and Z values in the exploded A and B files are in error when considered to be population values.

If the values of X, Y in exploded file A and of X, Z in exploded file B were the correct population data, then there would be a way to reorder the units in exploded file B such that the reordered units in file B would correspond exactly to the units in file A: the first unit in exploded file A would be the same as the first unit in exploded file B, and so on. If each unit in the population had a unique value of X, a reordering of exploded file B based on matching the files on the basis of X would correctly align the units in the two exploded samples. Of course, the exploded files are not based on population values, so matching the exploded files on X does not correctly align population units. In fact, usually there is no reordering of the exploded file B such that each unit in the reordered exploded file B will have the same value of X as the associated unit in exploded file A.

Table 1b. Rodger's (1984) Example Illustrating Statistical Matching for a Population of N = 24 Units: File B

Unit	*W*	X_1	X_2	*Z*
*B*1	4	0	33	6.932
*B*2	4	1	52	5.524
*B*3	4	1	28	4.224
*B*4	4	0	59	6.147
*B*5	4	1	41	7.243
*B*6	4	0	45	3.230

NOTE: See Note to Table 1a.

The constrained optimization algorithm essentially reorders the exploded file B to minimize an objective function. The objective function is a distance measure in X space between the exploded A file and the reordered exploded B

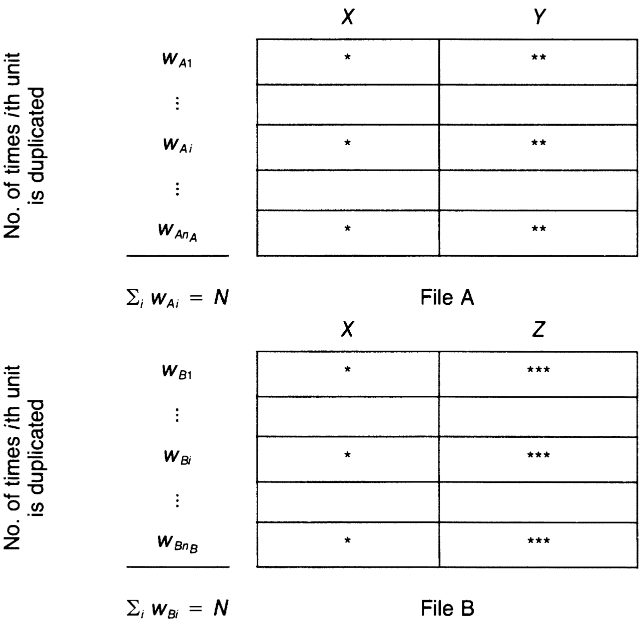


*Figure 3. Exploded Files A and B Used for "Constrained Optimization" File-Matching Approach. *, All rows within block have the same value of X; **, all rows within block have the same value of Y; ***, all rows within block have the same value of Z.*

Table 1d. Rodger's (1984) Example Illustrating Statistical Matching for a Population of N = 24 Units: Constrained Match for Files A and B

Units	W	X_1	X_{2A}	X_{2B}	Y	Z
A1 B2	1	1	42	52	9.156	5.524
A1 B5	2	1	42	41	9.156	7.243
A2 B3	1	1	35	28	9.149	4.223
A2 B5	2	1	35	41	9.149	7.243
A3 B4	3	0	63	59	9.287	6.147
A4 B2	3	1	55	52	9.512	5.524
A5 B1	3	0	28	33	8.494	6.932
A6 B4	1	0	53	59	8.891	6.147
A6 B6	2	0	53	45	8.891	3.230
A7 B1	1	0	22	33	8.425	6.932
A7 B6	2	0	22	45	8.425	3.230
A8 B3	3	1	25	28	8.867	4.223

NOTE: See Note to Table 1a.

file. The result of reordering the exploded B file is effectively one file of N units with four sets of variables: X_A = the values of X from file A, X_B = the values of X from file B, Y = the values of Y from file A, and Z = the values of Z from file B. Of course, this file can be stored without storing N records, since many records have identical values; a weighted file is used to summarize the resultant "population" file. Table 1d displays the results of a constrained match applied to the files of Tables 1a and 1b.

