LONGITUDINAL ITEM IMPUTATION IN A COMPLEX SURVEY

Michael E. Samuhel and Vicki Huggins, Bureau of the Census

INTRODUCTION

Missing data in sample surveys are of two general forms. Unit nonresponse occurs when no information is available to the survey for an entire sample unit, such as a person, or household, or hospital. Some information may, however, be available from other kinds of records such as those used to define the sample frame. The reasons for unit nonresponse vary; for example, a person may refuse to respond, be away from home, or be impossible to locate. Typically, this form of nonresponse is handled in part by a call-back strategy. That is, the interviewer makes repeated attempts to contact the unit. If the call-back strategy fails, or is not feasible, weights can be assigned to the responding units (Cochran, 1977).

The other type of nonresponse is item nonresponse. It occurs when the unit supplies information for some but not all of the variables. For example, a person may answer questions about age, race, and sex but not about income; or the information may be deleted by an edit failure. Depending upon the intended uses of the data, item nonresponse can be handled with two different but overlapping approaches. Either the data can be completed using imputation methods, or the recorded data can be used with modified estimation methods. The modified estimation methods may also be used to impute the missing data.

The focus of this paper is the imputation of categorical data in a longitudinal survey. Statistical research pertaining to missing categorical data has considered censored, discrete random variables and partially or completely unobserved data in contingency tables. Harley's (1958) solution to the problem of estimating the rate parameter for a censored Poisson random variable is a special case of what was later called the EM algorithm. Fuchs (1982) applied the EM algorithm to find maximum likelihood estimates for parameters in a log-linear model, when the values of one or more variables are missing for subsets of the cross-classified data. Chen and Fienberg (1974) developed models for analyzing contingency tables with supplemental marginal totals.

Unfortunately, none of these methods offer solutions to the problem of missing categorical data in complex, longitudinal surveys such as the Survey of Income and Program Participation (SIPP). Although a contingency table could be constructed from monthly responses to a categorical survey item over a year, the resulting twelve dimensional table would be exceedingly sparse. In addition, the application of log-linear models or the EM algorithm to such tables would be computationally difficult.

In this paper we describe a general method for imputing missing categorical items in longitudinal surveys. We show that the longitudinal data, completed according to the method, provides unbiased estimates of the probability of occurrence of the various response patterns, assuming that the data are observed at random and missing at random (Rubin, 1976). The importance of longitudinal information for imputing missing data is discussed, and a statistic measuring the amount of information available is described.

The imputation methodology described here was developed from data collected by the Income Survey Development Program (ISDP). The method is suggested as the fundamental tool for imputing missing, longitudinal, categorical items in the Survey of Income and Program Participation (SIPP). However, its implementation can occur only after further development and modifications. Here, it is described as a general, statistical approach applicable to any longitudinal survey. The data from the ISDP is utilized only to explain the method and provide examples.

The Income Survey Development Program (ISDP) was initiated to gain experience with the data collection and data analysis requirements of SIPP. The ISDP is a longitudinal survey consisting of two national panels (1978, 1979). The sample design is a multi-stage stratified sample of the United States population. Sampling elements are housing units not households (which may move) or persons. The first sampling stage involves the definition of the United States in terms of counties or groups of counties called primary sampling units (PSU's), which are stratified. At the second stage, a sample of addresses within the PSU's is selected. To minimize the inconvenience to sample participants, interviews are conducted every three months. Each household is assigned to one of three rotation groups (A,B,C). Every three months all the households in a rotation group are interviewed and data is collected for each of the previous three months. A wave is the time period during which each rotation group is interviewed once. Data from each wave is published by the United States Bureau of the Census as a cross-sectional file. The longitudinal data for our imputation research is an annual file, constructed by merging five waves of ISDP data from the 1979 panel.

THE IMPUTATION OF MISSING LONGITUDINAL CATEGORICAL SURVEY ITEMS

Many of our activities today are the direct result of events which occurred yesterday. Last night we may have arrived home late, returning from a long trip. Today, it is likely that we will need to stop off at the gas station to refill our car's fuel tank. Or perhaps yesterday we were layed-off from our job. Today we are reading the employment opportunities section of the newspaper.

