STATISTICAL MATCHING WITH ASSESSMENT OF UNCERTAINTY IN THE PROCEDURE: NEW FINDINGS

Chris Moriarity, Fritz Scheuren

Chris Moriarity, U.S. General Accounting Office,
411 G Street NW, Washington, DC 20548

KEY WORDS: data fusion, constrained matching

Abstract: In several articles (2001, 2003), we have described a method for statistical matching that allows assessment of the uncertainty introduced during the match that is due to relationships in variables that are not jointly observed. These articles focussed on matches of bivariate normal (X,Z) samples with bivariate normal (X,Y) samples. In this paper we present new findings, with a focus on results from matching when all components of X,Y,Z are bivariate normal.

1. Introduction

We begin with a brief overview of statistical matching in the next section. We then summarize previously-published descriptions of our procedure in Section 3. Then, we present a new generalization of our procedure to higher dimensions in Section 4. We summarize our new findings in Section 5, and provide conclusions and areas for future research in Section 6.

2. Statistical Matching - An Overview

Suppose there are two sample files, File A and File B, taken from two different surveys. Suppose further that File A contains potentially vector-valued variables (X,Y), while File B contains potentially vector-valued variables (X,Z). The objective of statistical matching is to combine these two files to obtain at least one file containing (X,Y,Z).

In contrast to *record linkage, or exact matching*, the two files to be combined are not assumed to have records for the same entities. In statistical matching the files are assumed to have little or no overlap; hence, records for similar entities are combined, rather than records for the same entities. For example, one may choose to match individuals who are similar on characteristics like gender, age, poverty status, health status, etc.

All statistical matches described in the literature have used the X variables in the two files as part of the matching process. To illustrate, suppose File A consisted, in part, of records

$$X_1, Y_1$$
$$X_2, Y_2$$
$$X_3, Y_3$$

while File B has records of the form

$$X_1, Z_1$$
$$X_3, Z_3$$
$$X_4, Z_4$$
$$X_5, Z_5$$

If only the X variables are used to define matches, this is akin to assuming that Y and Z are uncorrelated, given X; if the variables have normal distributions, then the assumption is that Y and Z are conditionally independent, given X. This "conditional independence" assumption has been discussed extensively in the statistical matching literature (e.g., Rodgers (1984), and references given therein).

Given the assumption of conditional independence, it would be immediate that one could create

$$X_1, Y_1, Z_1$$
$$X_3, Y_3, Z_3$$

Notice that matching on X_1 and X_3 (where X is, say, age) does not imply that these are the same entities.

This paper does not necessarily reflect the views or position of the U.S. General Accounting Office.

What to do with the remaining records is less clear and techniques vary. Broadly, the various strategies employed for statistical matching can be grouped into two general categories: "constrained" and "unconstrained." Each is described in turn.

Constrained statistical matching requires the use of all records in the two files and basically preserves the marginal Y and Z distributions. In the above example, for a constrained match one would have to end up with a combined file that also had additional records that used the remaining unmatched File A record (X_2,Y_2) and the two unmatched File B records (X_4,Z_4) and (X_5,Z_5). In other words, all of the records on both files get used. Notice that, as would generally be the case, one could not limit the role of X in the matching so as to require identical values of X to allow a match; in at least some cases, matches would have to be allowed where X's were close (similar) to one another.

Unconstrained matching does not have the requirement that all records are used. Referring to the above example, one might stop after creating (X_1,Y_1,Z_1) and (X_3,Y_3,Z_3). Usually in an unconstrained match, though, all the records from one of the files (say File A) would be used (matched) to "similar" records on the second file. Some of the records on the second file may be employed more than once, or not at all. Hence, in the unconstrained case, the remaining unmatched record on File A, the observation (X_2,Y_2), would be matched to make the combined record $(X_2,Y_2,Z_{??})$. The observations (X_4,Z_4) and (X_5,Z_5) from File B might or might not be included.

