CROSS-SECTIONAL IMPUTATION AND LONGITUDINAL EDITING PROCEDURES IN THE SURVEY OF INCOME AND PROGRAM PARTICIPATION

Steven G. Pennell and James M. Lepkowski, Survey Research Center, University of Michigan
Steven G. Pennell, Survey Research Center, University of Michigan, Ann Arbor, MI 48106

KEY WORDS: Imputation, Longitudinal Editing

This paper describes the major cross-sectional imputation and longitudinal editing procedures applied to data collected in the Survey of Income and Program Participation. Cross-sectional imputation procedures in the SIPP are used to compensate for item nonresponse and for two types of person-level noninterviews in otherwise cooperating SIPP households. Longitudinal editing procedures in the SIPP are designed to remove inconsistencies in a sample person's longitudinal record introduced through independent wave imputations of item missing data and person-level noninterviews, to adjudicate occasional disagreements in reported information across waves and to reconcile reported information with Census demographic definitions.

Overview of the SIPP Design

The SIPP was initiated in late 1983 (the start of the 1984 panel) by the U.S. Bureau of the Census with the principal objective to provide policy-makers with more accurate and comprehensive information on income and program participation in government programs than were available through other data sources.

Interviews of panel members by self or proxy reports are conducted every four months for seven or eight consecutive interviews. Original panel members are defined as persons age 15 or older who are living in sampled households on the date of the Wave 1 interview or persons under the age of 15 who become age eligible in subsequent waves. In subsequent waves age eligible persons who join a SIPP sampled household are also interviewed.

The data collection instruments in the SIPP include the Control Card, the Core questionnaire and one or more Topical Modules. The Control Card is the basic record for each sample unit and contains demographic and household composition information, items transcribed from prior wave interviews as well as administrative data. The Core questionnaire contains questions which are repeated at each interview and are asked of each sample person. Topical Modules contain questions which generally are not repeated at each wave and cover special topics not included in the Core questionnaire.

Sequence of Imputations and Longitudinal Edits

The cross-sectional processing begins by first imputing item missing data on the Control Card. Missing items on the Control Card are imputed first because many of the demographic variables located there are used in subsequent imputation steps and need to be nonmissing for all cases. Next, Core questionnaire records are imputed in full from a single donor for two types of person-level noninterviews. Because person-level noninterviews are imputed before donor records are processed for item missing data, imputed noninterview records initially retain the pattern of item missing data on the donor record. Missing items on the Core questionnaire are subsequently imputed for responding sample persons and for noninterviews whose records were previously imputed. The processing of Core questionnaire items is also sequenced so that missing items in earlier steps can be used to impute missing items in later steps. Item missing data on the Core questionnaire are imputed section by section in the following sequence: 1) labor force and recipiency; 2) other cash income; 3) wage and salary and self-employment variables; 4) asset variables; and 5) program participation variables.

Item missing data on Topical Modules are imputed at the same time missing items on the Core questionnaire are imputed. Once the data for each wave in a panel has been processed, selected groups of items are extracted from each wave and longitudinally edited. The process of extracting and editing is performed separately for the following groups of items: 1) demographic and household variables; 2) employment variables; 3) general amount variables; and 4) other variables. As each group of items is edited they are joined together to create the SIPP Longitudinal file.

Goals of Imputation

There are two general goals of imputation, one is statistical and the other is practical. The statistical goal of imputation is to minimize the mean square error of survey estimates. The mean square error has both a variance and a bias component. All imputation procedures increase the variance of estimates but some imputation procedures increase the variance less than others. Imputation can reduce the bias component of

the mean square error to the extent systematic patterns of item nonresponse are identified and correctly modeled. The ability of an imputation scheme to correctly guess the missing values of individual items is of lesser importance; although, the better an imputation scheme is able to does this, the smaller will be the error due to imputation. The statistical goal of imputing missing data in the SIPP is also more general than specific. The SIPP imputation procedures are not designed to address estimation of specific parameters, but rather to provide reasonable estimates for a variety of analytical purposes. No single imputation procedure is likely to be ideal for all analytical purposes.

