Regression-Based Data Fusion: Robust Residual Imputation

Chris Moriarity[1], Fritz Scheuren[2]
[1]National Center for Health Statistics, 3311 Toledo Road, Hyattsville, MD 20782 USA
[2]NORC, 55 East Monroe Street, Chicago, IL 60603 USA

Abstract
We have described a method (2001a, 2003a, 2004) for merging two independent samples using data fusion (also known as statistical matching). One sample contains (X,Z) and the other contains (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 merge 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. An essential part of our algorithm is to add random residuals to the regression estimates. Our initial approach for estimating the residual variance could fail (be negative) because it used subtraction of estimates from both files. Innovations due to D'Orazio, et al. (2006a) and Kiesl and Raessler (2009) give improved results, yielding a robust algorithm.

Key Words: Statistical matching

1. Introduction

We provide a brief introduction to data fusion (also known as statistical matching) in this section. The statistical matching book by D'Orazio, et al. (2006a) is a comprehensive reference, and is recommended reading for anyone who wishes to know more about the subject.

Suppose there are two sample files, File A and File B, taken from two different surveys. Suppose further that File A contains variables (X,Y), while File B contains variables (X,Z). X, Y, and/or Z may be vectors. The objective of data fusion is to combine the information in these two files to obtain at least one synthetic file containing variables (X,Y,Z).

In contrast to record linkage (also known as exact matching), the two files to be combined are not assumed to have records for the same entities. In data fusion 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 historic literature have used the X variables that appear in both files as part of the matching process. To illustrate with a simple example, suppose File A has 3 records with X, Y values:

(x1, y1), (x2, y2), (x3, y3), where x1, y1, etc. are specific values of X, Y,

while File B has 4 records with X, Z values:

(x1, z1), (x3, z3), (x4, z4), (x5, z5), where x1, z1, etc. are specific values of X, Z.

In this example, x1 and x3 are given values of X, and they appear in both files. x2, x4, and x5 are given values of X that appear only in one file.

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 data fusion literature (e.g., Rodgers (1984), and references given therein). It has been shown (e.g., Moriarity and Scheuren (2001a)) that the "conditional independence" value of the covariance of (Y,Z) is just one of many possible values that are consistent with the observed data. Assuming conditional independence can lead to final synthetic files with very different distributions than final synthetic files created with other assumed, feasible values of the covariance of (Y,Z).

Given the assumption of conditional independence, one could merge (match) the records with identical values of X to create:

(x1, y1, z1), (x3, y3, z3).

Notice that matching on the given values of x1 and x3 (where X is, say, age) does not imply that the (y1, z1) and (y3, z3) paired data came from the same entities in the population.

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 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 (x2, y2) and the two unmatched File B records (x4, z4) and (x5, z5). 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 (x1, y1, z1) and (x3, y3, z3). 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 (x2, y2), would be matched to make the combined record (x2, y2, z??). The observations (x4, z4) and (x5, z5) from File B might or might not be included.

A number of practical issues, not discussed in this brief overview, often need to be addressed in data fusion; for example, alignment of universes (i.e., the sums of the weights in the data files should be equal) and alignment of units of analysis (i.e., individual records represent the same units, e.g., persons).

Rodgers (1984) includes a more detailed example of combining two files, using both constrained and unconstrained matching, than the brief sketch 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.

Some more recent literature (e.g., Raessler (2002)) includes examples of data fusion where an actual match (a "statistical match") of the data files does not occur. For this reason, we have decided that the terminology "data fusion" is preferable to "statistical matching", because it is more general.

2. The Use of Regression in Data Fusion

Regression-based data fusion was described in Kadane (1978) and Rubin (1986). Moriarity and Scheuren (2001a, 2001b, 2003a, 2003b, 2004) described innovations and extensions of the methods described by Kadane and Rubin, Kadane's method in particular.