Analogously, in the ISDP, there are strong dependencies between the monthly values of the survey items. For example, fitting a logistic regression of the receipt of wages and salaries in July on the receipt reported in other months, we found the parameters for the months June, August, and November to be significantly different from zero. Similiar results where obtained in regressions of each month on the remaining months.

Define a longitudinal record for a survey unit to be the set of responses recorded over a fixed time period. In the ISDP as well as SIPP, the survey unit is a household, but other examples of survey units include the person, family, and employer. In this paper, the survey person is the unit of analysis. The set of responses on the longitudinal record may be any combination of survey items. Here, we restrict ourselves to a single item recorded monthly for one year. For example, the receipt of wages and salaries.

The following example illustrates the imputation process. Consider the ISDP survey item indicating whether a person had a job or business during a month. Further, consider the set of individuals in rotation group A who responded "yes" from January thru November 1979, but did not respond in December, 1979. The longitudinal record for these individuals is given by

$$X = (0,0,0,0,0,0,0,0,0,0,0,2),$$

where $X_t = 0$ (t=1,...,12), if the response in the t^{th} month is "yes", $X_t = 1$ if the response is "no", and $X_t = 2$ indicates missing data. Either "0" or "1" is an admissible imputation value for X_{12}. Based on the individuals in rotation group A who reported data in every month from January to December we estimate

$$\text{Prob}\,(X_{12} = 0 \mid X_1 = 0, X_2 = 0,\ldots,X_{11} = 0)$$
$$= \frac{2313}{2379} = 0.9723, \text{ and}$$

$$\text{Prob}\,(X_{12} = 1 \mid X_1 = 0, X_2 = 0,\ldots,X_{11} = 0)$$
$$= 1 - .9723 = 0.0277\,.$$

Generating a random number between zero and one, we impute $X_{12} = 0$ if the random number is less than or equal to 0.9723, otherwise we impute $X_{12} = 1$.

This imputation procedure can be applied to any categorical survey item with any combination of missing months. Consider the sample item indicating the monthly receipt of wages and salaries and the following longitudinal record for persons in rotation group A

$$X = (0,0,0,0,0,0,0,2,2,2,0,0)\,.$$

Based on those persons responding in all twelve months, we estimate

$$\text{Prob}\,(X_8 = x_8, X_9 = x_9, X_{10} = x_{10} \mid X_1 = 0,\ldots,X_7 = 0, X_{11} = 0, X_{12} = 0) \quad (1)$$

$$= \frac{1120}{1140} = 0.9823 \text{ if } X_8 = 0,\ X_9 = 0,\ X_{10} = 0\,,$$

$$= \frac{10}{1140} = 0.0088 \text{ if } X_8 = 1,\ X_9 = 0,\ X_{10} = 0\,,$$

$$= \frac{4}{1140} = 0.0035 \text{ if } X_8 = 0,\ X_9 = 0,\ X_{10} = 1\,,$$

$$= \frac{3}{1140} = 0.0026 \text{ if } X_8 = 1,\ X_9 = 1,\ X_{10} = 0\,,$$

$$= \frac{2}{1140} = 0.0018 \text{ if } X_8 = 0,\ X_9 = 1,\ X_{10} = 0\,,$$

$$= \frac{1}{1140} = 0.0009 \text{ if } X_8 = 1,\ X_9 = 0,\ X_{10} = 1\,.$$

Here, we impute the entire subvector (x_8, x_9, x_{10}) based on a random draw from a uniform (0,1) distribution.

The imputation process is formalized by letting the random variable X represent the responses (and missing data) on a longitudinal record. The vector X = x can be partitioned into subvectors x_m and x_r, representing the missing and recorded monthly valves, respectively. On the i^{th} longitudinal record, we impute the missing items X_{mi} based on the reported values x_{ri}. The imputed values are a random draw from the conditional distribution $f(x_m \mid X_r = x_{ri})$, emperically estimated from the longitudinal records with values reported in every month.

AN UNBIASED ESTIMATE OF THE OCCURRENCE PROBABILITY OF A LONGITUDINAL PATTERN

Response patterns to survey items are singularly important in longitudinal surveys. The longitudinal data is collected so that changes over time of the survey items can be analyzed. For example, a researcher may wish to accurately estimate the average duration of unemployment or the length of time an individual participates in a social welfare program. It is important that the imputations do not disrupt the frequency distribution of response patterns and bias these longitudinal estimates.