2.5 Comparing the Constrained and Unconstrained Approaches

I regard the automatic matching of margins to the original files as a relatively minor benefit of the constrained approach in most circumstances, especially considering that the real payoff in matching margins arises when samples are not large and census data on the margins exist; and then this constrained approach, which matches sample margins, is not as appropriate as methods designed to match population margins such as ratio and regression adjustment, which can be applied after the matched file is created.

Of particular importance, the optimal constrained approach suffers from the same major defect as the unconstrained approach, namely, it does not represent uncertainty about which values to impute due to either sampling variability or model uncertainty. Simply put, both the unconstrained and constrained approaches are methods of single imputation—that is, one value is imputed for each missing value—and one value cannot represent uncertainty. Within the imputation context, multiple imputation is the natural way to reflect the uncertainty about the correct value to impute.

3. FILE CONCATENATION WITH ADJUSTED WEIGHTS AND MULTIPLE IMPUTATIONS

The method for creating one file that is proposed here treats the two data bases as two probability samples from the same population and creates one concatenated data base with missing data, where the missing data are multiply imputed to reflect uncertainty about which value to impute. The step of file concatenation is described first, and then the creation of multiply imputed values is described, assuming the files have been concatenated.

3.1 Assumptions

Suppose that (a) file A was drawn from population P by sampling scheme A, (b) file B was drawn from population P by sampling scheme B, and (c) the sampling weights (the inverse probabilities of selection) under both sampling scheme A (w_{Ai}) and sampling scheme B (w_{Bi}) are known for each unit in both files. In the artificial example given by Tables 1a and 1b, $w_{Ai} = 3$ and $w_{Bi} = 4$ for all i, since both sampling schemes are simple random samples from a population of 24 units.

We assume that both samples are very large in the sense that Horvitz–Thompson estimators have small enough sampling variance that issues concerning more efficient estimation techniques are not of interest. We also assume that duplicate units cannot be identified; that is, if a unit in the population appears in both the A and B samples, we will not be able to detect this occurrence. Judging from reviews received on Rubin (1980b) this situation seems to be the dominant case in practice (I thank F. Scheuren, G. Sliwa, and an anonymous referee for their thoughtful comments to this effect).

3.2 File Concatenation and Associated Weights

There is a very natural way to combine the two files into one file of $n_A + n_B$ units; simply concatenate the files and calculate a new weight for each unit as exhibited in Figure 2. A straightforward weight for the ith unit in the concatenated file is

$$w_{ABi} = (w_{Ai}^{-1} + w_{Bi}^{-1})^{-1}. \tag{1}$$

To see that these weights are reasonable, let $j = 1, \ldots, N$ index the population P of N units; let V_j, $j = 1, \ldots, N$, be the values of a variable V in the population (V could be an X, Y, or Z variable or squares, products, etc., of these); and let $I_{Aj} = 1$ if unit j is drawn according to sampling scheme A and zero otherwise, and $I_{Bj} = 1$ if unit j is drawn according to sampling plan B and zero otherwise. With weight w_{ABj} for the jth unit in the population, the standard weighted estimate of the total V based on the units in the concatenated sample is

$$\sum_{j=1}^{N} V_j(I_{Aj} + I_{Bj})w_{ABj}, \tag{2}$$

which is the weighted sum of the V values in the sample. Since the expectation of I_{Aj} is the probability of selection, w_{Aj}^{-1}, and analogously for sampling scheme B, the expectation of the estimator (2) over both sampling schemes is