Kadane (1978) described a model for data fusion. One model assumption was that the variables in the sample files had a multivariate normal distribution. He pointed out that this model was not universal; sample files could contain variables that were binary, variables that took integer values only, etc. The data fusion literature contains a number of contributions that consider these situations, e.g., D'Orazio, et al. (2006b). Clearly, regression-based data fusion is not always the most appropriate method to apply. Given the current state of development of regression-based data fusion, alternatives should be considered anytime the sample files contain data that do not appear to have a multivariate normal distribution, and it is not feasible to apply transformations to the data to obtain something reasonably close to a multivariate normal distribution.

Kadane's use of regression was to produce estimates of the "missing" values in the sample files: for File A, containing (X,Y), the "missing" value to be estimated was Z; for File B, containing (X,Z), the "missing" value to be estimated was Y. In order to produce these estimates, Kadane needed to make some assumption about the (Y,Z) relationship, and he correctly noted that a number of different assumptions about the (Y,Z) relationship were consistent with the observed data. He emphasized the necessity for exploring a range of assumptions about the (Y,Z) relationship. His emphasis of this necessity was the genesis of the paradigm of assessing the uncertainty in data fusion.

Kadane's model, with some minor changes and innovations, has been shown to be a sound theoretical framework when the variables in the sample files have a multivariate normal distribution (Moriarity and Scheuren (2001a)). Thus, regression-based data fusion can be considered a feasible and defensible approach for sample files with continuous variables that have a multivariate normal distribution.

3. Adding Random Residuals to Regression Estimates

The most essential innovation we developed for Kadane's and Rubin's methods is adding random residuals to the regression estimates. The methods did not work correctly without this innovation (Moriarity and Scheuren (2001a)).

Our initial approach for this step (e.g., Moriarity and Scheuren (2001a)), which used subtraction of estimates created with information from both files, was not guaranteed to work correctly. It was not always feasible because it could yield a residual variance estimate that was negative. It became apparent (Moriarity and Scheuren (2003b)) that this was a crippling limitation when the residual variance being estimated was close to 0.

The method for generating residuals for RIEPS (Raessler (2002, p. 100) defines RIEPS as "regression imputation with random residual") discussed in Raessler (2002) used a different approach that had the advantage of always being feasible, but also had shortcomings (Moriarity and Scheuren (2004)).

D'Orazio, et al. (2006a) suggested using maximum likelihood estimation. Research we conducted in 2009 indicated that maximum likelihood estimation gave better results than our initial method, but there still were occurrences of negative residual variance estimates. Specifically, for the simulation described in Moriarity and Scheuren (2001a), our initial method gave negative residual variance estimates for the Y variable in File B for 143 out of the 1873 repetitions of the simulation; the maximum likelihood estimation method gave negative residual variance estimates for the Y variable in File B for 107 out of the 1873 repetitions of the simulation.

Kiesl and Raessler (2009) suggested an innovation of Raessler's RIEPS algorithm to address the shortcomings that were noted in Moriarity and Scheuren (2004). The essence of the innovation was to iterate one part of the RIEPS algorithm to improve the statistical properties of the residual variance estimate. The original algorithm used what we referred to as the "primary" regression estimates (with no residual added) to produce "secondary" regression estimates; a sum of squares calculation involving the secondary regression estimates was the basis of the RIEPS residual variance estimate. The innovation iterates the primary/secondary process by taking the residuals estimated by the first set of secondary estimates, adding these residuals to the primary regression estimates, producing a new set of secondary regression estimates, generating an updated set of residuals, etc.

We implemented Kiesl and Raessler's suggested innovation in the simulation framework described in Moriarity and Scheuren (2001a). For each occurrence of the simulation, the iterative process was implemented with the following rules/parameters:

1. Maximum number of iterations allowed: 25
2. If the updated residual variance estimate was less than or equal to the previous one, the iteration process was terminated
3. If the relative increase in the updated residual variance estimate [(update-current)/current] was less than a specified tolerance (we used 0.005) the iteration process was terminated

We allowed these rules to operate independently for the Y and Z iterations.

