Some Statistical Problems in Merging Data Files

Joseph B. Kadane

Comment

Christopher A. Sims

Some Statistical Problems in Merging Data Files

Joseph B. Kadane

Suppose that two files are given with some overlapping variables and some variables unique to each of the two files. Notationally, let X represent the common variables, Y, the variables unique to the first file, and Z, the variables unique to the second file. Thus the basic data consist of a sample of pairs (X, Y) and a sample of pairs (X,Z).

Merging of such microdata files may occur in two contexts. In the first, the files are known to consist of the same objects or persons, although their identities may be obscured by measurement errors in the common variables X. In the other case, the two files are random samples from the same population, but only accidentally will the same object or person be on both lists.

To want to merge data files in the first context is a very natural impulse. A merged file permits statements about (Y,Z) cross-classifications that are unavailable without merging. If the measurement errors in the variables X are low (for instance, if X includes accurate social security numbers), the merging can be very accurate, and the meaning of an item in a merged file is clear. It represents the (X,Y,Z) information on the object or person in question.

Merging data files in the second context requires greater caution. Again, facts are sought about (Y,Z) cross-classifications, but the items in the merged file have no natural meaning. The information on the Z variables for persons in the first file and on

---

JOSEPH B. KADANE is with the Department of Statistics at Carnegie-Mellon Unievrsity.

the Y variables for persons in the second file are missing. A mechanical method of merging can be seductive in this context because it will produce a file of records with X,Y, and Z entries inviting treatment as if they refer to the same persons. Yet it is clear that information cannot be created by the merging process where none existed before. Great care must be exercised in the second context.

One important method, reported by Okner (1972a), sets up "equivalence classes" of X's and makes a random assignment of an (X,Y) with an (X,Z) among "equivalent" (X,Z)'s that achieve a minimum closeness score. Sims (1972a, 1972b) stresses the need for a theory of matching and criticizes the Okner procedure for making the implicit assumption that Y and Z, given X, are independent. Peck (1972) defends the assumption, while Okner (1972b) discusses the validity of the assumption in various cases. Budd (1972) compares Okner's procedure to one then being used in the Commerce Department.

A second round of discussion—Okner (1974), Ruggles and Ruggles (1974), and Alter (1974)—shows some improvements in method but a continuing concentration on equivalence classes. Sims (1974) again stresses his belief that the methods proposed will not perform well in sparse X-regions.

The first section of this report considers the case in which the lists are known to consist of the same objects or persons, and the second section takes up the case in which the lists are unrelated random samples from the same population. Although the final section, "Why Match?", is obviously speculative, that term really describes all of the work in this paper.

Files Consist of the Same Objects or Persons

A Statistical Model

We assume that originally there were true triples (X_i,Y_i,Z_i) that had a normal distribution with means (μ_X,μ_Y,μ_Z) and some covariance matrix. These were broken into two samples, (X_i,Y_i) and (X_i,Z_i), and then independent normal measurement error $(\epsilon_i^1,\epsilon_i^2)$ was added. Let

$$X_i^1 = X_i + \epsilon_i^1$$

and

$$X_i^2 = X_i + \epsilon_i^2,$$

where $(\epsilon_i^1, \epsilon_i^2)$ has a normal distribution with zero mean. Suppose, also, ϵ_i^1 has covariance matrix Ω_{11} and ϵ_i^2 has covariance matrix Ω_{22}, and that ϵ_i^1 and ϵ_i^2 have covariance matrix Ω_{12}. Then we observe a permutation of the paired observations (X_i^1, Y_i) and (X_i^2, Z_i).

There are two ways in which the assumed joint normality of X, Y, and Z is restrictive. First, some of our data is binary or integer-valued. Second, this implies that all the regressions are linear, which is not likely to be the case, as pointed out by Sims (1972a, 1972b, 1974). One way around that problem might be to assume joint normality region-by-region in the X space. This thought is not pursued further here.

Let $T_i = (X_i^1, Y_i)$ and $U_i = (X_i^2, Z_i)$ be vectors of length k and l respectively, where without loss of generality we take $k \leq l$. Also without loss of generality, take $\mu_X = 0$, $\mu_Y = 0$, $\mu_Z = 0$. The covariance matrix of T and U can be written as

$$\Sigma = \left[\begin{array}{cc:cc} \Sigma_{XX}+\Omega_{11} & \Sigma_{XY} & \Sigma_{XX}+\Omega_{12} & \Sigma_{XZ} \\ \Sigma_{YX} & \Sigma_{YY} & \Sigma_{YX} & \Sigma_{YZ} \\ \hdashline \Sigma_{XX}+\Omega_{12} & \Sigma_{XY} & \Sigma_{XX}+\Omega_{22} & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_{ZY} & \Sigma_{ZX} & \Sigma_{ZZ} \end{array}\right] = \left[\begin{array}{c:c} \Sigma_{11} & \Sigma_{12} \\ \hdashline \Sigma_{21} & \Sigma_{22} \end{array}\right].$$

Let

$$\Sigma^{-1} = \left[\begin{array}{c:c} C_{11} & C_{12} \\ \hdashline C_{21} & C_{22} \end{array}\right]$$

so that, in particular, we have

$$C_{12} = (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1}.$$

Note that all these covariances can be estimated easily except Σ_{YZ} and $\Sigma_{XX}+\Omega_{12}$. Treatment of them is deferred.

Now suppose that $v_1, \ldots, v_n$ is the random permutation of $T_1, \ldots, T_n$ which is observed, and $w_1, \ldots, w_n$ is the random permutation of $U_1, \ldots, U_n$ which is observed. Let $\phi = [\phi(1), \ldots, \phi(n)]$ be a permutation of the integers $1, \ldots, n$.