$$\sum_{j=1}^{N} V_j(w_{Aj}^{-1} + w_{Bj}^{-1})w_{ABj},$$

which equals the total V, $\Sigma_{j=1}^{N} V_j$, when the weights w_{ABj} are chosen as suggested in (1). Thus the weights w_{ABi} yield unbiased estimates of population totals. Over the $n_A + n_B$

units in the concatenated file, the weights w_{ABi} in expectation add to N (i.e., let all $V_j = 1$) but might not add exactly to N, and so it might be convenient to form ratio adjusted weights that do add to N:

$$w^*_{ABi} = w_{ABi} \times N \Big/ \sum_{i=1}^{n_A + n_B} w_{ABi},$$

where i indexes the units in the concatenated sample and the summation is over the units in the concatenated sample.

For the simple example of Tables 1a and 1b, $w_{ABi} = (\frac{1}{3} + \frac{1}{4})^{-1} = \frac{12}{7}$ for all i, so $w^*_{ABi} = w_{ABi}$, since $\sum_{i=1}^{14} w_{ABi} = 24 = N$.

When the populations for files A and B are not identical, in general some units represented in sample A have zero probability of being selected under sampling scheme B and vice versa. For example, if the ith unit in file A fell outside population B, then this unit's inverse sampling weight under sampling scheme B, w_{Bi}^{-1}, is zero, and so its weight in the concatenated AB file is simply its weight from file A, w_{Ai}. The population from which the concatenated AB file is a sample is the union of the A and B populations. Consequently, research efforts that focus on a target population smaller than $A \cup B$ will want to eliminate from the concatenated file those units that fall outside the target population.

3.3 Imputation of Missing Data

Having concatenated the A and B files, the plan is to impute Y to file B and Z to file A using the values of (w_{AB}, X), which are observed in both files. When the sampling weights are determined by X, using w_{AB} in addition to X is redundant, but in many practical cases the weights are not functions of X alone (as was drawn to my attention by Scheuren in his comments on Rubin 1980b). Such imputation procedures essentially estimate the conditional distribution of Y given (w_{AB}, X) from file A and the conditional distribution of Z given (w_{AB}, X) from file B and then use these estimates and the values of (w_{AB}, X) to impute missing Y and Z values. Nearly all such procedures commonly used implicitly assume that Y and Z are conditionally independent given (w_{AB}, X)—actually, commonly given just X. Assessing sensitivity of inferences to this assumption of conditional independence is important and can be carried out by using multiple imputation.

3.4 Multiple Imputation

Because the sizes of A and B files are usually very large, the matched file is often manipulated and summarized as if it perfectly represented the target population. This treatment could be acceptable if the matched file is based entirely on real data, but it is not acceptable when vast amounts of data are effectively created by imputation. The proposal here is to use the method of multiple imputation (introduced by Rubin 1977, 1978 and extended in Rubin 1979a, 1980b,c, 1986, Herzog and Rubin 1983, and Rubin and Schenker in press) to display the uncertainty due to the use of fabricated data.

Suppose that K imputation procedures are being considered, each corresponding to a different assumption. For example, in the simple case with scalar Y and Z, the various assumptions could be about the partial correlation r, between Y and Z given (X, w), say $K = 2$, and $r = 0$ and $r = .5$. Suppose that one set of imputations is created under each assumption (when sampling variability is of concern, at least two sets of imputations should be made under each model—see the previously referenced papers on multiple imputation for justification). Each of these files is analyzed by whatever methods would be used on one file; thus instead of one vector of summary statistics, there are K vectors of summary statistics, $T_1, \ldots, T_K$, one for each of the multiply imputed data sets. For example, each vector T_k could represent a table of counts, a table of means, variances, and correlations, and so on. The variability among the T_k displays the uncertainty of estimation due to the uncertainty about which assumption (e.g., which partial correlation) is appropriate. Figure 2 depicts this case with $K = 2$. The procedure is now illustrated using the artificial example of Rodgers (1984).