A comparison of Kiesl and Raessler's suggested innovation with the methods we have previously investigated (our initial method, the original RIEPS method, the maximum likelihood method) shows clearly that Kiesl and Raessler's innovation is the best method to use. In Table 1, "d_resy" is the average difference between the Y residual variance estimate and the true value, and "d_resz" is the corresponding Z average difference. "min_y" shows the largest underestimate of the Y residual variance estimate, and "min_z" shows the corresponding Z underestimate.

Table 1

	d_resy	min_y	d_resz	min_z
our initial method	0.006	-0.34	0.04	0.01
original RIEPS	-0.04	-0.33	-0.02	-0.19
maximum likelihood	0.02	-0.18	0.02	-0.10
Kiesl/Raessler innovation	0.03	-0.14	0.02	-0.05

The innovation provides a notable improvement in the properties of the residual variance estimates, compared to the original RIEPS method. The downward bias of the original RIEPS method (negative values of d_resy, d_resz) is not present in the innovation. As with the original RIEPS method, the innovation always produces a nonnegative residual variance estimate. The statistical properties of the residual variance estimates from the innovation are similar to those generated from the maximum likelihood method.

Table 2 provides detailed information about the number of iterations that occurred in the 1873 simulations:

Table 2

Number of Y iterations	Number of Z iterations	Frequency	Cum.
0	0	79	79
1	1	295	374
1	2	121	495
2	2	386	881
3	1	135	1016
3	3	231	1247
4	1	44	1291
4	2	19	1310
5	3	217	1527
6	5	48	1575
6	7	79	1654
7	5	5	1659
10	9	93	1752
10	10	38	1790
13	10	82	1872
19	16	1	1873

Table 2 shows that the maximum number of allowed iterations (25) was an appropriate choice for the specified tolerance level of 0.005. In all but one case (when there were 19 Y iterations and 16 Z iterations) the outcome would have been the same if we had used a maximum of 15 instead of a maximum of 25.

Table 2 also shows that it is useful to allow the stopping rules to operate independently for the Y and Z iterations.

Note that there were 79 occurrences of "0" iterations for both Y and Z in Table 2. This actually corresponds to the situation where there is no convergence of the iterative process, which can and does occur; the residual variance estimates continue to increase with every iteration. Thus, it is important to implement stopping rules that both check for convergence (e.g., relative growth less than a small value) and control the total number of iterations (to halt iterative processes that are diverging). Not surprisingly, almost all of the occurrences of divergence occur for simulated values when Corr(Y,Z) is far from the conditional independence value; the (X,Y,Z) distribution is still non-singular, but approaching singularity. Of course, in the absence of auxiliary information about Corr(Y,Z), these areas should be included when assessing the uncertainty in data fusion, even if they are "hard" areas to work with.

4. Determining the Range of Admissible Values For the (Y,Z) Relationship

Kiesl and Raessler's innovation for generating random residuals solves one of the two major previously unsolved problems in regression-based data fusion. The remaining major unsolved problem is analytically determining the range of admissible values for the (Y,Z) relationship (the covariance matrix $\sum_{YZ}$ or the correlation matrix ρ_{YZ}), so that a lattice of admissible values that represent all areas of the admissible space can be efficiently generated. (By "admissible", we mean a value of $\sum_{YZ}$ or ρ_{YZ} (Cov(Y,Z) or Corr(Y,Z) in the univariate case) that yields a positive definite covariance matrix for (X,Y,Z).) Solutions currently exist for a number of special cases, but a general result that has been proven mathematically is not yet known to exist.

For univariate (X,Y,Z), the range of admissible Corr(Y,Z) is known. The values must fall in the interval:

$$\text{Corr}(X,Y)*\text{Corr}(X,Z) \pm \sqrt{(1-(\text{Corr}(X,Y))^2)}*\sqrt{(1-(\text{Corr}(X,Z))^2)}$$

The midpoint of this interval, Corr(X,Y)*Corr(X,Z), is the "conditional independence" value of Corr(Y,Z). That is, if Y and Z are independent given X, then Corr(Y,Z)=Corr(X,Y)*Corr(X,Z).

