REGRESSION-BASED STATISTICAL MATCHING: RECENT DEVELOPMENTS

Chris Moriarity, Fritz Scheuren
Chris Moriarity, U.S. Government Accountability Office,
411 G Street NW, Washington, DC 20548

KEY WORDS: data fusion, RIEPS

Abstract:

We have described a method in several articles (2001a, 2003a, 2003b) for statistically matching two samples. One sample is assumed to contain (X,Z) and the other is assumed to contain (X,Y), both drawn from a common nonsingular normal (X,Y,Z) distribution. Following Kadane (1978) and Rubin (1986), we employ regression in our approach. We assess the uncertainty introduced during the match that is due to the unobserved (Y,Z) relationship by repetition over a range of (Y,Z) values that are consistent with the observed data. In the final step of our algorithm, we replace predicted values with observed data by a match of the two samples to obtain final files consisting only of observed data, consistent with traditional statistical matching procedures. Prior to matching, we add random residuals to our regression-based estimates, an essential step in our method. Our approach for estimation of the amount of residual to add, using subtraction and estimates from both files, can be negative. Raessler (2002) suggests a different approach for residual estimation, which always is nonnegative. We compare the two methods and discuss other recent developments.

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 discuss Raessler's perspective on regression-based statistical matching and our review in Section 4. We discuss some new findings in Section 5, and provide conclusions and areas for future research in Section 6.

This paper does not necessarily reflect the views or position of the U.S. Government Accountability Office.

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 synthetic 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.

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 thus it 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 brief sketch 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 (2001a, 2003a, 2003b), 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} .

Also, as we have mentioned previously (2003a), for a given value of Σ_{YZ} , the process of residual imputation can be repeated several times if there is interest in assessing the impact of imputing various sets of random residuals.

In the initial development of our method, the amount of residual variance to impute was estimated by subtracting the estimated 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. Thus, the random residual imputation step described above was not guaranteed to occur.

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.

4. Raessler's Presentation of Regression-Based Statistical Matching; Discussion

Raessler (2002) provides some history of statistical matching, discussion of regression-based statistical matching, and then advocates use of Monte Carlo Markov Chain (MCMC) procedures and multiple imputation. Our focus here is on Raessler's discussion of regression-based statistical matching.

Note that Raessler (2003) contains a condensed version of the regression-based statistical matching discussion that appears in Raessler (2002).

It is not possible to do a complete comparison between our procedure and any of the results presented by Raessler, because none of her algorithms include a final step of replacing predicted values with observed data by a match of the two samples to obtain final files consisting only of observed data. This final step is consistent with traditional statistical matching procedures, and is part of our method.

Raessler (2002, p. 100) uses the acronym RIEPS (regression imputation with random residual) in her presentation, and we use the term in this paper to denote her algorithm.

RIEPS is similar to the method proposed by Rubin (1986), and discussed by us (2003a). One important difference is that Rubin carried out a final step of replacing predicted values with observed data by a match of the two samples to obtain final files consisting only of observed data; RIEPS does not include this step. Another important difference is that RIEPS imputes a residual to the predicted values, whereby Rubin did not do this.

Following Rubin, Raessler's Equation 4.40 (2002, p. 99) proposed the creation of what we referred to (2003a) as "secondary" predictions of Y in File A (Y_{SEC}) and Z in File B (Z_{SEC}), using X and the "primary" predictions of Z in File A (Z_{PRI}), and X and the "primary" predictions of Y in File B (Y_{PRI}), respectively, to obtain (X,Y,Y_{SEC},Z_{PRI}) for File A and (X,Y_{PRI},Z,Z_{SEC}) for File B. Raessler then

proposed using the secondary predictions to estimate residual variances, specifically, using a function of $(Y-Y_{SEC})^2$ in File A to estimate the residual variance of Y given (X,Z), and using a function of $(Z-Z_{SEC})^2$ in File B to estimate the residual variance of Z given (X,Y).

We showed (2003a) that the variance of the secondary prediction Y_{SEC} is not equal to the variance of Y given (X,Z), and that the variance of the secondary prediction Z_{SEC} is not equal to the variance of Z given (X,Y). Thus, while using the secondary predictions to estimate residual variances of Y and Z has the advantage of always giving nonnegative results, it is not an unbiased method. We conducted some simulation research that provided evidence that Raessler's method tends to underestimate the residual variances, sometimes by substantial amounts.

Raessler (2002, Chapter 4) presents results from a simulation study using multivariate normal data. To be consistent with the framework we described above, we denote the variables as (X_1,X_2,Y,Z), all with mean 0 and covariance matrix

$$\begin{pmatrix} 1.0 & 0.2 & 0.5 & 0.8 \\ 0.2 & 1.0 & 0.5 & 0.6 \\ 0.5 & 0.5 & 1.0 & 0.8 \\ 0.8 & 0.6 & 0.8 & 1.0 \end{pmatrix}. \qquad \textbf{(1)}$$

(Raessler uses (Z_1,Z_2,X,Y) in her presentation.)