There are also several practical goals for imputing missing data. Consistency is maintained between the results from different analyses when missing data are imputed because cases with missing data are not necessarily excluded. In the absence of imputation, and in the presence of missing data, different analyses will be based on different subsets of cases depending on the pattern of missing data.

Although the statistical goal of imputation is to reduce the bias component of the mean square error, there is no guarantee that estimates based on imputed data are less biased than estimates based only on nonmissing data. In fact, the bias associated with estimates based on imputed data could be greater depending on the type of imputation used and the parameter being estimated. Imputation also has the distinct disadvantage of creating the impression that the data are complete. All statistical imputation procedures fabricate data which increase the variance of estimates. Because the increase in variance due to imputation is difficult to incorporate into variance estimates, the precision of survey estimates is often overstated.

Type Z Imputation Procedure

The U.S. Bureau of the Census classifies noninterviews at both the household and person level. Person-level noninterviews are defined only in households in which at least one person was interviewed and occur when one or more sample persons, but not all sample persons in the household, refuse to be interviewed or are unavailable and a proxy report is not obtained. The two types of person-level noninterviews imputed in the SIPP are: 1) Type Z noninterviews, which occur when a member of an interviewed household is not interviewed because they are unavailable for an interview or refuse and a proxy interview is not obtained; and 2) Departure noninterviews, which occur when someone who was a member of a SIPP interviewed household sometime during the four-month reference period is no longer a household member on the date of interview. The phrase "Departure noninterview", which is not an official Census term, is used as a convenient way to distinguish between the two types of person-level noninterviews.

A statistical matching procedure referred to as Type Z imputation is used to impute both types of person-level noninterviews. Type Z imputation is based on a hierarchical sorting and merging operation which matches noninterviews with respondents on socioeconomic characteristics available for both. Once a matching donor is identified Core questionnaire values reported by the donor, or provided by a proxy, are assigned to the noninterview record in full, except for identification variables or other variables not relevant for the household in which the noninterview occurred.

The socioeconomic variables which are used to match noninterviews with respondents are either taken from the current wave Control Card, extrapolated forward from previous wave Control Card information or, if missing on the Control Card, imputed for the current wave using an item by item hot-deck procedure. The variables used to match noninterviews with respondents include age, race, gender, marital status, household relationship, education, veteran status, parent/guardian status and income and asset sources. Income and asset variables which are used to match noninterviews with respondents are obtained from the previous wave Control Card for both current wave noninterviews and respondents if an interview was obtained in the previous wave; otherwise, income and asset variables are not used as matching variables. In practice, a noninterview cannot always be matched with a respondent on all matching variables. To account for situations where a match cannot be made on all variables, simultaneous matches are made at several lower levels of detail by omitting some matching variables and reducing the number of categories in others.

Compensation for Item Missing Data

Both logical, or deductive, and statistical imputation procedures are used in the SIPP to compensate for item missing data in the Core questionnaire and Topical Modules. Variations in the statistical imputation procedure are applied to item missing data in the Control Card. Logical imputation is used in the SIPP whenever missing or inconsistent items can be reasonably inferred from nonmissing items within the same record. Prior to the point in the cross-sectional editing process at which a missing item would be imputed a check is made for feasible values within the section the missing item is located. If a feasible value can be inferred from reported information, the

inferred value replaces the missing value and no statistical imputation of that item for that case is performed.

The statistical imputation procedure used in the SIPP to compensate for item missing data in the Core questionnaire, Control Card and Topical Modules is referred to as a sequential hot-deck. The hot-deck procedure used in the SIPP is sequential because the selection of replacement values is implemented one record at a time from an ordered file. The sequential hot-deck procedure used in the SIPP is carried out independently for each wave and by groups of related variables within the Core questionnaire, and involves five key steps: 1) specifying cold-deck or starting values; 2) sorting the sample cases; 3) preprocessing the data file to identify records with no item missing data within a group of related variables and to update cold-deck values.; 4) classifying cases into subclasses of the population, referred to as imputation classes or adjustment cells, according to values on a set of classification or auxiliary variables which are nonmissing for all cases; and 5) selecting replacement values from donor records to impute item missing data on recipient records.