Rodgers and DeVol (1982) derived a bound for the range of admissible Corr(Y,Z) for multivariate X, univariate (Y,Z). Let ρ_{XX} be the correlation matrix of X, ρ_{XY} be the correlation matrix (in this case, a column vector) of (X,Y), and ρ_{XZ} be the correlation matrix (in this case, a column vector) of (X,Z). Rodgers and DeVol showed Corr(Y,Z) must lie in the interval:

$$\rho_{YX} (\rho_{XX})^{-1} \rho_{XZ} \pm \sqrt{(1-(\rho_{YX} (\rho_{XX})^{-1} \rho_{XY})^2)}*\sqrt{(1-(\rho_{ZX} (\rho_{XX})^{-1} \rho_{XZ})^2)}$$

Again, the midpoint of this interval, $\rho_{YX} (\rho_{XX})^{-1} \rho_{XZ}$, is the conditional independence value of Corr(Y,Z).

Raessler (2004) determined via a grid search that the space of admissible values for Corr(Y,Z_1) and Corr(Y,Z_2) in the two-dimensional case (X,Y,Z_1,Z_2) is an ellipse, with the conditional independence value at the center.

Raessler and Kiesl (2009) determined analytically that for multivariate Y and univariate Z, the space of admissible values for the correlation matrix is an ellipsoid, with the conditional independence value at the center:

$$(\rho_{YZ} - \rho_{YX} (\rho_{XX})^{-1} \rho_{XZ})' \, C \, (\rho_{YZ} - \rho_{YX} (\rho_{XX})^{-1} \rho_{XZ}) < 1,$$

$$\text{where } C=((\rho_{ZZ} - \rho_{ZX} (\rho_{XX})^{-1} \rho_{XZ})^{-1}(\rho_{YY} - \rho_{YX} (\rho_{XX})^{-1} \rho_{XY})^{-1})$$

Note: since Z is univariate, $(\rho_{YZ} - \rho_{YX} (\rho_{XX})^{-1} \rho_{XZ})$ is a column vector, and $(\rho_{ZZ} - \rho_{ZX} (\rho_{XX})^{-1} \rho_{XZ})^{-1}$ is a scalar. Also, without loss of generality, Raessler and Kiesl assumed that the covariance matrix was a correlation matrix, so $\rho_{ZZ}=1$.

Raessler and Kiesl's formula also applies to the case of univariate Y and multivariate Z, it covers the case (X,Y,Z_1,Z_2) previously solved via a grid search, and it also yields the univariate (X,Y,Z) result and Rodgers and DeVol's result that are shown above.

Note that in general, the space of admissible values is convex. If $(\sum_{YZ})_1$ and $(\sum_{YZ})_2$ are two admissible values, then the corresponding covariance matrices for (X,Y,Z), say, $\sum_1$ and $\sum_2$, are both non-singular, positive definite matrices. Any linear combination of $\sum_1$ and $\sum_2$, that is, $a*\sum_1 + b*\sum_2$, where a, b are nonnegative scalars, a+b=1, can be shown to be positive definite, using the definition that A is positive definite if x'Ax>0 for all nonzero vectors x. Thus, any linear combination of two admissible values $(\sum_{YZ})_1$ and $(\sum_{YZ})_2$ is also an admissible value.

The conditional independence value of $\sum_{YZ}$, $\sum_{YX} (\sum_{XX})^{-1} \sum_{XZ}$, always is an admissible value. Thus, in the general case, the conditional independence value should be the center point of the space of admissible values.

A comparison of Raessler and Kiesl's result with $\sum_{(Y,Z)|X}$, the covariance matrix of (Y,Z) given X (the residual covariance matrix of (Y,Z) after regressing (Y,Z) on X), indicates a clear link. At the conditional independence value of $\sum_{YZ}$, $\sum_{YX} (\sum_{XX})^{-1} \sum_{XZ}$, $\sum_{(Y,Z)|X}$ is block diagonal, with the upper block equal to $\sum_{YY} - \sum_{YX} (\sum_{XX})^{-1} \sum_{XY}$, and the lower block equal to $\sum_{ZZ} - \sum_{ZX} (\sum_{XX})^{-1} \sum_{XZ}$. C in Raessler and Kiesl's result is the product of the inverses of these blocks. The column vector $(\sum_{YZ} - \sum_{YX} (\sum_{XX})^{-1} \sum_{XZ})$ in the Raessler/Kiesl formula is the upper diagonal element of $\sum_{(Y,Z)|X}$.