A number of practical issues, not discussed in this brief overview, need to be addressed in statistical matching; for example, alignment of universes (i.e., agreement of the weighted sums of the data files) and alignment of units of analysis (i.e., individual records represent the same units).

Rodgers (1984) includes a more detailed example of combining two files, using both constrained and unconstrained matching, than the example we have provided here. We encourage the interested reader to consult that reference for an illustration of how sample weights are used in the matching process, etc.

3. Our Statistical Matching Procedure

In several articles (2001, 2003), we have described a method for statistical matching that uses information from X, Y, and Z in the matching process. Our procedure is an extension of innovative ideas due to Kadane (1978) and Rubin (1986) that allows assessment of the uncertainty introduced during the match that is due to relationships in variables that are not jointly observed, as opposed to simply assuming conditional independence.

The covariance matrix Σ of the vector (X,Y,Z) can be written in partitioned form as

$$\Sigma = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} & \Sigma_{XZ} \\ \Sigma_{YX} & \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZX} & \Sigma_{ZY} & \Sigma_{ZZ} \end{pmatrix}.$$

All elements of Σ can be estimated from File A (containing (X,Y)) or File B (containing (X,Z)) except Σ_{YZ} and its transpose, Σ_{ZY}. Although Σ_{YZ} cannot be estimated directly from Files A or B, the assumption that (X,Y,Z) have a nonsingular distribution places some restrictions on the possible values of Σ_{YZ}, which we refer to henceforth as "admissible" values for convenience. Without loss of generality, Σ can be assumed to be a correlation matrix.

Our algorithm begins with selecting an admissible value of Σ_{YZ}. This value is used in regressions to estimate missing data values of Z in File A and of Y in File B; Z is regressed on X and Y in File A, and Y is regressed on X and Z in File B. Random residuals are imputed to the regression estimates of Z in File A and Y in File B to recover the variance lost during the regression step. The files are matched using constrained matching, with the metric being the Mahalanobis distance on (Y,Z). For matched records, estimated values are replaced with observed values from the other file.

This process is repeated for a range of admissible values of Σ_{YZ}, say *n* of them, to produce *n* distinct synthetic datasets that are available for *n* subsequent analyses that can display the effect of alternative assumptions on the value of Σ_{YZ}.

Note that the random residual imputation step described above is not guaranteed to occur. The amount of residual to impute is estimated by subtracting the variance of the regression estimate of a variable from the estimated variance of the variable; because both File A and File B contribute to this calculation, a nonpositive definite quantity can result.

For example, the variance of Z, Σ_{ZZ}, is estimated using File B. The variance of the regression estimates of Z in File A is given by

$$\begin{pmatrix}\Sigma_{ZX} & \Sigma_{ZY}\end{pmatrix}\begin{pmatrix}\Sigma_{XX} & \Sigma_{XY}\\ \Sigma_{YX} & \Sigma_{YY}\end{pmatrix}^{-1}\begin{pmatrix}\Sigma_{XZ}\\ \Sigma_{YZ}\end{pmatrix}$$

where Σ_{YZ} is specified, Σ_{XX}, Σ_{YY}, and Σ_{XY} are estimated using File A, and Σ_{XZ} is estimated using File B.

For one or both of Y and Z, when a nonpositive definite expression is obtained after the subtraction of the variance of the regression estimate of a variable from the estimated variance of the variable, no imputation of residuals is done for the regression estimate of the variable in the file where it is missing (e.g., the regression estimate for Z in File A). Thus far in our research, occurrences of this phenomenon generally correspond to when the assumed distribution of (X,Y,Z) is close to being singular, as indicated by the smallest eigenvalue of Σ being close to zero.

4. Generalizations Of Our Procedure To Higher Dimensions

Our algorithm, as outlined in Section 3, can be followed for (X,Y,Z) of any dimension.

As might be expected, the procedure is more complicated when the dimension of (X,Y,Z) exceeds 3. In particular, the specification of an admissible value of Σ_{YZ} requires some effort.