According to DeGroot and Goel (1976), the likelihood function of ϕ is

$$L(\phi) = \exp\left\{-\tfrac{1}{2}\sum_{i=1}^{n} v_i' C_{12} w_{\phi(i)}\right\}.$$

Thus the maximum likelihood ϕ minimizes

$$C(\phi) = \sum_{i=1}^{n} v_i' C_{12} w_{\phi(i)}.$$

Let

$$p_{ij} = v_i' C_{12} w_j.$$

Then minimizing $C(\phi)$ is equivalent to minimizing

$$C = \Sigma p_{ij} a_{ij},$$

subject to the conditions

$$\Sigma_i a_{ij} = 1,$$

$$\Sigma_j a_{ij} = 1,$$

and

$$a_{ij} = 0 \text{ or } 1,$$

which is a linear assignment problem (Degroot & Goel, 1976).

There may be cases in which v_i and w_j occur several times in the files and consequently are recorded together. In general, suppose that v_i occurs q_i times $(i = 1, \ldots, n)$ and w_j occurs y_j times $(j = 1, \ldots, m)$, where we assume

$$\sum_{i=1}^{n} q_i = \sum_{j=1}^{m} w_j.$$

Then a simple transformation of $C(\phi)$ yields the minimization of

$$\sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} a_{ij},$$

subject to the conditions

$$\Sigma_i \, a_{ij} = y_j \text{ for } j = 1, \ldots, m,$$

$$\Sigma_j \, a_{ij} = q_i \text{ for } i = 1, \ldots, n,$$

and

$$a_{ij} = \text{non-negative integers.}$$

This minimization is in the form of a transportation problem. The matrix C_{12} appears to be a natural choice of a distance function in this context.

Information About Σ_{YZ}

One of the difficulties of this method is that it requires knowledge of Σ_{YZ}. There are several possible sources of such informa-

tion. First, from a coarse but perfectly matched sample, certain elements of Σ_{YZ} may be known. If so, surely this information should be used. Second, the assumption may be made, as is customary in the literature on matching, that Y and Z are conditionally independent given the X's. That is,

$$f(Y,Z|X^1,X^2) = f(Y|X^1,X^2)f(Z|X^1,X^2).$$

The covariance matrix of $(Y,Z|X^1,X^2)$ is (Anderson, 1958, pp. 28, 29)

$$\begin{pmatrix} \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZY} & \Sigma_{ZZ} \end{pmatrix} - \begin{bmatrix} \Sigma_{YX^1} & \Sigma_{YX^2} \\ \Sigma_{ZX^1} & \Sigma_{ZX^2} \end{bmatrix} \begin{pmatrix} \Sigma_{X^1X^1} & \Sigma_{X^1X^2} \\ \Sigma_{X^2X^1} & \Sigma_{X^2X^2} \end{pmatrix}^{-1} \begin{bmatrix} \Sigma_{X^1Y} & \Sigma_{X^1Z} \\ \Sigma_{X^2Y} & \Sigma_{X^2Z} \end{bmatrix}.$$

Conditional independence occurs iff the upper-right partitioned submatrix is zero, i.e., iff

$$\Sigma_{YZ} - (\Sigma_{YX^1}\ \Sigma_{YX^2}) \begin{pmatrix} \Sigma_{X^1X^1} & \Sigma_{X^1X^2} \\ \Sigma_{X^2X^1} & \Sigma_{X^2X^2} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{X^1Z} \\ \Sigma_{X^2Z} \end{pmatrix} = 0.$$

Thus this assumption gives a condition that uniquely defines Σ_{YZ} in terms of the other Σ's. Some simplification of this answer is possible. Using

$$\Sigma_{YX^1} = \Sigma_{YX^2} = \Sigma_{YX} \text{ and } \Sigma_{ZX^1} = \Sigma_{ZX^2} = \Sigma_{ZX},$$

we have

$$\Sigma_{YZ} = (\Sigma_{YX}\ \Sigma_{YX}) \begin{pmatrix} \Sigma_{X^1X^1} & \Sigma_{X^1X^2} \\ \Sigma_{X^2X^1} & \Sigma_{X^2X^2} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{XZ} \\ \Sigma_{XZ} \end{pmatrix}.$$

Suppose, without loss of generality, that

$$\begin{pmatrix} \Sigma_{X^1X^1} & \Sigma_{X^1X^2} \\ \Sigma_{X^2X^1} & \Sigma_{X^2X^2} \end{pmatrix}^{-1} = \begin{pmatrix} R & S \\ S' & V \end{pmatrix}.$$

Then

$$\begin{aligned} \Sigma_{YZ} &= (\Sigma_{YX}\ \Sigma_{YX}) \begin{pmatrix} R & S \\ S' & V \end{pmatrix} \begin{pmatrix} \Sigma_{XZ} \\ \Sigma_{XZ} \end{pmatrix} \\ &= (\Sigma_{YX}R + \Sigma_{YX}S' \quad \Sigma_{YX}S + \Sigma_{YX}V) \begin{pmatrix} \Sigma_{XZ} \\ \Sigma_{XZ} \end{pmatrix} \\ &= \Sigma_{YX}R\Sigma_{XZ} + \Sigma_{YX}S'\Sigma_{XZ} + \Sigma_{YX}S\Sigma_{XZ} + \Sigma_{YX}V\Sigma_{XZ} \\ &= \Sigma_{YX}(R + S' + S + V)\Sigma_{XZ}. \end{aligned}$$

A well-known fact about inverses of partitioned matrices (Rao, 1965, p. 29) is

$$\begin{bmatrix} A & B \\ B' & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + F\,E^{-1}F' & -FE^{-1} \\ -E^{-1}F' & E^{-1} \end{bmatrix},$$

where

$$E = D - B'A^{-1}B \text{ and } F = A^{-1}B.$$

Then

$$\begin{aligned} R+S'+S+V &= A^{-1}+FE^{-1}F'-FE^{-1}-E^{-1}F'+E^{-1} \\ &= A^{-1}+(I-F)E^{-1}(I-F)' \\ &= A^{-1}+(I-A^{-1}B)E^{-1}(I-B'A^{-1}) \\ &= A^{-1}(A+(A-B)(D-B'A^{-1}B)^{-1}(A-B'))A^{-1}. \end{aligned}$$