4. EXAMPLE

A matching method based on a linear regression model is now applied to the example of Tables 1a–1d to illustrate how multiple imputation in a concatenated file can be used to display sensitivity to assumptions about parameters of conditional association. This imputation procedure is purely illustrative and does not represent an ideal method of imputation for a variety of reasons (e.g., the regression of log income on age is probably less sensible than one on log age; imputations should be created by drawing from an approximate posterior predictive distribution of missing values; more than one randomly drawn value should be made under each model to reflect sampling variability). The multiple imputation method used here was chosen because (a) under the conditional independence assumption, it reduces to the method used by Rodgers (1984) to create Table 1c and (b) it is easily generalized to reflect conditional dependence.

4.1 The Method Assuming Conditional Independence of Y and Z given X

Suppose in the concatenated file that $Y = \log(E)$ is regressed on X_1 = sex and X_2 = age, and that $Z = \log(P)$ is regressed on X_1 and X_2 [because the w_{ABi} are constant

Table 2. Ordinary Least Squares Regression Estimates of Outcomes (Y, Z) on X_1 and X_2 in the Concatenated File

	Regression Coefficients			Residual Covariance Matrix	
	1	X_1	X_2	Y	Z
Y	7.943	4.418×10^{-1}	2.002×10^{-2}	6.269×10^{-3}	*
Z	4.767	3.052×10^{-1}	1.465×10^{-2}	*	4.022

NOTE: The Y regression is computed on A units (in Table 1a), the Z regression is computed on B units (in Table 1b), and the results are displayed as a submatrix of the covariance matrix swept on X = (1 = intercept, X_1, X_2).

* Residual covariance between Y and Z given X cannot be computed for the concatenated file because Y and Z are never jointly observed.

Table 3. Predicted Y and Z Values for All Units in Files A and B, Based on Regressions in Table 2 of Y on X and Z on X

Unit	Sex	Predicted Y	Predicted Z	Matches*
*A*1	1	9.226	5.688	*B*5
*A*2	1	9.086	5.585	*B*5
*A*3	0	9.205	5.690	*B*4
*A*4	1	9.486	5.878	*B*2
*A*5	0	8.504	5.177	*B*1
*A*6	0	9.005	5.544	*B*4
*A*7	0	8.384	5.090	*B*1
*A*8	1	8.886	5.438	*B*3
*B*1	0	8.604	5.251	*A*5
*B*2	1	9.426	5.834	*A*4
*B*3	1	8.946	5.483	*A*8
*B*4	0	9.125	5.632	*A*3
*B*5	1	9.206	5.673	*A*1
*B*6	0	8.844	5.427	*A*6

* Matches are obtained for *A* units (*B* units) by finding, within sex, the closest *B* unit (*A* unit) with respect to predicted *Z* (predicted *Y*).

functions of (X_1, X_2), conditioning on the weights is irrelevant]. Using the data of Tables 1a and 1b, we find the results given in Table 2, where the Y regression is obtained from the units with Y observed and the Z regression is obtained from the units with Z observed. Suppose that these regressions are then used to create predicted Y, Z values for all units in the concatenated file. The results are in the middle columns of Table 3.

Each unit missing Y (each B unit) is now matched to the same-sex unit with Y observed who is closest to the predicted Y; thus unit $B1$ is matched with unit $A5$. Analogously, each unit missing Z (each A unit) is now matched to the same-sex unit with Z observed who is closest to the predicted Z; thus from Table 3, unit $A1$ is matched with unit $B5$. The matches' observed values are then imputed for the missing values. Because of the linearity of the regressions on X and the restriction that the matches have identical values of X_1 = sex, this procedure produces the same matches as found by simply matching on X_2 = age within sex, as shown in Table 1c for file A. Imputation 1 in Table 4 exhibits the results for the concatenated file.