One possible strategy is to begin with the "conditional independence value" $\Sigma_{YX}(\Sigma_{XX})^{-1}\Sigma_{XZ}$, which always is an admissible value for Σ_{YZ}. This provides a starting point for generating perturbations that would then need to be checked for admissibility. This option is a good choice if no auxiliary information about the (Y,Z) relationship is available.

Another possibility that might be useful for generating admissible values of Σ_{YZ} either when the dimension of (Y,Z) is low (e.g., univariate Y, bivariate Z), or when many of the (Y,Z) relationships can be estimated using auxiliary sources, is to use the following recursion formula for partial correlations (e.g., Anderson, 1984, p. 43):

$$\rho_{ij\cdot k,l,m,\ldots,p} = \frac{\rho_{ij\cdot l,m,\ldots,p}-\rho_{ik\cdot l,m,\ldots,p}\,\rho_{jk\cdot l,m,\ldots,p}}{\sqrt{1-\rho^2_{ik\cdot l,m,\ldots,p}}\,\sqrt{1-\rho^2_{jk\cdot l,m,\ldots,p}}}$$

Note that in the simplest case of univariate (X,Y,Z), with i=Y, j=Z, and k=X, this formula reduces to

$$\rho_{YZ\cdot X} = \frac{\rho_{YZ} - \rho_{YX}\,\rho_{ZX}}{\sqrt{1 - \rho^2_{YX}}\,\sqrt{1 - \rho^2_{ZX}}}$$

solving for ρ_{YZ} and allowing $\rho_{YZ\cdot X}$ to vary from -1 to 1, a bound for ρ_{YZ} is obtained in terms of ρ_{YX} and ρ_{ZX}, namely

$$\rho_{YZ} = \rho_{YX}\,\rho_{ZX} \pm \sqrt{1 - \rho^2_{YX}}\,\sqrt{1 - \rho^2_{ZX}}$$

This equation displays the range of admissible values for ρ_{YZ}, given the observed values of ρ_{YX} and ρ_{ZX}. The midpoint of the interval is the "conditional independence" value

$$\rho_{YZ} = \rho_{YX}\,\rho_{ZX}$$

Consider again a situation where the dimension of (Y,Z) is low, e.g., (X,Y,Z_1,Z_2). Here, a collection of admissible values for the pair (ρ_{YZ_1} , ρ_{YZ_2}) needs to be generated. One way of doing this, using the recursion formula, is to first select a value for ρ_{YZ_1} in accordance with a bound analogous to what is outlined above; i.e., from the interval,

$$\rho_{YX}\,\rho_{Z_1X} \pm \sqrt{1 - \rho^2_{YX}}\,\sqrt{1 - \rho^2_{Z_1X}}$$

This value of ρ_{YZ_1} then determines $\rho_{YZ_1\cdot X}$ via the equation

$$\rho_{YZ_1\cdot X} = \frac{\rho_{YZ_1} - \rho_{YX}\,\rho_{Z_1X}}{\sqrt{1 - \rho^2_{YX}}\,\sqrt{1 - \rho^2_{Z_1X}}}$$

The recursion formula then is used to obtain the relation

$$\rho_{YZ_2\cdot X,Z_1} = \frac{\rho_{YZ_2\cdot X} - \rho_{YZ_1\cdot X}\,\rho_{Z_2Z_1\cdot X}}{\sqrt{1 - \rho^2_{YZ_1\cdot X}}\,\sqrt{1 - \rho^2_{Z_2Z_1\cdot X}}}$$

which specifies $\rho_{YZ_2\cdot X,Z_1}$ in terms of $\rho_{YZ_1\cdot X}$, $\rho_{Z_2Z_1\cdot X}$, and $\rho_{YZ_2\cdot X}$. $\rho_{Z_2Z_1\cdot X}$ is estimated from File B. Then, by allowing $\rho_{YZ_2\cdot X,Z_1}$ to vary from -1 to 1, the allowable range of values for $\rho_{YZ_2\cdot X}$ can be determined. Applying these bounds to the equation

$$\rho_{YZ_2\cdot X} = \frac{\rho_{YZ_2} - \rho_{YX}\,\rho_{Z_2X}}{\sqrt{1 - \rho^2_{YX}}\,\sqrt{1 - \rho^2_{Z_2X}}}$$

the allowable range of values for ρ_{YZ_2} can be determined that correspond to the selected value of ρ_{YZ_1} , and an admissible value selected.