Thus, in our case,

$$\Sigma_{YZ}=\Sigma_{YX}\Sigma^{-1}_{x^1x^1}(\Sigma_{x^1x^1}+(\Sigma_{x^1x^1}-\Sigma_{x^1x^2})(\Sigma_{x^2x^2}-\Sigma_{x^2x^1}\Sigma^{-1}_{x^1x^1}\Sigma_{x^1x^2})^{-1} \cdot(\Sigma_{x^1x^1}-\Sigma_{x^2x^1})\Sigma^{-1}_{x^1x^1}\Sigma_{XZ}.$$

Thus Σ_{YZ} is given by this equation as a function of Σ_{YX}, Σ_{XZ}, $\Sigma_{x^1x^1}$, $\Sigma_{x^2x^2}$, and $\Sigma_{x^2x^1}$. All of these can be estimated directly except the last, $\Sigma_{x^2x^1}=\Sigma_{XX}+\Omega_{12}$.

Estimation of $\Sigma_{x^2x^1}=\Sigma_{XX}+\Omega_{12}$

There are really two topics in this section. First I consider the elicitation of the measurement error process variance-covariance matrix Ω. Then I consider how to use that with other information to obtain $\Sigma_{x^2x^1}$.

In the elicitation of Ω, I must first emphasize what it is *not*. It does not refer to the levels of the common variables X. That is, we are dealing only with the spread in measured X's caused by the measurement process. Second, it does not refer to any systematic bias there may be in the measurement error process, but refers only to variability around what would be expected, taking into account both the level of the X variable and the measurement bias, if any.

Begin, then, with the diagonal elements of Ω, which are variances. Each variance refers to a specific measurement error variable, that is, to a specific X-variable and the associated source (one of the two). Choose any value for the true underlying X variable, for instance x. Write down what you think the measurement bias b is. (This must be independent of the value you gave for the X-variable, x. While this is not exactly the case, take for b a typical value). Not everyone with this true value x will have a measured value $x+b$. Write down the number y such that only 33.3 percent of such people will lie below y and 66.7 percent, above. Write down the number w such that 66.7 percent will lie below w and 33.3 percent, above. These numbers should line up so that $y<x+b<w$. There are now two measures for the standard deviation: 2.17 $(w-x-b)$ and 2.17$(x+b-y)$. These values should be close. The variance is then the square of the standard deviation. This vari-

ance should not, according to the model, depend on x, so try it for a number of x's and hope that the results are close. If they are, take the median as the best value. If they are not, the model is not a good representation of reality.

Now we turn to the off-diagonal elements of Ω, which have to do with the relationship between two variables. Suppose that those variables are A and B. Then the work above defines for us the following: x_A, b_A, σ_A, w_A, and y_A, and similarly, x_B, b_B, σ_B, w_B and y_B. We now are trying to capture the extent to which A and B affect one another. The characteristic we focus on is the proportion p of times a measurement error on A is smaller than w_A and, simultaneously, a measurement error on B is smaller than w_B. If A and B have nothing to do with one another, this proportion would be $2/3 \times 2/3 = 4/9 = .44$, slightly under 50 percent. However, if A and B are related to one another, this proportion p may vary from .44. Write down the number you think is correct, and then convert it into a correlation between A and B using table 1.

This yields a ρ_{AB} for each pair of variables A and B. The proper element for Ω is then the covariance of A and B, which is $\sigma_A \sigma_B \rho_{AB}$.

Not every matrix formed in this way is positive definite, as a covariance matrix must be. Hence, additional checks must be made to ensure that the covariance matrix is positive definite. One convenient way to achieve this is to augment Ω one row and column at a time, making use of the following simple fact:

If A is positive definite, then

$$\begin{pmatrix} A & b \\ b' & c \end{pmatrix}$$

is positive definite, iff $c - b' A^{-1} b > 0$. The proof is simple (see Kadane et al., 1977.)

In this way, every element of Ω can be elicited. Now the sample also has some information about Ω, which can be used as a check on the process. The variance-covariance matrix of X^1 is $\Sigma_{X^1X^1} = \Sigma_{XX} + \Omega_{11}$ and of X^2, $\Sigma_{X^2X^2} = \Sigma_{XX} + \Omega_{22}$. This gives two independent estimates for Σ_{XX}, namely $\Sigma_{X^2X^2} - \Omega_{22}$ and $\Sigma_{X^1X^1} - \Omega_{11}$. These should be very close. I suggest rechecking the work if they are not. If they are, then an estimate for Σ_{XX} is at hand. Finally we obtain $\Sigma_{X^2X^1} = \Sigma_{XX} + \Omega_{12}$, for we now have estimates of both of the latter.

TABLE 1.—*Relation between p and ρ*

p	.33	.35	.37	.40	.42	.44	.46	.48	.50	.54	.59
ρ	—.9	—.7	—.5	—.3	—.1	0	.1	.3	.5	.7	.9

Source: National Bureau of Standards (1959).

Some Concluding Remarks About This Case

The case in which the files are known to consist of the same objects or persons is not well understood. Recently DeGroot and Goel (1975) obtained the astonishing result that such matched samples contain information about Σ_{YZ}. Their results suggest that there may not be a lot of information, and we do not know whether the amount of information in some relevant sense increases or decreases (or stays constant) with n. In particular, we do not know if a consistent estimate of Σ_{YZ} can be found in this case, although this writer's intuition is that it cannot.