4.2 The Method Assuming a Nonzero Partial Correlation Between *Y* and *Z* Given *X*

The method just described would not be of much interest if all that it could do was match on age within sex. It is useful because it suggests a more general procedure applicable for demonstrating sensitivity. Suppose that we know the partial correlation between Y and Z given X is equal to .5. Then this value, coupled with the values in Table 2, allows us to calculate regressions of Y on X and Z and Z on X and Y using standard relationships between such regression coefficients. A convenient form with multivariate X, Y, and Z uses the sweep operator and the formulation in Rubin and Thayer (1978), which leads to taking the submatrix in Table 2 with $.5\,(6.269 \times 10^{-3} \times 4.022)^{1/2} = .07939$ substituted for the asterisks and sweeping on Z to obtain the regression coefficients of Y on $(1, X_1, X_2, Z)$, and analogously sweeping the submatrix on Y to obtain the regression coefficients of Z on $(1, X_1, X_2, Y)$. The resultant regression coefficients are summarized in Table 5.

The coefficients in Table 5 can now be used with the observed data to calculate (a) predicted Z values for all units missing Z and (b) predicted Y values for all units missing Y. For matching purposes, corresponding predicted Y values for units with Y observed and predicted Z values for units with Z observed can be found using the regression coefficients in Table 5 with observed X data and the predicted Y and Z data just calculated. The results are displayed in the middle columns of Table 6. Note that the Y values in Table 6 are close to the Y values in Table 3 because Y is very well predicted from X alone, whereas the Z values in Table 6 differ more from the corresponding values in Table 3 because Z is relatively poorly predicted by X alone.

The same procedure as applied to Table 3 is now applied to Table 6 to create within-sex matches, and then the matches' actual values are imputed. The result is exhibited as Imputation 2 in Table 4.

4.3 Between Imputation Uncertainty

Uncertainty due to the assumed partial correlation between Y and Z given X, which cannot be addressed by the data at

Table 4. Concatenated File With Adjusted Weights and Multiple Imputations for the Example of Tables 1a–d and the Imputation Method of Section 4

				Imputation 1		Imputation 2	
Unit	w_{AB}	X_1	X_2	Y	Z	Y	Z
*A*1	12/7	1	42	9.156	7.243	9.156	4.223
*A*2	12/7	1	35	9.149	7.243	9.149	7.243
*A*3	12/7	0	63	9.287	6.147	9.287	6.147
*A*4	12/7	1	55	9.512	5.524	9.512	7.243
*A*5	12/7	0	28	8.494	6.932	8.494	3.230
*A*6	12/7	0	53	8.891	6.147	8.891	3.230
*A*7	12/7	0	22	8.425	6.932	8.425	6.932
*A*8	12/7	1	25	8.867	4.223	8.867	4.223
*B*1	12/7	0	33	8.494	6.932	8.494	6.932
*B*2	12/7	1	52	9.512	5.524	9.512	5.524
*B*3	12/7	1	28	8.867	4.223	8.867	4.223
*B*4	12/7	0	59	9.287	6.147	9.287	6.147
*B*5	12/7	1	41	9.156	7.243	9.156	7.243
*B*6	12/7	0	45	8.891	3.230	8.891	3.230

Table 5. Regression Coefficients of Y on X and Z and of Z on X and Y Found From Table 2, Assuming the Partial Correlation Between Y and Z Given X is .5

	Predictors				
Outcome	1	X_1	X_2	Y	Z
Y	7.849	4.358×10^{-1}	1.974×10^{-2}	—	1.974×10^{-2}
Z	-9.583×10^{1}	-5.290	-2.389×10^{-1}	1.266×10^{1}	—

hand, is exposed by comparing the results of standard complete-data analyses using Imputation 1 with the analogous results using Imputation 2. For example, suppose that with complete data we would have calculated (a) the regression equation of log income, $\log[\exp(Y) + \exp(Z)]$, on X and (b) the regression equation of Z on X and Y. In the notation of Section 3, the statistic T then has nine components—three coefficients and a residual variance for the first regression and four coefficients and a residual variance for the second regression. Table 7 gives the values of T_1 and T_2 for the concatenated multiply imputed file of Table 4, T_1 for Imputation 1 and T_2 for Imputation 2.