Returning to the strategy of beginning with the "conditional independence value" $\Sigma_{YX}\,(\Sigma_{XX})^{-1}\,\Sigma_{XZ}$ as a starting point of generating a range of admissible values for Σ_{YZ} , one can determine the maximum range of admissible values for each component of Σ_{YZ} by using the fact that a necessary condition for Σ to be positive definite is that the covariance matrix of (Y,Z) given X is positive definite:

$$\Sigma_{(Y,Z)|X} = \begin{pmatrix} A & B \\ B' & C \end{pmatrix}$$

$$\begin{aligned} \textit{where } A &= \Sigma_{YY} - \Sigma_{YX}\,(\Sigma_{XX})^{-1}\,\Sigma_{XY} \\ B &= \Sigma_{YZ} - \Sigma_{YX}\,(\Sigma_{XX})^{-1}\,\Sigma_{XZ} \\ B' &= \Sigma_{ZY} - \Sigma_{ZX}\,(\Sigma_{XX})^{-1}\,\Sigma_{XY} \\ C &= \Sigma_{ZZ} - \Sigma_{ZX}\,(\Sigma_{XX})^{-1}\,\Sigma_{XZ} \end{aligned}$$

Note that $\Sigma_{(Y,Z)|X}$ is the residual covariance matrix of (Y,Z) after regressing (Y,Z) on X, and that this matrix is block diagonal at the "conditional independence" value of Σ_{YZ} .

For each element of Σ_{YZ} , the maximum amount of "wiggle room", or allowable perturbation, about the conditional independence value can be determined via the square root of the product of the appropriate variance components from A and C in the $\Sigma_{(Y,Z)|X}$ matrix. This follows from the requirement that every principal submatrix of a positive definite matrix must be positive definite. For example, the maximum amount of allowable perturbation about the conditional independence value of $\mathrm{Covar}(Y_i,Z_j)$ is given by the square root of $(\mathrm{Var}(Y_i|X) * \mathrm{Var}(Z_j|X))$. After making this determination for all elements of Σ_{YZ} , one can then iterate within these boundaries to find the widest possible range of admissible values.

We used this strategy successfully to carry out simulation research using bivariate X, Y, and Z. Corresponding to our previously published simulation research, we first specified the distribution of (X,Y,Z). Then, we generated samples of size 1000 for (X,Y) (File A), and for (X,Z) (File B), and then our algorithm was applied.

5. Summary of research using X_1,X_2,Y_1,Y_2,Z_1,Z_2

A total of 1930 simulations were done. Without loss of generality, zero means and unit variances were assumed. First, the (X_i,Y_j) and (X_i,Z_k) correlations were generated

over a range of values such as 0, 0.25, 0.50, and 0.75. For a set of generated correlations, the conditional independence values of Corr(Y_i,Z_j) were computed. If the resulting matrix was positive definite, this set of values was saved for later processing, giving a very large set (~400,000). A 1/1000 subsample of this set was drawn, and then 4 perturbations were generated around each conditional independence value. The combined collection of conditional independence values and perturbations gave a total of 1930 sets of values.

Of the 1930 simulations, residuals were imputed for the regression estimates for both Y=(Y_1,Y_2) in File B and Z=(Z_1,Z_2) in File A in 1595 cases; 202 cases had no imputation for one of (Y,Z); and 133 cases had no imputation at all.

Performance was robust for reproducing the specified value of Σ_{YZ} over the 1930 simulations, with the best results occurring when residuals were imputed for the regression estimates of one or both Y and Z. Table 1 provides details.