Another case, one in which the lists may or may not contain the same individuals, is called record linkage. A few important papers in record linkage have been written by DuBois (1969), Fellegi and Sunter (1969), Newcombe and Kennedy (1962), and Tepping (1968).

Matching When the Files Are Random Samples from the Same Population

We assume here that there were true triples (X_k,Y_k,Z_k) that had a normal distribution with means (μ_X,μ_Y,μ_Z) and some covariance matrix. Suppose that in some of these triples the X coordinates were lost, yielding a sample $(X_j,Y_j),(j=1,\ldots,m)$, and that for others the Y coordinates were lost, yielding a sample (X_i,Z_i), $(i=1,\ldots,n)$. The parameters μ_X, μ_Y, μ_Z, Σ_{XX}, Σ_{XY}, Σ_{XZ}, Σ_{YY}, and Σ_{ZZ} can all be estimated consistently, and so we will take them as known. However, the covariance matrix of Y and Z, Σ_{YZ}, cannot be consistently estimated from such data.

In fact, in the domain in which Σ_{YZ} is such that the matrix

$$\begin{pmatrix} \Sigma_{YY} & \Sigma_{YZ} \\ \Sigma_{ZY} & \Sigma_{ZZ} \end{pmatrix}$$

is positive semidefinite, nothing is learned from the data about Σ_{YZ}. In Bayesian terms, whatever our prior on Σ_{YZ} was, the posterior distribution will be the same (see Kadane, 1975 for other examples of this).

Hence we cannot hope to make realistic progress on this problem without a prior probability distribution on Σ_{YZ}. Our intention is to trace through the analysis using a particular value for Σ_{YZ}, for the purpose of obtaining results that would ultimately yield the expected value of some quantity—for instance, the expected amount of taxes a particular kind of tax schedule would raise. The taxes

raised would then be a random variable, where the uncertainty would arise from the uncertainty about Σ_{YZ}. Hence we may assume that the distribution of Σ_{YZ} is known, and we may take values of Σ_{YZ} from the distribution, weighting the final results with the probability of that particular value of Σ_{YZ}. We proceed, then, with a value for Σ_{YZ} sampled in this way.

A natural first thing to do is to estimate the missing values, and the obvious way to do that is by the conditional expectation:

$$E(Z_j|X_j,Y_j) = \mu_Z + (\Sigma_{ZX}\ \Sigma_{ZY})\begin{pmatrix}\Sigma_{XX} & \Sigma_{XY}\\ \Sigma_{YX} & \Sigma_{YY}\end{pmatrix}-1\begin{pmatrix}X_j-\mu_X\\ Y_j-\mu_Y\end{pmatrix}.$$

Let

$$\Sigma_{RS\bullet T} = \Sigma_{RS} - \Sigma_{RT}\Sigma_{TT}^{-1}\Sigma_{TS}$$

for any matrices R, S, and T.

Then

$$\begin{aligned}\mathrm{E}(Z_j|X_j,Y_j) &= \mu_Z + (\Sigma_{ZX}\ \Sigma_{ZY})\begin{pmatrix}\Sigma_{XX\bullet Y}^{-1} & -\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY\bullet X}^{-1}\\ -\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX\bullet Y}^{-1} & \Sigma_{YY\bullet X}^{-1}\end{pmatrix}\begin{pmatrix}X_j-\mu_X\\ Y_j-\mu_Y\end{pmatrix}\\ &= \mu_Z + \Sigma_{ZX\bullet Y}\Sigma_{XX\bullet Y}^{-1}(X_j-\mu_X) + \Sigma_{ZY\bullet X}\Sigma_{YY\bullet X}^{-1}(Y_j-\mu_Y).\end{aligned}$$

Similarly, we may predict missing Y_i with its conditional expectation

$$E(Y_i|X_i,Z_i) = \mu_Y + \Sigma_{YX\bullet Z}\Sigma_{XX\bullet Z}^{-1}(X_i-\mu_X) + \Sigma_{YZ\bullet X}\Sigma_{ZZ\bullet X}^{-1}(Z_i-\mu_Z)\bullet$$

Then the joint distribution of $(X_j,Y_j,\hat{Z}_j)$ is normal with mean vector (μ_X,μ_Y,μ_Z) and covariance matrix

$$S_1 = \begin{bmatrix}\Sigma_{XX} & \Sigma_{XY} & T'_1\\ \Sigma_{YX} & \Sigma_{YY} & T'_2\\ T_1 & T_2 & T_3\end{bmatrix},$$

where

$$T_1 = \Sigma_{ZX.Y}\Sigma_{XX.Y}^{-1}\Sigma_{XX} + \Sigma_{ZY.X}\Sigma_{YY.X}^{-1}\Sigma_{YX},$$

$$T_2 = \Sigma_{ZX.Y}\Sigma_{XX.Y}^{-1}\Sigma_{XY} + \Sigma_{ZY.X}\Sigma_{YY.X}^{-1}\Sigma_{YY},$$

and

$$\begin{aligned}T_3 = {}&\Sigma_{ZX.Y}\Sigma_{XX.Y}^{-1}\Sigma_{XX}\Sigma_{XX.Y}^{-1}\Sigma_{XZ.Y} + \Sigma_{ZY.X}\Sigma_{YY.X}^{-1}\Sigma_{YY}\Sigma_{YY.X}^{-1}\Sigma_{YZ.X}\\ &+ \Sigma_{ZX.Y}\Sigma_{XX.Y}^{-1}\Sigma_{XY}\Sigma_{YY.X}^{-1}\Sigma_{YZ.X} + \Sigma_{ZY.X}\Sigma_{YY.X}^{-1}\Sigma_{YX}\Sigma_{XX.Y}^{-1}\Sigma_{XZ.Y}.\end{aligned}$$

This is a singular distribution, of course, since $\hat{Z}_j$ is a linear function of X_j and Y_j.