The sequential hot-deck imputation procedure used in the SIPP begins by filling a large matrix with starting or cold-deck values. The cells in the matrix are defined by the cross classification of auxiliary variables. Each cell in the matrix corresponds to respondent cases with the same set of values on the classification variables.

The matrix is initially referred to as the cold-deck matrix. During subsequent stages of processing, as the cold-deck values are replaced with information from the current wave, the array of cells is referred to as the hot-deck matrix. Historically, cold-deck values in a sequential hot-deck procedure served as the initial set of replacement values for missing items in the first record processed; missing items in subsequent records would typically receive replacement (hot-deck) values from the current data set. In the SIPP, however, cold-deck values are not frequently used as replacement values for either the first or subsequent records processed. The primary purpose of cold-deck values in the SIPP is to initialize the cold-deck matrix.

The records in the sample file are sorted by three geographic variables prior to imputing item missing data. The three geographic sort variables are primary sampling unit, segment number and serial number. The cases are sorted prior to processing and are not resorted at any other time during the imputation process. The sorting operation creates a file in which neighboring records represent geographically proximate households. Once the cases have been sorted they are processed through a series of edit programs. During the first pass against the edit programs the cold-deck values in the matrix are updated with information from the current wave, but missing data are not imputed. The imputations are performed during the second pass through the edit programs. The initial processing is done separately (but simultaneously) for each group of related Core questionnaire variables outlined above.

During the initial pass against the edit programs the first record in the sorted file with consistent and nonmissing data for a particular section is identified and the values from this case replace the cold-deck values for that section in the matrix. The values for each subsequent record with consistent and nonmissing information in a section update the previous set of consistent and nonmissing values written to the matrix. The last set of values written to the matrix serve as the starting values in the subsequent sequential hot-deck procedure. In this way, cold-deck values are rarely used as replacement values in the SIPP because the initial processing usually replaces all starting values with values from the current wave of data collection. In the next step of the imputation procedure each respondent and noninterview record in the sorted file is allocated to one of the imputation classes or adjustment cells according to its values on the set of classification or auxiliary variables. Each matrix is defined by a different set of classification variables and corresponds to a single item or to a series of related items whose missing data are imputed.

The selection of classification variables is determined by the subject matter specialists at the Bureau of the Census who base their selections on the extent to which the nonmissing values of the variable being imputed are correlated with the classification variables, the extent to which the classification variables are nonmissing for all cases and the linkages through edits. Ideally, the set of classification variables should account for a large proportion of the variance in the variable being imputed and be associated with variations in response rates.

The allocation of sample cases into imputation classes (also known as subclasses or strata) according to a set of classification variables serves several purposes. The classification procedure creates homogeneous adjustment cells such that cases within an adjustment cell are more similar than cases between adjustment cells. In this way donors and recipients are similar under the assumption that the nonresponse mechanism within the imputation class is not related to the item being imputed; that is, an underlying assumption is made that item missing data are distributed randomly within the subclass defined by the cross classification of the auxiliary variables. The implicit stratification

created by the sort order of the file further improves the opportunity for a better quality imputation to the extent that the sort order creates positive autocorrelation, where nearby cases are more similar to each other than cases which are further apart in the file.

As the file is processed through the set of edit programs the second time and the imputations are performed, the set of hot-deck values is updated once again, but this time the updating procedure is item by item rather than case by case. Missing items in the first record processed receive the final set of replacement values which were obtained from the initial updating procedure. The nonmissing values in the first record processed update the corresponding set of current hot-deck values. These current hot-deck values, in turn, donate information to any missing items in the next record processed. The records are processed one at a time in a sequential fashion according to the sort order of the file. A missing item is imputed the value of the item in the last nonmissing record processed for that imputation class. If the value of an item in the current record is nonmissing it replaces the previous hot-deck value for that imputation class. In this way the hot-deck value for each imputation class is constantly being updated with the value of the last nonmissing case.