The regression of log income on X is insensitive to the assumption about the partial correlation between Y and Z given X. In contrast, the regression of Z on Y and X is extremely sensitive to this assumption, in fact leading to opposite signs for all regression coefficients. For example, consider the implied effect of sex (X_1) on property income (Z) for individuals with the same age (X_2) and personal earnings (Y): if the partial correlation is zero, males are estimated to have typically exp(.5360) = 1.71 times as much property income as comparable females; whereas if the partial correlation is .5, females are estimated to have typically exp(2.148) = 8.57 times as much property income as comparable males. If both underlying values for the partial correlation are plausible, these results suggest that extreme caution must be used when drawing conclusions from such analyses.

Table 6. Predicted Y and Z Values Given X and Z or Y for All Units in Concatenated File Using Regressions in Table 5 to Predict Missing Values, Assuming That Partial Correlation Between Y and Z Given X is .5

Unit	Sex	Predicted Y	Predicted Z	Match
A1	1	9.209[a]	4.800	B3
A2	1	9.102[a]	6.384	B5
A3	0	9.225[a]	6.732	B4
A4	1	9.493[a]	6.203	B5
A5	0	8.501[a]	5.052	B6
A6	0	8.976[a]	4.106	B6
A7	0	8.394[a]	5.612	B1
A8	1	8.881[a]	5.202	B3
B1	0	8.637	5.679[b]	A5
B2	1	9.420	5.757[b]	A4
B3	1	8.921	5.168[b]	A8
B4	0	9.135	5.761[b]	A3
B5	1	9.237	6.066[b]	A1
B6	0	8.801	4.877[b]	A6

[a] Missing Z is first predicted using observed X and Y, and then this value is used in the prediction equation for Y given X and Z.

[b] Missing Y is first predicted using observed X and Z, and then this value is used in the prediction equation for Z given X and Y.

The method of multiple imputation can thus help to avoid the drawing of unwarranted conclusions.

5. CONCLUSION

I do not regard the method of file concatenation with adjusted weights and multiple imputations to be a panacea for the problems of statistical matching, especially in the simple form described here. For instance, finding appropriate sample weights for both files can be difficult in complex multipurpose surveys, and deciding which sets of assumptions to use when creating multiple imputations can be demanding with highly multivariate data and users with diverse interests.

Nevertheless I believe that the ideas presented here are on target. If multi-user resource files are to be created from two or more source files, and these are to be accessible to users whose analytic tools are standard complete-data procedures, then uncertainty in the process of creating the files must be displayed by such complete-data analyses. Multiple imputation appears to be the only currently available technique that simultaneously satisfies the dual constraints of (a) displaying sensitivity to assumptions used to create the resource files while (b) requiring only standard complete-data methods of analysis.

ACKNOWLEDGMENTS

This work was partially supported by National Science Foundation Grant SES-8311428. The original work on which this article is based was supported by Social Security Administration Order SA-80-6240.

[Received June 1984. Revised May 1985.]

Table 7. Sensitivity of Simple Regression Statistics to Assumptions About the Partial Correlation Between Y and Z Given X

	Statistics From Multiply Imputed Data Set in Table 4	
Regression	Imputation 1, T_1	Imputation 2, T_2
$\log(e^Y + e^Z)$ on X		
Coefficients		
1	8.087	7.992
X_1	3.818×10^{-1}	4.144×10^{-1}
X_2	1.946×10^{-2}	2.084×10^{-2}
Residual Variance	9.615×10^{-3}	1.365×10^{-2}
Z on X and Y		
Coefficients:		
1	1.953×10^{1}	-4.649×10^{1}
X_1	5.360×10^{-1}	-2.148
X_2	3.399×10^{-2}	-1.292×10^{-1}
Y	-1.692	6.482
Residual Variance	2.138	2.943

REFERENCES

Alter, H. (1974), "Creation of a Synthesis Data Set by Linking Records of the Canadian Survey of Consumer Finances With the Family Expenditure Survey 1970," *Annals of Economic and Social Measurement*, 3 (2), 373–394.