Similarly, the joint distribution of $(X_i,\hat{Y}_i,Z_i)$ is normal with mean vector (μ_X,μ_Y,μ_Z) and covariance matrix

$$S_2 = \begin{bmatrix}\Sigma_{XX} & T'_4 & \Sigma_{XZ}\\ T_4 & T_6 & T'_5\\ \Sigma_{ZX} & T_5 & \Sigma_{ZZ}\end{bmatrix},$$

where

$$T_4 = \Sigma_{YX.Z}\Sigma^{-1}_{XX.Z}\Sigma_{XX} + \Sigma_{YZ.X}\Sigma^{-1}_{YY.X}\Sigma_{YY},$$
$$T_5 = \Sigma_{YX.Z}\Sigma^{-1}_{XX.Z}\Sigma_{XZ} + \Sigma_{YZ.X}\Sigma^{-1}_{ZZ.X}\Sigma_{ZZ},$$

and

$$T_6 = \Sigma_{YX.Z}\Sigma^{-1}_{XX.Z}\Sigma_{XX}\Sigma^{-1}_{XX.Z}\Sigma_{XY.Z} + \Sigma_{YZ.X}\Sigma^{-1}_{ZZ.X}\Sigma_{ZZ}\Sigma^{-1}_{ZZ.X}\Sigma_{ZY.X}$$
$$+ \Sigma_{YX.Z}\Sigma^{-1}_{XX.Y}\Sigma_{XZ}\Sigma^{-1}_{ZZ.X}\Sigma_{ZY.X} + \Sigma_{YZ.X}\Sigma^{-1}_{ZZ.X}\Sigma_{ZX}\Sigma^{-1}_{XX.Z}\Sigma_{XY.Z},$$

which again is a singular distribution. Now a natural impulse is to pool these two samples $w_j = (x_j, y_j, \hat{z}_j)$, $(j = 1, \ldots, m)$ and $v_i = (x_i, \hat{y}_i, z_i)$, $(i = 1, \ldots, n)$. However the covariance matrices S_1 and S_2 are not the same, and all such data would lie on two hyperplanes in (X,Y,Z) space. Another impulse is to match the data. Suppose now that $m = n$, so that simple matching has some hope of making sense.

Observe that $w_j - v_i$ has a normal distribution with mean of zero and covariance matrix $S_1 + S_2$, which is nonsingular.

Hence, using the Mahalanobis distance, we may define the distance from w_j to v_i to be d_{ij}, where

$$d_{ij} = (w_j - v_i)' \, (S_1 + S_2)^{-1} \, (w_j - v_i).$$

Thus a match would minimize

$$C' = \sum_{i,j} d_{ij} a_{ij}$$

over choices of a_{ij} subject to the conditions

$$\sum_i a_{ij} = 1,$$

$$\sum_j a_{ij} = 1,$$

and

$$a_{ij} = 0 \text{ or } 1,$$

which again is a linear assignment problem. In the case in which the observations have weights, we relax the condition $n = m$ and suppose v_i has weight q_i $(i = 1, \ldots, n)$ and w_j has weight y_j $(j = 1, \ldots, m)$. The condition $n = m$ is replaced by the condition $\Sigma q_i = \Sigma y_j$. Then the national generalization is to minimize

$$\sum_{i,j} d_{ij} a_{ij}$$

over choices of a_{ij} subject to the conditions

$$\sum_i a_{ij} = y_j,$$

$$\sum_j a_{ij} = q_i,$$

and

$$a_{ij} \geqq 0,$$

which is a transportation problem.

An interesting alternative to the matrix $S_1 + S_2$ to use in the Mahalanobis distance is the matrix

$$\begin{bmatrix} \Sigma_{XX} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

This alternative avoids "bias" that might be introduced by paired Y_j and Z_i, at the cost of not using some of the available information. I regard the relative benefits of these two methods as an open question.

Once the merging is complete, suppose—with slight abuse of notation—that w_j and v_i have been matched. Then it might be natural to take (x_j, y_j, z_i) and (x_i, z_j, y_i) as simulations of the underlying distributions.

Now the expected taxes can be computed. Again I stress that this is conditional on a value of Σ_{YZ}. Many such matchings and averagings should be done, to explore the sensitivity of the results to Σ_{YZ}.

Another aspect of this problem that is not well understood is the relation of matching to the prior reduction of the files (Turner & Gilliam, 1975). Perhaps the two processes can be combined into one, or mutually rationalized.

Why Match?

At first, matching seems to be a peculiar way to treat data. If Σ_{YZ} were known in either framework, the complete joint distribution of the data would be consistently estimated, and any devised probabilities or expectations could in principle be calculated from that estimated jointly normal distribution or, if necessary, simulated on a computer. This approach is less than satisfactory because the variables are in truth not normally distributed. Hence we use the matched sample as if it were a sample from the true distribution and estimate, for instance, the expected value of some tax variable as if by simulation. The normality assumption is used to derive the matching methodology but need not be relied on for the rest of the estimation.

The soundness of this approach is very difficult to assess, and that question will not be settled in this paper. It is clear that a matched sample cannot be treated uncritically as if it were a joint sample that had never been split nor had missing values. Thus the question is not the quality of the match itself, but rather the correct use and interpretation of statistics derived from the matched sample. Our understanding of this question is in its infancy.

References

Alter, Horst E. "Creation of a Synthetic 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 (1974), pp. 373–394.

Anderson, Theodore W. *Introduction to Multivariate Statistical Analysis.* New York: John Wiley and Sons, 1958.

Budd, Edward C. "Comments." *Annals of Economic and Social Measurement* 1 (1972), pp. 349–354.

DeGroot, Morris H., and Prem Goel. "Estimation of the Correlation Coefficient from a Broken Random Sample" (Technical Report No. 105). Pittsburgh: Carnegie-Mellon University, Department of Statistics, 1975. (Mimeo)

DeGroot, Morris H., and Prem Goel. "The Matching Problem for Multivariate Normal Data." *Sankyā*, 38 (1976) (Series B, Part 1), pp. 14–28.

