DISTRIBUTIONAL CHARACTERISTICS OF A MERGED MICRODATA FILE

Kevin Hollenbeck, Urban Systems Research & Engineering
Pat Doyle, Mathematica Policy Research

In a study of the size distribution of income, Budd, Radner, and Hinrichs (1973) combined the observations from the March 1965 Current Population Survey and a sample of 1964 federal personal income tax returns to derive a more complete universe and definition of income. Similarly, Okner (1972) linked the 1967 Survey of Economic Opportunity to a sample of 1966 tax returns and Ruggles and Ruggles (1974) merged the 1970 Public Use Sample of the Census of Population and Housing and the Social Security Longitudinal Employer-Employee Data file. These studies all relied on a microdata file merging technique called attribute matching. The list of applications of attribute (or statistical) matching has grown rapidly as researchers have attempted to enrich data samples for analysis purposes.[1] But the question of how meaningful are the synthetic distributions created in these exercises has not been seriously addressed.

The Matching Problem

The general problem for which attribute matching has been used may be stated as follows:

(X, Y, Z) is a $(k_1+k_2+k_3)$-tuple distributed in the population U as $f(X, Y, Z)$

S_1 is a sample of X, Y of size m

S_2 is a sample of X, Z of size n.

The question then is whether it is possible to associate the two samples S_1 and S_2 in such a way as to allow inferences about f, the joint distribution of X, Y, and Z.

For example, one household sample survey may gather data about household characteristics such as education of head and family income while another body of data will contain information about family income and asset holdings. An economist may wish to combine the samples to analyze the relationships between age of head and asset holdings.[2] More likely, the economist will be interested in the age of head/asset distribution conditional on income. Does a 30 year old with an income of $30,000 have a different level of assets than a 60 year old?

Although the algorithms for achieving it differ, the basic solution to the matching problem has been an association operator A which chooses a data point in S_2 for every point in S_1 and appends the Z data to the S_1 observation. The assumption is that the observed moments of the S_1 and S_2 sample distributions unbiasedly estimate the moments of the true distribution. Notationally, this is as follows:

$$(1) \quad \{S_1\ (X_i, Y_i)\} \ \overset{m}{\underset{i=1}{A}}\ \{S_2\ (X_j, Z_j)\} = \hat{f}(X,Y,Z) \quad ,$$

where $\{S_1\ (X_i, Y_i)\}$, $\{S_2\ (X_j, Z_j)\}$ are the sets of data points in S_1 and S_2.

A is the matching association operator,

$\hat{f}$ is the derived synthetic sample, and

$g(X,Y)$, $h(X,Z)$ are the observed density functions in S_1 and S_2, respectively.

It is assumed that for the means,

$$(2) \quad \sum_y \sum_z x\, g(x,y) = \sum_z \sum_x x h(x,z) = \int_z \int_y \int_x x f(x,y,z)\,dxdydz$$

$$(3) \quad \sum_y \sum_x y\, g(x,y) = \int_z \int_y \int_x y f(x,y,z)\,dxdydz$$

$$(4) \quad \sum_z \sum_x z\, h(x,z) = \int_z \int_y \int_x z f(x,y,z)\,dxdydz$$

Similar assumptions hold for the other moments of the sample distributions assuming the moments of f are finite.

Constrained and Unconstrained Matching

There are essentially two types of asscia-tion operators -- a constrained match and an unconstrained match. Let X_m be a subset of the X variables and M_{ij}; $i \in S_1$, $j \in S_2$, be a metric measuring the weighted distance between X_{m_i} and X_{m_j}, data points in S_1 and S_2. An unconstrained match chooses the set of X_{m_j} which minimizes M_{ij} for all i, where a point in S_2 may be chosen any number of times. The resultant file is of size m.

A constrained match results in a sample size much larger than m or n by introducing a set of variables W_{ij} which are the weights assigned to matching observation i in S_1 to observation j in S_2. It then minimizes $\sum_i \sum_j M_{ij} \cdot a_{ij}$, where $a_{ij} = 1$ if i, j match or 0 otherwise, subject to the constraints that $\sum_j W_{ij} = W_i$, $i=1,\ldots,j$; $\sum_i W_{ij} = W_j$; $j=1, \ldots, n$; and $W_{ij} \geq 0$. The advantage to a constrained match is that it preserves the moments of the S_1 and S_2 samples.[3]

However, it can be shown that the minimum of the objective function of a constrained match exceeds the minimum in an unconstrained match.[4] Thus in a sense, the quality of the match in X_m-space is not as high as in the unconstrained case.