Consider a simple random sample of a size n without nonresponse. The longitudinal records for individuals in the labor force every month are represented by

$$X = (0,0,0,0,0,0,0,0,0,0,0,0) \qquad (2)$$

Let the binomial random variable T represent the number of times the pattern (2) occurs. It follows that

$$\frac{1}{n}\, T\, (X_1 = 0, X_2 = 0,\ldots,X_{12} = 0) \qquad (3)$$

is an unbiased estimate of

$$\text{Prob}\,(X_1 = 1, X_2 = 1,\ldots,X_{12} = 1).$$

Of course, in longitudinal surveys with complex sample designs like SIPP, the statistic (3) would need to be modified to reflect the particular survey design.

In longitudinal files, completed according to the imputation method described above, statistics analagous to (3) are also unbiased estimates of the probability that the particular pattern occurs; provided the data are missing at random and observed at random (Rubin, 1976). We prove this result for longitudinal records containing two time periods. Without loss of generality the result extends to longitudinal records of any length.

THEOREM

Consider the longitudinal record (X_1 = a, X_2 = b), where a and b represent the only values of the categorical random variables X_1 and X_2. In a simple random sample of size n, completed by imputation, let the binomial random variable $T'(X_1 = a, X_2 = b)$ represent the number of occurrences of the longitudinal record. Assuming the data are observed at random and missing at random.

$$\frac{1}{n}\, T'(x_1 = a, x_2 = b)$$

is an unbiased estimate of

$$\text{Prob}\,(X_1 = a,\ X_2 = b)$$

Proof:

The pattern $(X_1 = a,\ X_2 = b)$ can arise in the imputed sample in four ways:

1) $(X_1 = a,\ X_2 = b)$ is reported,
2) $X_1 = a$ is imputed given $X_2 = b$ is reported
3) $X_2 = b$ is imputed given $X_1 = a$ is reported,
4) $(X_1 = a,\ X_2 = b)$ is imputed.

Define the binomial random variable T( ) as the number of occurrences of the event in parentheses. For example, using an astrisk to indicate imputed counts,

$$T^*(X_1 = a \mid X_2 = b)$$

represents the number of times that $X_1 = a$ is imputed given that $X_2 = b$ is reported.

The total number of times the pattern

$$(X_1 = a,\ X_2 = b)$$

occurs in the sample, completed by imputation, can be decomposed into terms corresponding to the four ways the pattern (a, b) arises,

$$\begin{aligned} T'(X_1 = a,\ X_2 = b) \ = \ & T(X_1 = a,\ X_2 = b) + \qquad (4)\\ & T^*(X_1 = a \mid X_2 = b) + T^*(X_2 = b \mid X_1 = a) + \\ & T^*(X_1 = a,\ X_2 = b). \end{aligned}$$

Let the indicator vector $Y = (Y_1,\ Y_2)$ represent the reporting status of the elements in the longitudinal record. That is,

$$\begin{aligned} Y_i \ &= \ 1 \ \text{ if } X_i \text{ is reported } (i=1,2),\\ &= \ 0 \ \text{ otherwise }. \end{aligned}$$

The expected value of the sum (4) with respect to the data reported in the sample is

$$\begin{aligned} & E(T'(X_1 = a,\ X_2 = b) \mid T(X_1 = a,\ X_2 = b)\,) = \qquad (5)\\ & T(X_1 = a,\ X_2 = b,\ Y_1 = 1,\ Y_2 = 1) + \\ & T(X_2 = b,\ Y_1 = 0, Y_2 = 1)\cdot \\ & \left[\frac{T(X_1 = a,\ X_2 = b,\ Y_1 = 1,\ Y_2 = 1)}{T(X_2 = b,\ Y_1 = 1,\ Y_2 = 1)}\right] + \\ & T(X_1 = a,\ Y_1 = 1,\ Y_2 = 0)\cdot \\ & \left[\frac{T(X_1 = a,\ X_2 = b,\ Y_1 = 1,\ Y_2 = 1)}{T(X_1 = a,\ Y_1 = 1,\ Y_2 = 1)}\right] + \\ & + T(Y_1 = 0,\ Y_2 = 0)\cdot \\ & \left[\frac{T(X_1 = a,\ X_2 = b,\ Y_1 = 1,\ Y_2 = 1)}{T(Y_1 = 1,\ Y_2 = 1)}\right]. \end{aligned}$$

Note that the random variables in the conditional expectation (5) are multimonial. The expectation with respect to all possible samples is found by applying the following result.

