SOME PROBLEMS IN AND APPROACHES TOWARD HANDLING MISSING DATA IN LONGITUDINAL FILES[1]

Harvey A. Goldstein, University of Pennsylvania

Several recent papers have suggested that longitudinal data can provide the basis for the development of powerful and efficient analyses of process-oriented social problems.[2] In this regard, the development of longitudinal data files for all households and dwelling units in a single American metropolitan area (Wichita - Sedgwick County, Kansas) has presented the opportunity to investigate several aspects of this argument in terms of classes of questions relating to residential mobility and neighborhood change which have been precluded by the use of the kinds of cross-sectional data currently available from the federal decennial census and other similar sources. There are, however, several problems associated with the development and utilization of longitudinal files, which are not in the use of cross-sectional data. Thus, while the latter engenders no problems concerning the inter-temporal consistency of each record, for longitudinal data to be useful such consistency is of paramount importance. There are at least two general sources of inter-temporal inconsistency: (a) measurement or technical erros and, (b) missing items. The latter can, in turn, be caused by (i) respondent non-response, (ii) the addition of variables not previously in the enumeration, or (iii) treating an error, once discovered, as a missing item to be estimated. Each of the two general classes of inconsistency implies a particular set of procedures for 'longitudinal editing'. The former issue has been treated in an earlier paper;[3] the remainder of this discussion will outline a typology of missing items and methods for treating each type of problem within longitudinal files.

I. Consistency Constraints on Longitudinal Files

There are at least three types of consistency constraints in longitudinally-structured files which can be identified:

1. Logical constraints. Given a temporally organized file, certain variables have specific logical constraints on the allowable changes between recorded observations. In effect, due to a time-homogeneous progression, many classes of observations must vary unambigously and systematically with time. Examples include:

   (a) Age. Age, once recorded, is determined for all previous and subsequent time periods. Thus, if age is given, missing observations can be supplied with the rule that the value must increase incrementally with time for each consecutive observation period. There is one qualification, however: the rule is subject to variations in the period of observation, e.g., for a given household, the observation is assumed to take place on the same day in a yearly enumeration.

   (b) Sex. This variable can be presumed to remain stable over time.

   (c) Identification Number. In many data files, there are control variables which uniquely identify different households and individuals. For the same individual and household, such an indicator can thus be unambiguous assigned when missing. An exception is the case where, say, a household has undergone structural change (e.g., through dissolution) and new; identification numbers must be assigned to each member of the household.

2. Quasi - logically constrained. Constraints of this type are characterized by similar properties as logical constraints, though there is a possibility of greater temporal variation -- particularly with substantially longer periods between observations. Characteristically, only a narrow range of assignments to missing variables are possible. Examples include:

   (a) Level of education. Completed years of education can not decrease temporally and, furthermore, unless other information is furnished, the value cannot increase by more than 1 for consecutive years of observation. True, if the 'flanking' observations are given, the value of the missing observation can be interpolated as the mean difference between the two. Note, however, that where the difference is small, a unique value cannot always be given.

   (b) Race. The race variable is often troublesome in survey work for several reasons, but where it is recorded as a household (rather than as an individual) variable, as in the Wichita Enumeration, measurement is often judgmental (on the part of the interviewer) and/or ambiguous (if the household is racially mixed). The following guidelines seem to be reasonable: (i) if, there is no reference there is no basis for supplying any value for the missing items (short of inferring

it from other characteristics such as area of residence); (ii) if the race is consistent for all given observations save the missing item, then it is assumed to be the same; and (iii) if race is inconsistent over the given observations and the head of household has not changed, supply the race of the household head if the head had been the interviewee for any of the given observations.

3. Logically unconstrained. These variables tend to exhibit missing items most frequently; and because they must be estimated from correlated variables, they also present the most interesting problems from both statistical and theoretical points of view. Variables of this class include family income, monthly rent/valuation, occupation, and weeks worked per year.

II. Schemes for Estimating Missing Items for Logically - Unconstrained Variables

Missing values for logically and quasi-logically constrained variables can be handled in quite straightforward ways. This is not the case, however, for logically - unconstrained variables. For purposes of explication, we shall use family income to illustrate procedures for estimating missing items in this class; we will assume here that the record contains at least one temporal reference. There are at least three possible approaches to this problem.

The first, which assumes the existence of at least two temporal references, is the use of one or another linear or non-linear interpolation scheme. Since inter-temporal variation or family income tends to be large (particularly over longer time-intervals) an assumption of linearity, though technically simpler, tends to impose unrealistic theoretical constraints. Non-linear interpolation, on the other hand, requires more sophisticated techniques and an interpretive theory for selecting among particular non-linear functions.