A potential generalization of Raessler and Kiesl's result for vector Y (say, with dimension m) and vector Z (say, with dimension n) is to assume without loss of generality that the covariance matrix is a correlation matrix, and construct a column vector (call it YZ) of length mn from the rows of $(\rho_{YZ} - \rho_{YX} (\rho_{XX})^{-1} \rho_{XZ})$, and construct the direct product (or Kronecker product) of $(\rho_{YY} - \rho_{YX} (\rho_{XX})^{-1} \rho_{XY})^{-1}$ and $(\rho_{ZZ} - \rho_{ZX} (\rho_{XX})^{-1} \rho_{XZ})^{-1}$, which is dimension mn by mn; call this direct product C. Then, construct the following:

$$(YZ)' \, C \, (YZ) < 1.$$

Note that this always gives an expression that is conformable, i.e., the matrices have dimensions that are suitable for matrix multiplication.

A limited empirical evaluation of this formula for a range of $(X_1,X_2,Y_1,Y_2,Z_1,Z_2)$ correlation matrices we generated for the research we presented at the 2003 Joint Statistical Meetings (Moriarity and Scheuren (2003b)) gave good results. We plan to do a more extensive empirical evaluation.

5. Conclusion

Regression-based data fusion has progressed significantly in the last ten years from its ad hoc origins. It has been shown to be a methodology with a sound theoretical basis for large sample files that are simple random samples from multivariate normal distributions. However, open questions remain. There are opportunities for future contributions to this area.

When a proven analytic result is available for determining the range of admissible values for $\sum_{YZ}$ in all cases, the existing contributed R package 'StatMatch' (D'Orazio (2009)) could become more powerful than it already is by incorporating this result and the Kiesl/Raessler innovation for generating random residuals.

It is important to remember that data fusion, in the absence of auxiliary information, cannot be expected to provide any sort of "best estimate" of the (Y,Z) relationship. What data fusion can do is create synthetic datasets for a range of plausible values of the (Y,Z) relationship, which allows sensitivity analyses to be carried out.

References

D'Orazio, M., Di Zio, M., and Scanu, M. (2006a): Statistical Matching: Theory and Practice, Chichester: Wiley.

D'Orazio, M., Di Zio, M., and Scanu, M. (2006b): "Statistical Matching for Categorical Data: Displaying Uncertainty and Using Logical Constraints", Journal of Official Statistics, 22, 137-157.

D'Orazio, M. (2009): "Package 'StatMatch'", available online at one of the mirror sites shown at http://cran.r-project.org/mirrors.html, e.g., http://lib.stat.cmu.edu/R/CRAN/web/packages/StatMatch/StatMatch.pdf

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 2001 in Journal of Official Statistics, 17, 423-433.)

Kiesl, H. and Raessler, S. (forthcoming, cited in Raessler and Kiesl, 2009): "How Valid Can Data Fusion Be?", Journal of Official Statistics.

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.

Moriarity, C. and Scheuren, F. (2004): "Regression-Based Statistical Matching: Recent Developments", ASA Proceedings of the Joint Statistical Meetings, American Statistical Association, 4050-4057.

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

Raessler, S. (2004): "Data Fusion: Identification Problems, Validity, and Multiple Imputation", Austrian Journal of Statistics 33(1-2), 153-171.

Raessler, S. and Kiesl, H. (2009): "How Useful Are Uncertainty Bounds? Some Recent Theory With an Application to Rubin's Causal Model", Proceedings of the 57th Session of the International Statistical Institute.

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

Rodgers, W.L. and DeVol, E.B. (1982): "An Evaluation of Statistical Matching". Report submitted to the Income Survey Development Program, Department of Health and Human Services, by the Institute for Social Research, University of Michigan.

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