Althauser, R., and Rubin, D. B. (1970), "The Computerized Construction of a Matched Sample," *American Journal of Sociology*, 76 (2), 325–346.

Barr, R. S., and Turner, J. S. (1980), "Merging the 1977 Statistics of Income and the March 1978 Current Population Survey," technical report, U.S. Dept. of the Treasury, Office of Tax Analysis.

Cochran, W. G. (1968), "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies," *Biometrics*, 24, 295–313.

Cochran, W. G., and Rubin, D. B. (1973), "Controlling Bias in Observational Studies: A Review," *Sankhya*, Ser. A, 35, 417–446.

Dempster, A. P., Laird, N., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Ford, B. L. (1983), "An Overview of Hot-Deck Procedures," in *Incomplete Data in Sample Surveys (Vol. 2): Theory and Bibliographies*, eds. W. G. Madow, I. Olkin, and D. B. Rubin, New York: Academic Press, pp. 185–207.

Herzog, T., and Rubin, D. B. (1983), "Using Multiple Imputations to Handle Nonresponse in Sample Surveys," in *Incomplete Data in Sample Surveys (Vol. 2): Theory and Bibliographies*, eds. W. G. Madow, I. Olkin, and D. B. Rubin, New York: Academic Press, pp. 209–245.

Kadane, J. B. (1978), "Statistical Problems of Merged Data Files" (OTA Paper 6), in *Compilation of OTA Papers* (Vol. 1), Washington, DC: U.S. Dept. of the Treasury, Office of Technology Assessment.

Klevmarken, N. A. (1981), "Missing Variables and Two-Stage Least Squares Estimation From More Than One Data Set," in *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, pp. 156–161.

Little, R. J. A. (1982), "Models for Nonresponse in Sample Surveys," *Journal of the American Statistical Association*, 77, 237–250.

——— (1986), "Survey Nonresponse Adjustments," *International Statistical Review*.

Little, R. J. A., and Rubin, D. B. (1986), *Statistical Analysis With Missing Data*, New York: John Wiley.

Madow, W. G., and Olkin, I. (1983), *Incomplete Data in Sample Surveys (Vol. 3): Proceedings of the Symposium*, New York: Academic Press.

Madow, W. G., Olkin, I., and Nisselson, H. (1983), *Incomplete Data in Sample Surveys (Vol. 1): Report and Case Studies*, New York: Academic Press.

Madow, W. G., Olkin, I., and Rubin, D. B. (1983), *Incomplete Data in Sample Surveys (Vol. 2): Theory and Bibliographies*, New York: Academic Press.

Okner, B. (1972), "Constructing a New Data Base From Existing Microdata Sets: The 1966 Merge File," *Annals of Economic and Social Measurement*, 1 (3), 325–362.

——— (1974), "Data Matching and Merging: An Overview," *Annals of Economic and Social Measurement*, 3 (2), 347–352.

Paass, G. (1982), "Statistical Match With Additional Information," Internal Report IPES.82.0204, Gesellschaft fur Mathematik und Datenverarbeitung, Bonn, W. Ger.

Radner, D. B. (1983), "Adjusted Estimates of the Size Distribution of Family Money Income," *Journal of Business & Economic Statistics*, 1, 136–146.

Rodgers, W. L. (1984), "An Evaluation of Statistical Matching," *Journal of Business & Economic Statistics*, 2, 91–102.

Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

——— (1985), "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score," *The American Statistician*, 39, 33–38.