DuBois, N. S. D'Andrea. "A Solution to The Problem of Linking Multivariate Documents." *Journal of the American Statistical Association* 69 (1969), pp. 163–174.

Felligi, Ivan P., and Alan B. Sunter. "A Theory for Record Linkage." *Journal of the American Statistical Association* 64 (1969), pp. 1183–1210.

Kadane, Joseph B. "The Role of Identification in Bayesian Theory." In S. E. Fienberg and A. Zellner, *Studies in Bayesian Econometrics and Statistics.* Amsterdam: North Holland Publishing Co., 1975, pp. 175–191.

Kadane, Joseph B., James M. Dickey, Robert L. Winkler, Wayne S. Smith, and Steven C. Peters. "Interactive Elicitation of Opinion for a Normal Linear Model." Pittsburgh: Carnegie-Mellon University, June 8, 1977. (Unpublished fifth draft)

National Bureau of Standards. *Tables of the Bivariate Normal Distribution Function and Related Functions* (Applied Mathematics Series, No. 50), 1959.

Newcombe, Howard B., and James M. Kennedy. "Record Linkage: Making Maximum Use of the Discriminating Power of Identifying Information." *Communications of the Association for Computing Machinery* 5 (1962), pp. 563–566.

Okner, Benjamin. "Constructing a New Data Base from Existing Microdata Sets: the 1966 Merge File." *Annals of Economic and Social Measurement* 1 (1972), pp. 325–342. (a)

Okner, Benjamin. "Reply and Comments." *Annals of Economic and Social Measurement* 1 (1972), pp. 359–362. (b)

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

Peck, Jon K. "Comments." *Annals of Economic and Social Measurement* 1 (1972), pp. 347–348.

Rao, C. Radhakrishna. *Linear Statistical Inference and Its Applications* (1st ed.). New York: John Wiley and Sons, 1965.

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

Sims, Christopher. "Comments." *Annals of Economic and Social Measurement* 1 (1972), pp. 343–345. (a)

Sims, Christopher A. "Rejoinder." *Annals of Economic and Social Measurement* 1 (1972), pp. 355–357. (b)

Sims, Christopher A. "Comment." *Annals of Economic and Social Measurement* 3 (1974), pp. 395–397.

Tepping, Benjamin J. "A Model for Optimum Linkage of Records." *Journal of the American Statistical Association* 63 (1968), pp. 1321–1332.

Turner, J. Scott, and Gary B. Gilliam. "A Network Model to Reduce the Size of Microdata Files." Paper presented to ORSA Conference, Las Vegas, 1975. (Mimeo)

COMMENT

Christopher A. Sims, University of Minnesota

Kadane's paper presents useful thoughts on the problem of matching heavily overlapping samples when the objective is to obtain exact matches and the resulting sample is not synthetic. I will make no further comment on that first section of the paper. When the paper turns to the problem of creating synthetic data files, however, I find it less convincing. In my opinion, the procedure suggested in the paper is likely not to be an improvement over the "equivalence class" procedures used by Okner and others, whom Kadane cites. My own view is that the purposes for which synthetic data files are currently being created would be better and more cheaply served without actual creation of synthetic files. While I have suggested as much in print before, I will try to be more specific here about a practical alternative to matching, before going on to criticize the method Kadane proposes and to suggest some sensitivity tests that might be done to check existing synthetic files.

My impression is that the main use of synthetic files arises as follows. We have, for example, a proposed modification of the income tax law. A vector of variables V determines the taxes an individual owes under the new version of the tax law according to a (possibly very complicated) function $g(V)$. If we knew V_i for every individual i in the United States and had a lot of computer time, we could compute the sum of $g(V_i)$ over all individuals and compute total income tax revenue under the new law. This sum can be written as

$$N\int g(V)\,dF(V), \tag{1}$$

where $F(V)$ is the population cumulative distribution function (*cdf*) of V, and N is the total population. If, instead, we had a stratified random sample from the population, we might estimate this integral as a weighted sum of $g(V_i)$ over our sample, with the weights determined by our sampling scheme. This weighted sum can be written as

$$N\int g(V)\,dF_s(V),$$

where $F_s(V)$ is what is known as the "sample *cdf*." In effect, we use the sample *cdf* as an estimator of the population *cdf*. The

sample *cdf* is a respectable estimator of the population *cdf*, and there are formal proofs that it has various good properties under certain circumstances.[1]

An Alternative to Matching

The desire to create a synthetic sample arises when V has three components, $V=(X,Y,Z)$, and we have two separate random samples from the population, with only X appearing in both samples, while Y appears in the first only and Z appears in the second only. To use an estimator of F that we can handle in the same way as F_s we must somehow make one sample out of the two samples. Since the two samples contain no information about the joint distribution of Y and Z conditional on X, it has been usual practice to form this synthetic sample in a way that will give good results, on the assumption that Y and Z are independent of each other conditional on X. But matching, on these assumptions, is not necessary.

With the conditional independence assumption, we can write

$$dF(V)=dF^{XY}(V)\,dF^{XZ}(V)/dF^{X}(V), \qquad (2)$$

where F^{XY} is the marginal *cdf* of X and Y (and hence $dF^{XY}(V)$ does not depend on Z), and the other terms on the right-hand side are analogously defined marginal *cdf*'s. The two separate samples are adequate to estimate all the terms on the right-hand side of equation (2) by any of a number of methods. One method that seems natural in this problem is to construct a histogram.[2] One forms a grid in V space and estimates a joint density function by counting the number (or weighted number) of sample points in each cell. Let j index X-categories, k index Y-categories, and m index Z-categories. Let n_{jkm} be the sum of weights of sample points in the jkm'th cell, and use dot notation according to $n_{jk\cdot}=\sum_m n_{jkm}$ to denote marginal distributions. Finally, let V_{jkm} refer to the value of V at the center of cell jkm. Now the estimator for the integral in expression (1) that I am suggesting is

[1] It is not likely to be so good, however, if V has an unbounded range or if g gets very large and varies a lot in regions where df is small.