When interpolation is combined with correlations with other variables, we can make better use of the information available and thus provide a basis for better estimates of the missing items. For family income, the employment status information on each member of the household, value of house or rental value, the occupation (if given), and possibly completed years of education can be helpful in estimating family income if these correlated variables indicate small or no temporal change. In this case, for example, we could employ simple interpolation procedures to estimate family income. The inverse, however, would not be as helpful: if the employment pattern of the members of the household was not time-homogeneous (in some sense), we would not have a sufficient basis for the relevant inferences.

The technique of stratifying households or individuals or given variables that are known to be correlated with the variable having a missing entry is well-known. For example, the U.S. Bureau of the Census frequently estimates missing data items by assigning the value of the same item of the last-scanned record belonging to the same stratum. However when stratification is used in conjunction with the information available for that particular record at other points in time we are often able to make a better estimate. Thus, we may use the inter-temporal difference in family income (as measured from the flanking observations of the missing item) as one of the stratifying variables; by further stratifying the population by residential location improved estimates are obtained (the assumption here being that households with similar levels of income tend to live in proximity to one another); and so on.

III. Problems in Designing Stratification Schemes

There are several theoretical and technical problems in the design of stratification schemes. The first is deciding upon the number of strata--which is, in turn dependent on requirements concerning the mean number of records per stratum needed to make inferences. The smaller number of strata, of course, the larger the within-group variance; with too many strata, too much weight may be placed on too few records. A second question involves the choice of stratifying variables. Again, this depends upon the particular variable to be estimated, the available information, and the theory one employs.

The problem of deciding where to make cuts (form classes) on each dimension (stratifying variable) presents one of the most crucial problems in designing stratification schemes. Clearly, this cannot be considered independently of the choice of the total number of strata to be formed. Furthermore, there may be strong theoretical reasons for certain a priori classes (e.g., in the case of age of head of household, 65 would be such a cut). But where this is not the case, classes should be selected so as to minimize within-group variance subject to the exogenously supplied minimum number of records per stratum. With some adaptations we have found that the AID program (automatic interaction detection) developed at Ann Arbor, Michigan by Morgan and Sonquist[4] can provide a 'locally optimal' set of classes when provided with initial interval data. Also, for some applications involving small numbers of records the 'smear and sweep analysis' developed by John Gilbert at the Harvard Computing Center and others, has been found to be useful.

There is an obvious problem in determining the appropriate sample for testing the stratification procedures discussed above: households with missing items cannot be included in the sample. Taking a random sample would, for example introduce biases in the estimating procedures. A reasonable way to proceed is based on drawing a stratified sample based from the records with missing items on a particular

variable (e.g., family income) and treating this sample as though the records in it had the missing items; estimation errors could then be calculated in the usual way.

One of the ultimate requirements for maintaining large longitudinal data files in an economical way is to institute automatic editing and estimating procedures. Thus, it would be important to have stratification schemes for use in estimating missing items on most variables, i.e., a system by which the strata definitions can be updated on the basis of changes in the population. This would entail, among other details, the addition of a stratum identification variable for each record and, efficiency, the organization of records by various strata. A general stratification design for estimating missing items for all variables could then be provided rather than a different design for each case of missing data items. Although a single stratification design would involve some trade-offs in this context, such a scheme could be extremely useful for other kinds of analysis and the monitoring and modelling of changes of a number of sub-populations.

IV. Conclusions

The most powerful method for estimating missing items on longitudinal records clearly depends upon using as much of the available information as possible: information from other variables for the particular record, previous observations, and information from other units of observation with similar properties. In this paper, we have suggested that the use of stratification schemes provides a general, efficient method for organizing this information for the estimation of missing items. In addition to the usual theoretical considerations, several design characteristics, such as choice of stratifying variables and choice of classes, were discussed. Alternative designs will also be needed in the development of fully automatic procedures for updating, editing, and correction.

Notes

1. The support of the National Science Foundation, Grant No. GS-39837 and the helpful comments of Stephen Gale are gratefully acknowledged.

2. H.S. Parnes (1972). "Longitudinal Surveys: Prospects and Problems", Monthly Labor Review, February, 11-15.

P. Baltes (1968). "Longitudinal Studies and Cross-Sectional Sequences in the Study of Age and Generation Effects", Human Development, 11, 145-171.

E.S. Dunn, Jr. (1974). Social Information Processing and Statistical Systems - Change and Reform. New York: John Wiley.

W.D. Wall and H.L. Williams (1970). Longitudinal Studies and the Social Sciences. London: Heinemann.

3. Harvey A. Goldstein and Stephen Gale (1975). "Some Problems in and Approaches Toward Editing and Maintaining Longitudinal Data Files", Research on Metropolitan Change and Conflict Resolution, Technical Paper No. 5. Peace Science Department, University of Pennsylvania.

4. James N. Morgan and John A. Songuist (1963). "Problems in the Analysis of Survey Data, and a Proposal", Journal of the American Statistical Association, June, 415-434.

John A. Songuist, Elizabeth Lauh Baker, and James N. Morgan, (1971). Searching for Structure. Ann Arbor, Michigan: Institute for Social Research, The University of Michigan.