LONGITUDINAL IMPUTATION FOR THE SIPP

Steven G. Heeringa and James M. Lepkowski,
The University of Michigan

I. Introduction

The problem of item nonresponse in a survey arises when an otherwise cooperative respondent does not or cannot provide a response to one or more survey questions. Imputation, the estimation of a value for a missing response, is commonly used to compensate for such item missing data. Item nonresponse and its compensation methods become more complex in the case of a panel survey where a sample of respondents provides data at a series of points in time. In a panel survey, the item nonresponse problem can be extended to include wave nonresponse, that is, failure to obtain any data from a respondent at one or more waves of the data collection sequence. Whether the data are missing for an entire wave or only for specific items within a wave, longitudinal survey data can provide additional information which may be used to improve the quality of imputation for missing values (Kalton and Lepkowski, 1982).

Since panel data are usually collected and processed one wave at a time, imputation of missing values is often conducted for each wave separately using only the information available within a wave to derive an imputed value. Such "cross-sectional" imputations do not take advantage of the information collected at other waves of the panel. In contrast, longitudinal imputation methods have the capability to use data collected at other waves, data which may be highly correlated with the item to be imputed.

The purpose here is to examine longitudinal and cross-sectional imputation methods for item missing data in the Survey of Income and Program Participation (SIPP). The investigation reported in this paper uses selected survey variables from the first three waves of the 1984 SIPP panel to compare the effectiveness of a simple longitudinal direct substitution technique and that of the Census cross-sectional hot-deck imputation method. After describing the SIPP design and cross-sectional hot-deck imputation method in Section II, we review some longitudinal imputation methods that could be applied to the SIPP in Section III. In Secton IV, the simple longitudinal imputation method that is applied to the SIPP file is described, and Section V compares the longitudinal and cross-sectional imputations. The paper concludes with remarks about further investigations that might be conducted.

II. SIPP Design

The SIPP is a national survey of U.S. households conducted by the Bureau of the Census. It is designed to provide comprehensive information on both households' and individuals' economic status and participation in government programs. It is a panel survey in which households that participate in a baseline interview are followed and interviewed at 4 month intervals for a total of eight interviews. Interviewing for the 1984 SIPP panel began in October 1983 with an equal probability sample of about 20,000 households. (See Nelson, McMillen and Kasprzyk (1985) for a full description.)

The SIPP is designed to meet a range of analytic objectives. Some analyses involve the data for a single wave while others require data from several waves (e.g., analyses of annual incomes). Cross-sectional data collected at each wave of the SIPP are used to provide important estimates for quarterly reports on income and program participation. For this purpose, each wave of the SIPP panel is processed as a separate cross-sectional survey, and item missing data at each wave are handled by cross-sectional imputations.

The Bureau of the Census currently uses a cross-sectional hot-deck (CSHD) imputation for selected item nonresponse on individual waves of the SIPP (Nelson, McMillen and Kasprzyk, 1985). The first step in the CSHD procedure is to define a "hot-deck" matrix based on a cross-classification of characteristics that are correlated with the item being imputed. Based on the cross-classifying variables, each individual record is uniquely linked to a cell of the hot-deck matrix. To initialize the procedure, a "cold-deck" or starting value is assigned to each cell of the hot-deck matrix. The complete SIPP data file is then sorted by geographic characteristics and is passed through the hot-deck imputation program two times. In the first pass, no imputations are made, but if an observation has a non-missing value for an item to be imputed, that value "updates" the current value for the item stored in the hot-deck matrix.

In the second pass of the data, the actual imputation of missing values takes place. In the sequential order of the file, each record is examined and if the item is missing, the current value stored in the hot-deck cell for that item replaces the missing value on the record. If the value of the record is not missing, the non-missing value for that case replaces the current donor value for the hot-deck matrix cell. Thus, missing values for a record are, for the most part, replaced by values from another record that has the same characteristics used to define the hot-deck cell. For each item receiving imputations, an indicator variable is added to the SIPP file identifying which values have been imputed (Bureau of the Census, 1985).

III. Longitudinal Imputation Methods and Models

Longitudinal methods are designed to utilize cross-wave data in imputing the value of a missing item (Kalton and Lepkowski, 1982). However, the exact form in which the cross-wave information is used differs from one techinque to another. Five general classes of longitudinal imputation methods might be considered as an alternative to the CSHD method:

1) **Longitudinal direct substitution.** For items that are stable over time, the value of a nonmissing item is substituted from one time period to another where the same item is missing. Direct substitution can be a highly accurate form of imputation in some situations.
2) **Deterministic imputation of change.** Additive or proportionate change from one time period to another can be computed from the survey data or obtained from an exogenous source. Imputed values are created by applying this change to a non-missing value from an another wave.
3) **Longitudinal regression imputation.** Missing values are predicted from a regression equation obtained by fitting a model to data with nonmissing values. In the prediction, the residual term in the model can be set to zero for a deterministic form of regression imputation, or it can be assigned a value through a hot-deck or other stochastic procedure.
4) **Longitudinal hot-deck.** Auxiliary cross-wave information available from the longitudinally linked records is used to form the cells of the hot-deck matrix, extending the characteristics used in the CSHD procedure. Continuous items must be categorized to form the cells of the hot-deck matrix, reducing the strength of the cross-wave correlations. Nonetheless, the strength of correlations over time for stable items improves the accuracy of the CSHD procedure.
5) **Longitudinal hot-deck imputation of change.** Longitudinal hot-deck procedures are used to impute change from a donor record to the case with the missing value. The imputed change can be added directly to a nonmissing value from a prior or succeeding wave or another wave's nonmissing value can be proportionately altered.