[2] Why not use sample *cdf*'s for the two subsamples as estimates of the components on the right of equation (2)? Because, since the sample *cdf's* put zero probability on values of X not observed in the sample, they leave the conditional distribution of, say, Y given X undefined except at values of X observed in the first sample. Another way to make the same point: the sample *cdf*'s do not define dF_s^{XY}/dF_s^{X} except at values of X found in the first sample.

$$\sum_{j,k,m} g(V_{jkm}) n_{jk\cdot} n_{j\cdot m} / n_{j} \ldots \quad (3)$$

Here $n_{jk\cdot}$ would be formed from the first sample, $n_{j\cdot m}$ from the second, and $n_{j\cdot\cdot}$ from the two together. It is assumed that observations are weighted to sum to N.

In practice the choice of cell sizes is important. They should be small enough that g varies little within cells, yet large enough so that the n's do not become very small in any cell. These two requirements may be incompatible for some g's, in which case the samples do not provide adequate information, even on the independence assumption, to evaluate expression (1). Also, the two requirements may balance out in different ways for different g's. For example, one might want narrow cells in the income dimension and wide cells in the "value of automobiles owned" dimension when evaluating the effects of a change in income tax law, but want exactly the opposite when evaluating the effects of an automobile excise tax. One could either (1) work with a single histogram and choose a fairly fine grid for a wide variety of applications, or (2) maintain the whole of the two samples in storage and provide a standard routine for forming histograms with user-supplied grids, depending on the application.

In most applications of these procedures, a reasonable and computationally efficient method for handling empty cells would be essential. If there were, for instance, 10 variables in X, each classified into deciles, there would be 10^{10}, or 10 billion, cells. If one indexed cells naively and stored information for each one, the computational costs would be very high. Since sample size is much less than 10 billion, the natural procedure is to index cells by observations they contain or are near to. In this way, the number of terms in the sum (3) would be kept to a number similar to sample size.

Since $n_{jk\cdot}$, $n_{j\cdot m}$ are independent of one another and have a binomial distribution, and since $n_{j\cdot\cdot}$ can be expressed in terms of $n_{jk\cdot}$ and $n_{j\cdot m}$, it would be possible to form an estimated covariance matrix for the terms $n_{jk\cdot} n_{j\cdot m} / n_{j\cdot\cdot}$ appearing in sum (3), and hence to provide a standard error for sum (3) that would warn when results are unreliable.

The advantages of a procedure like the one outlined here over matching to create a synthetic data set are, in summary: (a) the estimate it generates of the joint distribution is more economical of storage space; (b) the procedure lends itself to computation of standard errors indicating the reliability of computations based on it; (c) the procedure can be connected to the large statistical literature on estimating density functions and multidimensional

contingency tables,[3] and (d) it is likely to give more accurate results than matching.

Evaluating Matching Procedures

A critical difference between exact matching and synthetic file formation is that the local sparseness or denseness of the samples plays fundamentally different roles in the two problems. For exact matching, a dense region of the X-space is one in which we have many observations whose X values differ by less than the standard deviation of measurement error in X. Such regions present special problems because within them exact matching becomes difficult; in other (sparse) regions exact matching may be easy. For synthetic file formation, on the other hand, a dense region of X-space is one within which we expect that the distributions of Y and Z given X change little (both Y and Z are locally independent of X); it is, at the same time, a region within which we have many observations. Such regions are no problem at all, since within them any arbitrary matching procedure will produce results that do not distort the joint distribution of X, Y, and Z (except via the conditional independence assumption). In sparse regions we are almost bound to distort the joint distribution in synthetic file formation, unless we go beyond "matching" to more elaborate methods of generating synthetic observations.

The whole idea of actually minimizing a sum of distances over all matches seems computationally profligate in forming any very large synthetic data sets. Instead, dense regions should be treated by simply ensuring that all matches meet some minimum criterion, beyond which improvements in match will make little difference. This is the intuition underlying the "equivalence class" methodology. On the other hand, it is essential to identify sparse regions and either not match there at all or flag as unreliable the synthetic observations from such regions. This is the role of distance measures in synthetic matching. For these purposes, a good distance measure will be one that measures whether X's for the paired observations differ enough to make the conditional distribution of Y or Z given X differ across the observations. As a simple first approximation one might (assuming all data were scaled to have unit variance) measure the distance between two

[3] See, e.g., Bishop, Fienberg, and Holland (1975) and Moore and Yackel (1977) and references cited therein. Although these references do not deal explicitly with such highly multivariate problems as are usual in the matching literature, they do contain useful insights nonetheless.

observations as the sum of squared differences in the conditional means of Y and Z given X between the two observations, using a local normality assumption so that this becomes a quadratic function of the X's.

The distance measure Kadane proposes in the second part of his paper, because it uses information on Y and Z, is likely to produce systematic distortion in the estimated *cdf*. The implicit motivation for Kadane's distance measure is to match observations that are as "similar" as possible, where similarity in Y and Z, as well as similarity in X matters. This is an appropriate idea when one is doing exact matching, and the distance measure Kadane suggests in the first part of the paper is thus reasonable in that application, even though it, too, uses information on Y and Z. In forming synthetic matches, however, suppose we encounter an observation in sample 1 with a very unusual value of Y given the associated X. If Y depends strongly on X, we may be quite sure that an observation in sample 2 with nearly the same value of X as this sample 1 observation is unlikely to have a similar value of Y. Thus, using Kadane's distance measure we will prefer to match with an observation having a different value of X, giving up some X-similarity in hopes of improving Y-similarity. It is easy to see that this kind of tradeoff will result in a synthetic sample in which the conditional variance of Y given X will be biased downward if the synthetic sample uses sample 2 X values. If sample 1 X values are used, it is the conditional variance of Z given X that is biased downward.