Item missing data in Topical Modules are imputed using the same sequential hot-deck procedure used to impute item missing data in the Core questionnaire. Topical Module data for Type Z and Departure noninterviews are not Type-Z imputed, but rather imputed item by item using the sequential hot-deck procedure used to impute Core questionnaire item missing data. Other features of the implementation of the sequential hot-deck procedure for Topical Modules include: 1) more frequent changes in cold-deck values for variables sensitive to changes in economic activity; and 2) more frequent changes in the composition of classification variables. All other aspects of the Topical Module imputation procedure are similar to the features used to impute item missing data in the Core questionnaire.

The method for imputing item missing data in the Control Card is also a sequential hot-deck procedure but involves fewer steps than the Core questionnaire item missing data imputation procedure.

The first step in imputing item missing data on the Control Card involves specifying cold-deck values. In the second step the Control Card file is sorted by the same three geographic variables used to sort the Core questionnaire data file: primary sampling area, segment number and serial number. The preprocessing step to identify consistent and nonmissing records and to initially update cold-deck values, and the step which allocates cases into imputation classes in the Core questionnaire imputation procedure is omitted from the Control Card imputation procedure. No imputation classes are maintained in the Control Card procedure because the neighboring household with nonmissing information for an item is considered the best donor available. Another variation from the Core questionnaire procedure is that missing items on the Control Card are replaced with nonmissing values from the same donor, rather than from multiple donors.

Once cold-deck values have been specified and the file has been sorted the Control Card records are processed sequentially. Missing items in the first Control Card processed receive cold-deck replacement values. The cold-deck values are subsequently updated with information from the first Control Card record encountered without missing data. Each succeeding Control Card record encountered with no missing information updates the values in the hot-deck matrix. In turn, each Control Card record encountered with missing information is replaced with nonmissing information from the hot-deck. In this way any missing data on a Control Card record is replaced with information from the nearest neighboring record with no missing data.

Longitudinal Edits

To facilitate analysis of SIPP data across waves, the Bureau of the Census has developed a system which links together wave records to produce longitudinally processed data sets. The longitudinal edits are applied only for selected variables and only after all waves of a panel have been processed cross sectionally. This section provides a brief overview of the procedures which edit the data for consistency over time to produce the SIPP Full Panel Microdata Research files.

In general, the longitudinal edits do not replace missing data in one case with reported data from another case. Rather, when a data value is modified during longitudinal editing the replacement value is obtained: 1) from the same or different wave for the same case; 2) by extrapolation from a previous wave or by interpolation between waves for the same case; or 3) by some other procedure such as averaging which evens out fluctuations in a series of imputed values.

The longitudinal data sets are constructed and edited in several steps, each of which is performed independently on a subset of related variables. Each subset of variables is processed in a three-step sequence. First, the relevant variables for each section are extracted from the individual wave files and then moved to a record constructed for each sample person. Second, the longitudinal edits are applied. Third, the edited data

are added to the longitudinal file which is constructed in segments by joining together each subset of edited variables.

The longitudinal editing procedures are guided by several considerations including, the fundamental requirement to ensure cross-wave consistency, which only becomes apparent when multiple waves of SIPP data are examined together; the realization that not all possible edits and consistency checks can be implemented; the opportunity to address any problems associated with wave files; the preference to replace imputed values from one wave with reported values from another wave when available; and the need to reduce the number of variables carried from the wave files to the longitudinal files in order to condense the physical size of the data sets.

Many demographic variables are edited for consistency during wave processing by comparing survey responses provided in Wave 1 with responses provided in subsequent waves. For other demographic variables, and all household variables, inconsistencies are detected and corrected when the data are processed longitudinally. For example, persons reported as widowed in Wave 1 may be reported as never married in Wave 2, or two persons reported as parent and child in one wave may be reported as husband and wife in another wave. These and many other potential inconsistencies only become apparent when multiple waves of data are examined together.