Rubin, D. B. (1973a), "Matching to Remove Bias in Observational Studies," *Biometrics*, 29, 159–183; Correction (1974), 30, 728.

——— (1973b), "The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies," *Biometrics*, 29, 184–203.

——— (1974), "Characterizing the Estimation of Parameters in Incomplete Data Problems," *Journal of the American Statistical Association*, 69, 467–474.

——— (1976a), "Multivariate Matching Methods That Are Equal Percent Bias Reducing. I: Some Examples," *Biometrics*, 32, 109–120; Correction, 955.

——— (1976b), "Multivariate Matching Methods That Are Equal Percent Bias Reducing. II: Maximums on Bias Reduction for Fixed Sample Sizes," *Biometrics*, 32, 121–132.

——— (1977), "Formalizing Subjective Notions About the Effects of Nonrespondents in Sample Surveys," *Journal of the American Statistical Association*, 72, 538–543.

——— (1978), "Multiple Imputations in Sample Surveys—A Phenomenological Bayesian Approach to Nonresponse," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 20–34. (Also with addendum, discussion, and rejoinder in *Imputation and Editing of Faulty or Missing Survey Data*, Washington, DC: U.S. Dept. of Commerce, pp. 1–23.)

——— (1979a), "Illustrating the Use of Multiple Imputations to Handle Nonresponse in Sample Surveys," in *Proceedings of the ISI–IASS Manila*, Voorburg, The Netherlands: International Statistical Institute.

——— (1979b), "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies," *Journal of the American Statistical Association*, 74, 318–328.

——— (1980a), "Bias Reduction Using Mahalanobis' Metric Matching," *Biometrics*, 36, 295–298.

——— (1980b), "File Concatenation With Adjusted Weights and Multiple Imputations: A Solution to the File Matching Problems Different in Principle From the Constrained Optimization Approach," unpublished manuscript, Social Security Administration.

——— (1980c), *Handling Nonresponse in Sample Surveys by Multiple Imputations*, Washington, DC: U.S. Bureau of the Census.

——— (1983), Discussion of "Statistical Record Matching for Files" by M. A. Woodbury, in *Incomplete Data in Sample Surveys (Vol. 3): Proceedings of the Symposium*, eds. W. G. Madow and I. Olkin, New York: Academic Press, pp. 203–205.

——— (1986), *Multiple Imputation for Survey Nonresponse*, New York: John Wiley.

Rubin, D. B., and Schenker, N. (in press), "Multiple Imputation for Interval Estimation From Simple Random Samples With Ignorable Nonresponse," *Journal of the American Statistical Association*.

Rubin, D. B., and Thayer, D. T. (1978), "Relating Tests Given to Different Samples," *Psychometrika*, 43, 3–10.

Ruggles, N., and Ruggles, R. (1974), "A Strategy for Merging and Matching Microdata Sets," *Annals of Economic and Social Measurement*, 3 (2), 353–371.

Sims, C. A. (1972), Comments and Rejoinder, *Annals of Economic and Social Measurement*, 1, 343–345; 355–357.

——— (1974), Comment, *Annals of Economic and Social Measurement*, 3, 395.

Turner, J. S., and Gilliam, G. B. (1978), "Reducing and Merging Microdata Files" (OTA Paper 7), in *Compilation of OTA Papers* (Vol. 1), Washington, DC: U.S. Dept. of the Treasury, Office of Technology Assessment.

Wolff, E. N. (1977), "Estimates of the 1969 Size Distribution of Household Wealth in the U.S. From a Synthetic Database," paper presented at the Income and Wealth Conference, Williamsburg, VA.

Woodbury, M. A. (1983), "Statistical Record Matching for Files," in *Incomplete Data in Sample Surveys (Vol. 3): Proceedings of the Symposium*, eds. W. G. Madow and I. Olkin, New York: Academic Press, pp. 173–181.