It is true that if one drops the assumption of conditional indepedence of Y and Z, it is natural to use information on Y and Z in matching. This cannot, however, justify use of a distance measure based on reasoning appropriate to the exact-match case. If information on Y and Z is used in matching, it ought to be done on the basis of an explicit assumption about the nature of conditional dependence between Y and Z, and using a method that would make the distance measure unrelated to Y and Z under the conditional independence assumption.

It may be worthwhile to suggest ways to do sensitivity analysis of existing matching procedure without redoing the matching on a different conceptual base. To check whether the quality of match in sparse regions is making much difference to results, one could take an existing synthetic data set in which the X's from both samples (X_1 and X_2) were available for each observation and correct the matches for conditional means. For the purposes of this kind of test, one might use the assumption of joint normality, so that $E[Z|X]=a+Xb$. Assuming that X_1 is the X going in

to the synthetic observation in each case, with X_2 discarded, the original matched synthetic data set contains observations of the form (X_1,Y,Z). The proposal here is to form another data set with observations of the form $[X_1,Y,Z+(X_1-X_2)b]$. The parameter vector b, of course, would be calculated by the usual least-squares formulas from the sample variance-covariance matrix. The new sample might in most applications give the same results as the original synthetic sample, in which case the original matching procedure does not suffer much from the kinds of bias that have concerned me in this set of remarks. On the other hand, if results do differ across the two samples, systematic accounting for the effects of bad matches in sparse regions is esential.

Probably even more important would be sensitivity analyses for the conditional independence assumption. To check this assumption well, one would need some extraneous source of information on conditional dependence between Y and Z given X. In a sample with many variables it is likely to be difficult to obtain such information by introspection, since people are not likely to have reliable intuitions about interrelations among large numbers of variables. Sometimes it may happen that not all the variables actually in both samples are used in matching, either because of some missing observations or because it is felt that using all the X's makes matching too complicated. In that case, a sensitivity test for conditional independence given the X's actually used is possible by doing the match also with a larger set of X's and comparing the results with the smaller set.

In closing, let me repeat the point that in my opinion there is nothing that a matched synthetic data set can do that could not be done cheaper and better by the two original data sets sitting behind a routine for computing histograms, or, instead, a fine-grained histogram sitting behind a routine for aggregating it. In fact, once started down the road of considering this problem explicitly as one of joint-density-function estimation, specialists in this field are likely to improve substantially on the methods I outlined at the beginning of these remarks.

References

Bishop, Yvonne M. M., Stephen E. Fienberg, and Paul W. Holland. *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, Mass.: The M.I.T. Press, 1975.

Moore, David S., and James W. Yackel. "Consistency Properties of Nearest Neighbor Density Estimators." *Annals of Statistics* 5 (1977), pp. 143–154.

REPLY

Joseph B. Kadane

I welcome the constructive nature of Sims' "alternative to matching." It deserves to be taken seriously, and its properties should be investigated. However, Sims assumes the conditional independence hypothesis, which the second part of my paper eschews. Consequently, Sims' "alternative" is not an alternative to matching in the context in which I proposed that matching. Of the advantages advanced by Sims for his method, I agree with (b) and (c), but regard (a) and (d) as undemonstrated speculation.

It is easy to see that under the conditional independence assumption $\Sigma_{YZ}=\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XZ}$, so that all the parameters of the model may be estimated without resort to matching. Thus Sims' view that "nothing that a matched synthetic data set can do could not be done cheaper and better by the two original data sets . . ." is unsurprising in this context.

However, when conditional independence is not assumed, new problems develop. Sims believes that it would be difficult to obtain information on Σ_{YZ} by introspection, "since people are not likely to have reliable intuitions about interrelations among large numbers of variables." To the contrary, all that is needed to give an opinion on Σ_{YZ} is introspections on pairs of variables. Constraints must be imposed so that the information obtained is consistent with what is known about Σ_{YY} and Σ_{ZZ} from the sample. The methods to do this sort of thing are new but not impossible.[1]

With a single value for Σ_{YZ}, or several values with probability values attached, I believe that something like matching may be natural. In fact, Sims agrees that in this case "it is natural to use information on Y and Z in matching." However, I must agree with him, that should the introspected value of Σ_{YZ} satisfy the conditional independence hypothesis, it would be a good property for a matching measure not to put any weight on the Y and Z components. I do not know whether this is true of S_1+S_2, although it is of course true of the alternative measure

$$\begin{bmatrix} \Sigma_{XX}^{-1} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

[1] Joseph B. Kadane, James M. Dickey, Robert L. Winkler, Wayne S. Smith, and Steven S. Peters, "Interactive Elicitation of Opinion for a Normal Linear Model." Pittsburgh: Carnegie-Mellon University, June 8, 1977. (Unpublished fifth draft)

This is an area I intend to investigate further, and I thank Sims for raising the question.

The question of bias caused by different measures is also an interesting one deserving further study. If the metric

$$\begin{bmatrix} \Sigma_{XX}^{-1} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

is used, no bias is present but it is possible that information is wasted. If $S_1 + S_2$ is used, bias may be introduced. As I remark in my paper, I regard the relative merits of these two measures as an open question; after reading Sims' comment, I still do. Of course, bias in the conditional variance given X is not of great interest, as better estimates of that are available immediately without matching. Whether bias results in the estimation of the integral of functions g is something about which very little is known. I suspect that the systematic overestimation of tax revenue, because the ability of tax payers to adjust their behavior to minimize their taxes is not taken into account, is a far more serious source of bias.