The series of questions covering various aspects of each sample person's labor force situation during the four-month reference period is not longitudinally edited because: 1) a nonmissing response is required in the first item (whether or not the sample person worked during the four-month reference period) in order for the interview to be considered complete; and 2) item missing data rates for other key status indicators are low (generally less than 1 percent).

The SIPP Core questionnaire collects data on up to two wage and salary jobs and two self-employment businesses in a wave. Each job or self employment business is then uniquely identified across all waves by identification numbers. These identification numbers are subsequently used to link data about a particular job or business within and between waves and are longitudinally edited to identify and correct any errors in assigning the identification numbers. The edit also identifies jobs or business with imputed earnings amounts and replaces the imputed values with reported amounts obtained in previous or subsequent interviews.

Longitudinal editing of hourly wage and monthly earnings amounts are performed after the job or business identification numbers have been edited. Imputed hourly wage rates are replaced with the average of the reported values for a specific employer if at least one reported value is present. If no reported values are available for a specific job, the imputed values are replaced by the average imputed value. When an imputed hourly wage rate for a specific job is replaced with the average of the reported or imputed values, monthly amounts earned at that job are recalculated. The monthly amount earned for hourly wage jobs is calculated by multiplying the number of weeks with pay for that month by: 1) the usual number of hours worked per week; and 2) the edited hourly wage rate for that month.

The edit procedure for earnings amounts collected on a monthly basis is also based on an averaging algorithm which results in replacement of imputed monthly earnings values with either values derived from reported data, or with values derived from all cross-sectionally imputed values, if no reported data exist. The first step in the edit procedure involves calculating an "implied" hourly wage and salary amount for a specific job. The implied hourly wage amount is calculated by first replacing imputed monthly earnings amounts with either the average of the reported amounts, or if no reported amounts are present, by the average of the imputed amounts. Months with zero earnings are excluded from the calculation. The monthly earnings amounts are then summed and divided by the sum of the products of: 1) the number of weeks with pay; and 2) the usual hours worked per week for each month. The quotient is the implied hourly wage and salary amount. The replacement value for imputed monthly earnings amounts is obtained by multiplying the implied hourly wage rate by: 1) the number of weeks with pay; and 2) the usual number of hours worked per week for the month. An additional edit is performed on earnings amounts collected on a monthly basis for workers paid by the hour. This edit compares the reported monthly earnings amount with a calculated monthly earnings amount. If the reported monthly amount is 10 times greater than the calculated amount, the reported amount is replaced with the calculated amount. The purpose of this edit is to decrease the number of monthly amounts that have a high probability of being wrong.

The longitudinal edits for general amount variables are described separately for: 1) nonwage and salary income sources numbered 1-56; and 2) asset types numbered 100-150. Also described are designed to detect the presence of duplicate amounts. The edits for income amounts 1-56 are applied only to imputed amounts. No reported cross-sectional amounts are changed. If all monthly amounts for all reference periods for a specific income source were imputed, these

imputed amounts are averaged and the average imputed amount replaces the original imputed amounts; otherwise, imputed amounts are replaced by reported amounts obtained from other reference periods. When both reported and imputed amounts are present on a record, the imputed amounts are replaced with the nearest reported amount. The implementation of the nearest neighbor concept gives priority to the first month with a reported value preceding the month containing an imputed value. The monthly income amount which meets this criterion replaces the imputed amount. The first succeeding month with a reported value is used as a replacement value only when no month prior to the month requiring replacement contains a reported amount.

The longitudinal editing procedures for asset types 100-150 vary from those used for income types 1-56. Instead of using the nearest neighbor concept, any values for asset types 100-150 which were imputed during cross-sectional editing are replaced with the average of the reported values from other waves.

The primary means of detecting duplicate reporting of income amounts for AFDC, Food stamps and WIC by both the husband and wife are through item checks in the questionnaire. Any additional instances of duplicate reporting are identified during longitudinal processing by locating husbands and wives reporting amounts for the same income source for the same month and deleting either the husband's or wife's amount.