Properties of the Synthetic Estimators of the True Distribution

Unconditional Means and Variances. In a well-formulated constrained match, the unconditional means of all the sample X, Y, and Z variables are unbiased. In an unconstrained match of similar quality, there is a potential bias in the means of all the Z variables. Call b_{Z_i} the bias associated with the mean of Z_i. Then

(5) $E\,Z_i = Z_i + b_{Z_i} \quad ; i=1, \ldots, k_3$

The b_{Z_i} depend on the statistical properties of S_1 and S_2 and on the matching algorithm. Presumably this bias can be minimized by altering the algorithm.

The covariances between the individual elements of X and Y replicate those in S_1, so they are unbiased. The sample covariances between the elements of Z are unbiased in the constrained match; while in an unconstrained case, they may contain bias because of the distortion in the Z distribution caused by the match.

Of interest are the unconditional covariances between X_i and Z_i and Y_i and Z_i. Both the constrained and unconstrained match produce biased covariance estimators except under extreme independence assumptions. Because they come from different samples, there is a discrepancy between the X_i from S_1 and it's matched counterpart X_i and S_2.

(6) $X_{ik} = X_i^m + \varepsilon_k$, where

X_{ik} is the value of X_i for the k-th observation in $\hat{f}$

X_i^m is the value of X_i from S_2 which is associated with X_{ik}

ε_k is the discrepancy for k-th observation .

Because of these discrepancies, the sample covariance will not unbiasedly estimate the true population covariances in the constrained match. For that reason plus the bias in the Z_i distribution, the covariance of the sample resulting from the unconstrained match will be biased. The latter bias may be less than in the constrained match, however, because as the earlier discussion indicated, the ε_k's are lower.[5]

The pairwise covariances between individual Y_i and Z_i are more difficult to assess. If the Y_i is correlated with any set of X_i, then the discrepancies of equation (6) will come into play and there will be bias in the estimated covariances in a constrained match. As mentioned earlier, the bias in the unconstrained case may be lower even though it stems from the X-discrepancies and the distortion of the Z_i-distribution.

Conditional Means and Variances. The primary purpose of matching is to derive the conditional distribution of Y, Z, on X. The conditional covariances between Y_i, Z_i on a given X_i depend on the conditional means of Y_i on X_i and of Z_i on X_i. The former come from S_1 and so are unbiased. Since Z_i and X_i come from different samples, there is the X_i-discrepancy given in equation (6) which biases $E\,Z_i|X_i$ in both the constrained and unconstrained matches.

Conclusion

The technique of attribute matching has been applied when there are two independent samples from a population distribution of random variables X,Y,Z, one of which observes X, Y and the other of which observes X, Z. Two types of association operators have been applied to merge samples to provide synthetic samples from which to derive inferences about the distribution of X,Y,Z. They may be referred to as a constrained and an unconstrained match.

The distributional estimators derived from the merged sample may suffer from a bias because the individual X variables do not exactly match (call this X-bias) and if the unconstrained match is applied, there may be distortions in the distribution of the Z variables (Z-bias). The biases are functions of the statistical properties of the two samples, sample sizes, and the matching algorithms. The X-bias in an unconstrained match will not be as severe as in a constrained match, so it must be determined empirically which algorithm is better.

Because the technique has assumed a high degree of importance in policy analysis and because of its potential analytic usefulness, a set of Monte Carlo trials of each matching technique is suggested to determine the relationships between the properties of the samples and the X-bias and Z-bias.

APPENDIX

Proof that Constrained Match Minimum Exceeds Unconstrained Match Minimum

Let S_1, S_2 be samples of size m, n of data from the population distribution f(X,Y,Z), where S_1 contains observations on X, Y, and S_2 independent observations on X, Z. S_3 is a synthetic sample of X,Y,Z created through an unconstrained match and is thus of size m. S_4 is a synthetic sample of X,Y,Z created through a constrained match and is of size p, where $\max(m,n) \leq p \leq m \cdot n$. Let M_{ij} be a measure of distance between the common characteristics of observation i in S_1 and j in S_2.