The value of σ_{YZ}, 0.8, is in the range of admissible σ_{YZ} values, (0.2794, 0.8872). (The "conditional independence" value is the midpoint of this interval, 0.5833.)

Raessler's simulation consisted of 50 repetitions of the following steps:

1 Sequentially selecting a value of $\sigma_{YZ \mid X}$ from (-0.9, -0.8, -0.4, 0, 0.4, 0.8, 0.9), which corresponds to selecting σ_{YZ} from (0.3098, 0.3402, 0.4618, 0.5833, 0.7049, 0.8265, 0.8569), and repeating the subsequent steps for each value.

2. Generating File A and File B with 2500 observations each, randomly drawn from the specified multivariate normal distribution **(1)** with all means=0.

3. Blanking out Z values from File A and Y values from File B.

4. Using the postulated value of $\sigma_{YZ \mid X}$ and (X,Y) in File A to estimate Z_{PRI}, and using the postulated value of $\sigma_{YZ \mid X}$ and (X,Z) in File A to estimate Y_{PRI}.

5. Using the postulated value of $\sigma_{YZ \mid X}$ and (X,Z_{PRI}) in File A to estimate Y_{SEC}, and using the postulated value of $\sigma_{YZ \mid X}$ and (X,Y_{PRI}) in File B to estimate Z_{SEC}.

6. Using the secondary predictions Y_{SEC} in File A to estimate the residual variance of Y given (X,Z), and then using this estimate to impute residuals for the Y_{PRI} values in File B; similarly, using the secondary predictions Z_{SEC} in File B to estimate the residual variance of Z given (X,Y), and then using this estimate to impute residuals for the Z_{PRI} values in File A.

7. Repeat the previous step of imputing residuals 5 times.

8. Carry out various computations using the Y_{PRI} and Z_{PRI} values, after imputation of residuals.

A summary of Raessler's RIEPS simulations follows (2002, p. 121, from Table 4.6). Raessler also provided more detailed RIEPS results (2002, p. 212, Table A.1).

Assumed σ_{YZ}	Estimated σ_{YZ}
0.3098	0.3186
0.3402	0.3547
0.4618	0.4720
0.5833	0.5835
0.7049	0.7218
0.8265	0.8637
0.8569	0.8833

Note that all estimated σ_{YZ} are larger than the corresponding assumed σ_{YZ}, and the difference is substantial other than for the "conditional independence" value 0.5833.

The S-plus code that Raessler used for her RIEPS simulations (2002, pp. 214-216) includes the following features:

1. The estimated σ_{YZ} shown above, and the estimated means and variances shown in Raessler's more detailed results, all were calculated across the concatenation of File A and File B. This approach is not appropriate, because it combines "observed" data with "estimated/residual imputed" data for the calculation of means and variances, and in the case of σ_{YZ}, the calculation of the covariance of "observed" Y with "estimated/residual imputed" Z is combined with the calculation of the covariance of "estimated/residual imputed" Y with "observed" Z.

2. A correlation calculation was used to estimate σ_{YZ}, rather than a covariance calculation. Due to the tendency of the secondary predictions to underestimate variances, the underestimated variances in the denominator of the correlation calculation tended to inflate the correlations, leading to the upward distortions shown above. The only exception to this general tendency is at the "conditional independence" value of σ_{YZ}, where the regression of Y on (X,Z) reduces to the regression of Y on X, and the regression of Z on (X,Y) reduces to the regression of Z on X, thereby leading to more accurate estimation of residual variance. This can be seen clearly in Raessler's more detailed results (2002, p. 222):

Assumed σ_{YZ}	Estimated σ_{YY}	Estimated σ_{ZZ}
0.3098	0.9532	0.9846
0.3402	0.9335	0.9813
0.4618	0.9516	0.9885
0.5833	1.0008	0.9962
0.7049	0.9617	0.9870
0.8265	0.9297	0.9805
0.8569	0.9519	0.9853

Given the large sample sizes for File A and File B, the estimated σ_{YY} and σ_{ZZ} in all rows should have been very close to 1, that is, similar to the results obtained at conditional independence ($\sigma_{YZ} = 0.5833$), but it can be seen that this was not the case.

Noting our observation that the estimated correlations tended to be distorted upward, we decided to explore the effect of making no changes other than replacing the correlation calculation with a covariance calculation. The following table contains the results:

Assumed σ_{YZ}	Estimated σ_{YZ} (RIEPS)	Estimated σ_{YZ} (revised)
0.3098	0.3186	0.3162
0.3402	0.3547	0.3425
0.4618	0.4720	0.4655
0.5833	0.5835	0.5823
0.7049	0.7218	0.7035
0.8265	0.8637	0.8264
0.8569	0.8833	0.8498

Clearly, the use of the underestimated variances in the denominator of the correlation calculation had a notable effect.