LEMMA

Let $(X_1,\ldots,X_k)$ be multimonial $(n;P_1,\ldots,P_k)$ random variables, then X_1 and X_3 are independent given $X_1 + X_2 = \ell$ and

$$E\ X_3\ \frac{X_1}{X_1 + X_2} \ = \ nP_3\,\frac{P_1}{P_1 + P_2}\ .$$

The expectation of (5) with respect to all possible samples follows from the lemma. In addition, the assumption that the data are observed at random and missing at random asserts the independence of the indicator random vector Y and the random variables in the longitudinal record.

$$\begin{aligned} & E(T'(X_1 = a,\ X_2 = b)) \ = \\ & E_2\ E_1\ (T'(X_1 = a,\ X_2 = b) \mid T(X_1 = a,\ X_2 = b)) \ = \\ & n\ \text{Prob}\ (X_1 = a,\ X_2 = b)\ \text{Prob}\ (Y_1 = 1,\ Y_2 = 1) + \\ & n\ \text{Prob}(X_2 = b)\ \text{Prob}(Y_1 = 0,\ Y_2 = 1)\ \cdot \\ & \left[\frac{\text{Prob}(X_1 = a,\ X_2 = b)}{\text{Prob}(X_1 = a,\ X_2 = b) + \text{Prob}(X_1 = b,\ X_2 = b)}\right] + \\ & n\ \text{Prob}(X_1 = a)\ \text{Prob}(Y_1 = 1,\ Y_2 = 0)\ \cdot \\ & \left[\frac{\text{Prob}(X_1 = a,\ X_2 = b)}{\text{Prob}(X_1 = a,\ X_2 = a) + \text{Prob}(X_1 = a,\ X_2 = b)}\right] + \\ & n\ \text{Prob}(Y_1 = 0,\ Y_2 = 0)\ \ \text{Prob}(X_1 = a,\ X_2 = b) \\ & = n\ \text{Prob}(X_1 = a,\ X_2 = b) \end{aligned}$$

QED

The theorem is extended to longitudinal records of any length by adding the appropriate terms to equation (3).

THE EXPECTED NUMBER OF INCORRECT IMPUTATIONS

Longitudinal data by itself may not always be sufficient to accurately impute missing data. The amount of information available longitudinally can be measured by estimating the expected number of incorrect imputations. Consider the longitudinal record for the monthly receipt of wages and salaries,

$$X = (0,0,0,0,0,0,0,0,0,0,0,2), \qquad (6)$$

where $X_t = 0$ indicates receipt and $X_t = 2$ $(t=1,\ldots,12)$ indicates missing data. The probability

$$\text{Prob}(X_{12} = 1 \mid X_1 = 0,\ldots,X_{11} = 0)$$

is estimated from the completely reported cases as

8/1236 = 0.0065. This probability is independent of but equal to the probability of imputing $X_{12} = 1$. Consequently, the probability that $X_{12} = 1$ is imputed and is correct is $(0.0065)^2$. Similiarly, the probability that $X_{12} = 0$ is imputed and is correct is $(0.9935)^2$. It follows that the estimated probability of an incorrect imputation for the longitudinal record (6) is

$$1 - (0.0065)^2 + (0.9935)^2 = 0.0013.$$

Since, there are seventeen individuals in the file with this longitudinal record, it follows that the estimated number of incorrect imputations is 17(0.013) = 0.22.