An unconstrained match maps each point i in S_1 to a single point j in S_2. Call M_i the resulting distance. In a constrained match, each point in S_1 may be matched to several points in S_2 with each new observation assigned a sampling weight W_{ij}. Let W_i be the sampling weight for observation i in S_1.

CLAIM:

$$\sum_{i=1}^{m} W_i M_i \leq \sum_{i=1}^{m} \sum_{j=1}^{n} W_{ij} M_{ij} a_{ij}, \quad \text{where}$$

$$a_{ij} = \begin{cases} 1 & \text{if i and j match} \\ 0 & \text{otherwise} \end{cases}$$

PROOF:

$$\sum_{i=1}^{m} \sum_{j=1}^{n} W_{ij} M_{ij} a_{ij} = \sum_j W_{1j} M_{1j} a_{1j} + \sum_j W_{2j} M_{2j} a_{2j} + \ldots$$

$$= W_1 \sum_j \frac{W_{1j}}{W_1} M_{1j} a_{1j} + W_2 \sum_j \frac{W_{2j}}{W_2} M_{2j} a_{2j} + \ldots$$

$$\geq W_1 \sum_j \frac{W_{ij}}{W_1} M_1 a_{1j} + W_2 \sum_j \frac{W_{2j}}{W_2} M_2 a_{2j} + \ldots$$

since by definition $M_i \leq M_{ij}$, all j.

But by constraint, $\sum_j W_{ij} a_{ij} = W_i$, so the last expression equals:

$$= W_1 M_1 + W_2 M_2 + \ldots$$

$$= \sum_j W_i M_i . \qquad \text{QED}$$

REFERENCES

Alter, Horst. "Creation of a Synthetic Data Set by Linking Records of the Canadian Survey of Consumer Finances with the Family Expenditure Survey." Annals of Economic and Social Measurement. April 1974, pp. 373-94.

Armington, Catherine and Odle, Marjorie. "Research on Microdata Files Based on Field Surveys and Tax Returns: Creating the MERGE-70 File--Data Folding and Linking." Working Paper Washington ,D.C.: The Brookings Institution, 1975.

Barr, Richard S. and Turner, J. Scott. "A New Linear Programming Approach to Microdata File Merging." 1978 Compendium of Tax Research. Sponsored by the Office of Tax Analysis, U.S. Department of Treasury. Washington, D.C.: U.S. Government Printing Office, 1978. pp. 131-149; "Reply", pp. 152-155.

Budd, Edward C.; Radner, Daniel B., and Hinrichs, John C. "Size Distribution of Family Personal Income: Methodology and Estimates for 1964." Staff Paper No. 1. Bureau of Economic Analysis, U.S. Department of Commerce, 1973.

Hollenbeck, Kevin M. "A Design for Creating a New Data Base Using the Survey of Income and Education and Annual Housing Survey." Discussion paper. Washington, D.C.: Mathematica Policy Research, Inc., September 1978.

King, Jill A. The Distributional Impact of Energy Policies: Development and Application of the Phase I Comprehensive Human Resources Data System. Final Report submitted to the Federal Energy Administration. Washington, D.C.: Mathematica Policy Research, Inc., June 1977.

Okner, Benjamin. "Constructing a New Data Base from Existing Microdata Sets: the 1966 Merge File." Annals of Economic and Social Measurement. July 1972, pp. 325-42.

Ruggles, Nancy and Richard. "A Strategy for Merging and Matching Microdata Sets." Annals of Economic and Social Measurement. April 1974 pp. 353-72.

Springs, Ricardo and Beebout, Harold. The 1973 Merged SPACE/AFDC File: A Statistical Match of Data from the 1970 Decennial Census and 1973 AFDC Survey. Final Report submitted to the Social and Rehabilitation Service, U.S. Department of Health, Education, and Welfare. Washington, D.C.: Mathematica Policy Research, Inc., March 1976.

FOOTNOTES

[1] See Alter (1974), Armington and Odle (1975), Springs and Beebout (1976), King (1975), Barr and Turner (1978), and Hollenbeck (1978).

[2] Typically, however, k_1, k_2, and k_3 larger than one, or in other words, there are multiple joint characteristics in the samples and also several Y and Z characteristics.

[3] See Barr and Turner (1978), pp. 153-155.

[4] Proof is in the Appendix.

[5] The reader should note that unbiased estimates of $COV(X_i, Z_i)$ can be derived from S_2.