In addition, changing from a correlation calculation to a covariance calculation also affects the variability of the estimates:

Assumed σ_{YZ}	Estimated $s(\sigma_{YZ})$ (RIEPS)	Estimated $s(\sigma_{YZ})$ (revised)
0.3098	0.0182	0.0183
0.3402	0.0179	0.0169
0.4618	0.0135	0.0209
0.5833	0.0106	0.0205
0.7049	0.0096	0.0224
0.8265	0.0076	0.0209
0.8569	0.0081	0.0190

There is no reason, a priori, to suppose that the variance of the estimates would drift in a particular direction across the range of admissible values of σ_{YZ} , as occurred with RIEPS, and it can be seen that replacing the correlation calculation with a covariance calculation gives more stable behavior for the estimates of $s(\sigma_{YZ})$.

In summary, the residual imputation method in RIEPS has the advantage of being nonnegative. However, it appears to be downwardly biased. Calculations across observed data and modeled data are not appropriate, in our view, and the results shown in Raessler (2002) are affected significantly by the use of correlation calculations rather than covariance calculations.

5. New Findings

We explored the effect of using Raessler's residual imputation method in our previous simulation work.

For the trivariate normal simulation we discussed previously (2001a), we found that using Raessler's residual imputation method generally gave results similar to our method for retaining the specified value of σ_{YZ} . However, there were occurrences where Raessler's method led to significant distortions in σ_{XZ} in File A and σ_{XY} in File B; in these instances, our method performed better.

We also explored using a "hybrid" approach (use our method if it gave a nonnegative residual variance estimate, otherwise use Raessler's method). For this set of simulations, there were occurrences where our method gave a negative residual variance estimate for one variable but not the other, but no occurrences where our method gave negative residual variance estimates for both variables. For those instances where our method gave a negative residual variance estimate for one variable, the performance of our method and the hybrid method appeared to be about the same.

For the more general multivariate normal simulation we presented previously (2003b), using $(X_1,X_2,Y_1,Y_2,Z_1,Z_2)$, we also did simulations with a hybrid method, and where Raessler's method was the default. An important difference between the $(X_1,X_2,Y_1,Y_2,Z_1,Z_2)$ simulation and the previous (X,Y,Z) simulation is that former contained instances where our residual variance estimation method gave negative values for both Y and Z, where the latter did not. (Both simulations contained instances where our residual variance estimation method gave a negative estimate for one variable but not the other.) The hybrid method is the same as ours when our residual variance estimates are positive for both Y and Z, and it is the same as using Raessler's method as the default when our residual variance estimates are negative for both Y and Z.

Table 1 summarizes the results from simulations using the various methods. Working up from the bottom of the table, it is clear that when our residual variance estimation method ("2003b results" in Table 1) fails for both Y and Z, it is better to use Raessler's method ("Raessler's residuals" in Table 1) than to not do any residual imputation at all. In contrast to the trivariate simulation, for those occurrences where our method gave a negative residual variance estimate for one variable but not the other, the hybrid method ("hybrid" in Table 1) and Raessler's method both performed better than our

method for reproducing σ_{XZ} in File A and σ_{XY} in File B. Also, these methods did almost as well as ours to retain the specified value of σ_{YZ}. We did not see much difference in the outcomes when our residual variance estimate was positive for both Y and Z.

6. Conclusion, Areas of Future Research

Our statistical matching algorithm is improved by using the RIEPS residual imputation method when our imputation method fails for both Y and Z. Our findings to date give some indication that it is advantageous to use the RIEPS imputation method when our method fails for one but not both Y and Z, but we feel that more research is needed in this area.

It may be fruitful to consider the option of computing the residual variance estimate using our method and Raessler's method, and then always using the larger of the two estimates.

The performance of our method still needs to be assessed for smaller sample sizes, 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.

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

Moriarity, C. and Scheuren, F. (2001b): "Statistical Matching: Pitfalls of Current Procedures", ASA Proceedings of the Joint Statistical Meetings, American Statistical Association.

Moriarity, C. and Scheuren, F. (2003a): "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.

Moriarity, C. and Scheuren, F. (2003b): "Statistical Matching With Assessment of Uncertainty in the Procedure: New Findings", ASA Proceedings of the Joint Statistical Meetings, American Statistical Association, 2904-2909.

Raessler, S. (2002): "Statistical Matching: A Frequentist Theory, Practical Applications, and Alternative Bayesian Approaches", Lecture Notes in Statistics #168, Springer-Verlag.

Raessler, S. (2003): "A Non-Iterative Bayesian Approach to Statistical Matching", Statistica Neerlandica, 57, 58-74.

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)	
residuals imputed to regression estimates of both Y and Z (1595 simulations)					
2003b results (and hybrid)	0.02	0.02	0.02	0.02	good
Raessler's residuals	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)					
2003b results	0.01	0.01	0.02	0.02	not good in some instances
Raessler's residuals	0.02	0.02	0.02	0.02	good
hybrid	0.02	0.02	0.02	0.02	good
residuals not imputed to regression estimates of Y nor Z (133 simulations)					
2003b results	0.04	0.04	0.04	0.04	often not good
Raessler's residuals (and hybrid)	0.02	0.02	0.02	0.02	usually good