Some distortion in estimated covariances occurred when residuals were not imputed for one or both variables; e.g., if residuals were not imputed for the regression estimate of Z in File A, we sometimes observed distortion in Σ_{XZ} in File A. This finding was not unexpected, it was consistent with our previous research.

6. Conclusion, Areas of Future Research

We now have generalized our algorithm to multivariate normal (X,Y,Z) of any dimension. Our simulation results for (X_1,X_2,Y_1,Y_2,Z_1,Z_2) indicate that file sizes of 1000, the same file sizes we used in our previous research on (X,Y,Z), gave satisfactory results. If larger file sizes were used, we would expect better results reproducing specified Σ_{YZ}, and a lower proportion of cases where residuals would not be imputed for the regression estimates for one or both of Y and Z.

Although we have found a way to generate the full range of admissible values of Σ_{YZ} for a given set of observed relationships in File A and File B, more efficient methods deserve study. Another area of worthwhile research is to estimate the proportion of instances where residuals cannot be imputed for the regression estimates of Y and/or Z for various sample sizes and relationships in File A and File B. This information would help guide practitioners in the application of our method, as it is preferable to have file sizes large enough so that it is very likely one can impute residuals to the regression estimates for both Y and Z. Additionally, the performance of our method needs to be assessed when the variables do not have a normal distribution, when one or more variables to be matched are categorical, etc.

It is important to not ever lose sight of the fact that statistical matching, in the absence of auxiliary information, is unable to provide any sort of "best estimate" of the (Y,Z) relationship; the most that can be done is to exhibit variability for a range of plausible values of the (Y,Z) relationship, which allows for sensitivity analyses to be carried out.

Our procedure can be used for this purpose in the multivariate normal framework. Perhaps it is robust enough to extend to other situations as well, too. Additional research is needed, though, as stated above.

References

Anderson, T.W. (1984): An Introduction to Multivariate Statistical Analysis, Second Edition. New York: Wiley.

Kadane, J.B. (1978): "Some Statistical Problems in Merging Data Files," 1978 Compendium of Tax Research, U.S. Department of the Treasury, 159-171. (Reprinted in Journal of Official Statistics, 17, 423-433.)

Moriarity, C. and Scheuren, F. (2001): "Statistical Matching: A Paradigm for Assessing the Uncertainty in the Procedure," Journal of Official Statistics, 17, 407-422.

Moriarity, C. and Scheuren, F. (2001): "Statistical Matching: Pitfalls of Current Procedures," Proceedings of the Section on Survey Research Methods, American Statistical Association.

Moriarity, C. and Scheuren, F. (2003): "A Note on Rubin's 'Statistical Matching Using File Concatenation With Adjusted Weights and Multiple Imputations,' " Journal of Business and Economic Statistics, 21, 65-73.

Rodgers, W.L. (1984): "An Evaluation of Statistical Matching," Journal of Business and Economic Statistics, 2, 91-102.

Rubin, D.B. (1986): "Statistical Matching Using File Concatenation With Adjusted Weights and Multiple Imputations," Journal of Business and Economic Statistics, 4, 87-94.

Table 1: Summary of Simulation Results for $(X_1,X_2,Y_1,Y_2,Z_1,Z_2)$

Simulation Subset	Average absolute difference between specified value of Σ_{YZ} and values computed from matched (Y,Z) pairs				Performance reproducing specified values of Σ_{XZ} in File A and Σ_{XY} in File B
	(Y_1,Z_1)	(Y_1,Z_2)	(Y_2,Z_1)	(Y_2,Z_2)	
overall (1930 simulations)	0.02	0.02	0.02	0.02	usually good
residuals imputed to regression estimates of both Y and Z (1595 simulations)	0.02	0.02	0.02	0.02	good
residuals imputed to regression estimates of one of Y and Z but not the other (202 simulations)	0.01	0.01	0.02	0.02	not good in some instances
residuals not imputed to regression estimates of Y nor Z (133 simulations)	0.04	0.04	0.04	0.04	often not good