Bibliography

Coder, J.F. (1978) Income Data Collection and Processing from the March Income Supplement to the Current Population Survey. The Survey of Income and Program Participation Proceedings of the Workshop on Data Processing, February 23-24, 1978, (Eds. D. Kasprzyk), Chapter II. Washington, D.C.: U.S. Department of Health, Education and Welfare.

Doyle, P. and Dalrymple, R. (1987) The Impact of Imputation Procedures on Distributional Characteristics of the Low Income Population, Proceedings of the Bureau of the Census Third Annual Research Conference, Washington D.C., Department of Commerce, PP. 483-508.

Jabine, Thomas, B. (1990) Survey of Income and Program Participation Quality Profile, Second Edition, U.S. Bureau of the Census.

Jinn, Jann-Huei and Sedransk, J. (1987) Effect on Secondary Data Analysis of Different Imputation Methods, Proceedings of the Bureau of the Census Third Annual Research Conference, U.S. Department of Commerce, Washington D.C., pp. 509-530.

Kalton, G. (1983) Compensating for Missing Survey Data, Research Report Series, Institute for Social Research, The University of Michigan. Ann Arbor, Michigan

Kalton, G. and Kasprzyk, D. (1982) Imputing for Missing Survey Responses. Proceedings of the Section on Survey Research Methods, American Statistical Association. pp. 22-31.

Kalton, G. and Kasprzyk, D. (1986) The Treatment of Missing Survey Data. Survey Methodology, Vol. 12, No. 1, pp. 1-16.

Kalton, G., Kasprzyk, D. and Santos, R. (1981) Issues of Nonresponse and Imputation in the Survey of Income and Program Participation. In Current Topics in Survey Sampling, Proceedings of the International Symposium on Survey Sampling, (Eds. D. Krewski, R. Platek, J.N.K. Rao), New York, Academic Press, pp. 455-480).

Kish, L. (1965) Survey Sampling, John Wiley & Sons, New York.

Lepkowski, J.M, Landis, R.L., and Stehouwer, S.A. (1987) Strategies for the Analysis of Imputed Data Froma Sample Survey, Medical Care, Vol, 25, No. 8., pp.705-716.

Little, J.A. Roderick (1986) Missing Data in Census Bureau Surveys, Proceedings of the Bureau of the Census Second Annual Research Conference, U.S. Department of Commerce, Washington D.C., pp. 442-454.

Nelson, D., McMillen, D. and Kasprzyk (1985) An Overview of the Survey of Income and Program Participation: Update 1, SIPP Working Paper Series No. 8401, U.S. Bureau of the Census., Washington D.C.

Sande, I.G. (1982) Imputation in Surveys: Coping with Reality. The American Statistician, Vol. 36, pp. 145-152.

Sande, I.G. (1983) Hot-Deck Imputation Procedures. In Incomplete Data in Sample Surveys, Vol. 3, Proceedings of the Symposium, (Eds. W.G. Madow and I. Olkin), New York: Academic Press, pp.339-349.

Santos, R.L. (1981) Effects of Imputation on Regression Coefficients, Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 140-145.

Sedransk, J. (1985) The Objectives and Practice of Imputation, Proceedings of the Bureau of the Census First Annual Research Conference, Washington D.C., U. S. Department of Commerce, pp. 445-452.

Singh, R.P., Huggins, V., and Kasprzyk, D. (1990) Handling Single Wave Nonresponse In a Panel Survey, SIPP Working Paper Series No. 9009, U.S. Bureau of the Census, Washington D.C.

Singh, R.P. and Petroni, R.J. (1988) Nonresponse Adjustment Methods for Demographic Surveys at the U.S. Bureau of the Census, SIPP Working Paper Series No. 8823.

Survey of Income and Program Participation (SIPP) 1984 (and 1987) Full Panel Microdata Research File, Technical Documentation. U.S. Bureau of the Census, Washington, DC (1990).

U.S. Bureau of the Census Memoranda, SIPP 1984 Panel Cross-Sectional Imputation System Hot-Deck Matrices for Core Item Nonresponse (September 12, 1984).

Welniak, E.J. and Coder, J.F. (1980) A Measure of the Bias in the March CPS Earnings Imputation System. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 421-425.