The need to include demographic information would be indicated by an estimated number of incorrect imputations greater than some predetermined value. Consider again the longitudinal record for the monthly receipt of wages and salaries,

$$X = (0,0,0,0,0,0,0,2,2,2,0,0).$$

Thirty-eight persons in rotation group A had this pattern. Using the probabilities given in (1), the estimated expected number of incorrect imputation is

$$38(1-0.9825^2 - 0.0088^2 - 0.0035^2 - 0.0026^2 - 0.0018^2 - 0.0009^2) = 1.25 .$$

Here, we want to use demographic information to choose the most appropriate donor pattern. One approach is to include associated survey items as elements in the longitudinal record. For example, to impute the monthly receipt of wages and salaries, we can include in the longitudinal records survey items indicating seasonal or part time workers. A logistic model may also be useful, especially when the data are sparse. Letting the polychotomous variable Y represent the available donor patterns, we can regress Y on concomitant data, represented by the vector X. Based on the concomitant information, the probability of pattern h for the i^{th} longitudinal record is

$$\text{Prob}(Y_i = h) = \frac{e^{B_h X_i}}{1 + e^{B_h X_i}} .$$

The pattern selected for imputation can be the one with the highest probability, or the decision can be based on a random number generated between zero and one.

CODING PATTERNS

The responses on any longitudinal record can be summarized as a single number. Consider the longitudinal record

$$X = (0,0,0,0,0,0,1,1,1,2,2,2),$$

representing the receipt of wages and salaries from January (X_1) to December (X_{12}). This pattern can be represented in base ten as

$$377 = (2x3^0) + (2x3^1) + (2x3^2) + 3^3 + 3^4 + 3^5 .$$

In general, any pattern in an annual file of monthly categorical data can be represented by the polynomial

$$P_\ell = \sum_{k=1}^{12} c_k B^{k-1} ,$$

Each pattern has a unique base ten representation, because the transformation is one-to-one and onto, the index k represents the months in the longitudinal file in reverse order. That is, k=1 represents December, k=2 represents November, and so on. The coefficients c_k represent the monthly values of the item. The letter B represents the appropriate base. Typically, the base is one more that the highest coefficent (c_k).

Coding the longitudinal record patterns as base ten numbers operationally simplifies the imputation process. Consider the longitudinal record

$$X = (0,0,0,0,0,0,0,0,0,2,2,2),$$

indicating the receipt of wages and salaries in each month from January thru December. The receipt of wages and salaries in the t^{th} month is denoted by $X_t = 0$, and a missing monthly item is denoted by $X_t = 2$. This pattern is represented in base ten by the number 26. Because the transformation to base ten is unique, all individuals in the data file with the value 26 for their pattern have reported the receipt of wages and salaries from January to September, but did not respond to the item from October to December.

Donor patterns from the cases, reporting values in every month, are identified by subtraction. For example, the donor pattern.

$$(X_{10} = 0, X_{11} = 0, X_{12} = 0)$$

is identified by subtracting the base three number 222 from the longitudinal pattern

$$\begin{array}{r} 000000000222 \\ -222 \\ \hline 000000000000 \end{array}$$

The equivalent operation could also be done in base ten. Noting that 222 is represented in base ten by 26, the donor pattern ($X_9 = 0$, $X_{10} = 0$, $X_{11} = 0$) is the base three representation of (26-26) = 0. Similiarly, all possible donor patterns i.e., 000 thru 111 and found by subtracting from 26 the corresponding base ten numbers 26 through 0.

APPLICATIONS AND EXTENSIONS OF THE METHOD

Limitations on the number of pages available in these proceedings preclude a complete discussion of our research on longitudinal item imputation. A more extensive description, especially as it applies to the Survey of Income and Program Participation, can be found in Samuhel and Huggins (1984).

Acknowledgements

The author wishes to thank Dr. Patrick Kelly for many helpful discussions and acknowledge his contributions to the proof of the theorem.

References

Chen, T. and Fienberg, S.E. 1974. Two dimensional contingency tables with both completely and partially classified data. Biometries 30: 629-642.

Cochran, W.G. 1977. Sampling Techniques, New York: John Wiley and Sons, Inc.

Fuchs, C. 1982. Maximum likelihood estimation and model selection in contingency tables with missing data. Journal of the American Statistical Association 77: 270-278.

Hartley, H.O. 1958. Maximum likelihood estimates from incomplete data. Biometrics 14: 174-194.

Rubin, D.B. 1976. Inference and Missing data. Biometrika 63: 581-592.

Samuhel, M.E. and Huggins, V.J. 1984. Longitudinal item imputation in a complex survey. Survey of Income and Program Participation Working Paper Series. United States Department of Commerce.