Under these five general longitudinal imputation strategies, the value imputed for the ith respondent with missing data is derived as $y_i = f(x_{1i}, x_{2i}, \ldots, x_{pi}) + e_i$ where $f(\cdot)$ is a function of p auxiliary variables and e_i is an estimated residual. For the five general strategies the function $f(\cdot)$ can be expressed as a linear function where $y_i = b_0 + b_1 x_{1i} + \ldots + b_p x_{pi} + e_i$, and the b_j's are estimated from data for respondents with no missing values for y_i or the auxiliary variables.

Figure 1 presents simple linear models corresponding to the five general strategies to illustrate the relative features of the longitudinal imputation strategies. The simplest model is associated with the longitudinal direct substitution (LDS) method in which a nonmissing value is essentially "carried over" from another wave. Each of the other methods can be viewed as a modification of the LDS strategy incorporating proportionate change, additive change, and stochastic variation. For example, the deterministic imputation of change method can improve the LDS method by including an additive component of change (a), a proportionate change (cx_i), or both additive and proportionate change ($a + cx_i$) to the "carry-over" LDS imputation.

Figure 1

Models for Longitudinal Imputation Methods

Method	Model	Component of Change		
		Proportionate	Additive	Stochastic
Direct Substitution	$y_i = \mathbf{x}_i$	None	None	None
Deterministic Imputation of Change	$y_i = c\mathbf{x}_i$ or	$(1\text{-}c)\mathbf{x}_i$	None	None
	$y_i = c\mathbf{x}_i + a$	$(1\text{-}c)\mathbf{x}_i$	a	None
Longitudinal Regression	$y_i = b_0 + b_1\mathbf{x}_i + e_i$	$(1\text{-}b_1)\mathbf{x}_i$	b_0	e_i
Longitudinal Hot Deck	$y_i = b_0 + b_1\mathbf{x}_i + \overset{*}{e}_{j \neq i}$	$(1\text{-}b_1)\mathbf{x}_i$	b_0	$e_{j \neq i}$

From this perspective, the LDS method may be viewed as a base longitudinal imputation procedure to which modifications can be made to address deficiencies in the quality of the LDS imputations. As an initial investigation of the general longitudinal approach, a comparison of the LDS to the CSHD imputations will indicate whether longitudinal methods improve the quality of imputed values. Thus, the subsequent discussion examines the LDS as a base longitudinal imputation method relative to the CSHD imputations available in the SIPP data files.

Although the LDS method is conceptually simple, implementation can be complicated, because cross-wave information may not be available for each record with missing data on one wave. The general LDS strategy employed in this study was essentially a two step process. When an item could be carried over longitudinally, the imputation was made. Otherwise, the Census CSHD imputed value was used as the imputed value.

The LDS method has also been implemented somewhat differently for categorical and continuous types of variables. For categorical variables, the records with imputed responses (i.e., with missing data that has been replaced by the CSHD method) were scanned to determine if an actual value was available at a prior (or a subsequent) wave. If so, the actual value from the alternate wave was imputed for the missing item. If no value was available, the original CSHD imputed value was left unchanged. When two "donor" values were available, but different in value, the value from the "nearest" data collection wave was imputed for the missing item. For continuous variables, the LDS imputation algorithm also scanned the longitudinal data record to identify the full set of potential donor items for a missing value. But instead of selecting one member of the set as a "donor", the ***average*** of all nonmissing values was imputed for the missing item.

Finally, some SIPP variables such as earnings and wages undergo both systematic changes and random fluctuation across time. Therefore, short of performing the evaluation on a complete data set where both the amounts and patterns of missing values are simulated, it is difficult to choose an appropriate benchmark to measure the accuracy of imputations. Simulation can be a useful tool (Kalton and Lepkowski, 1982), but for the current study it has several drawbacks. First, since the simulation must operate on a data set with no missing values, the extension of the results to a full data set requires strong assumptions (or knowledge) about the distributions of the missing and non-missing values. Secondly, simulation of "missingness" would have to be carried out separately for each variable under study. This would require a large investment in set-up time and computing funds. By necessity then, the comparison of the CSHD and LDS imputation methods is presented here simply as a demonstration of what happens to actual distributions of these variables under the two imputation alternatives.

IV. Implementation of the Longitudinal Direct Substitution Method

Using data from 1984 SIPP Panel, an empirical investigation was conducted to test the feasibility and effectiveness of simple longitudinal imputation as an alternative to imputations based solely on cross-sectional hot deck methods.

The empirical study used a longitudinal file created from the first three waves of the 1984 SIPP panel. The Bureau of the Census cross-sectional public use files of data collected in the first three waves were merged to create longitudinal records of various types. The fourth rotation group of the original 1984 SIPP sample was excluded from the longitudinal file because data were not collected for the group in the second wave. From the sample households included in the first three rotation groups, a total of 31,161 individuals aged 15 and older by the end of Wave 3 had data on at least one of the three waves. A total of 26,992 of these persons had data at all three waves.

Each person could have had up to four wage-earning jobs on each wave. Each job is represented by a Wage and Salary record which can be linked to a person in the file. CSHD imputations were made to a limited number of items on these Wage and Salary records. The empirical work reported here focuses on three categorical and two continuous variables from the Wage and Salary record for which CSHD imputations were made where needed. The categorical variables were 1) occupation code, 2) employer category, and 3) frequency of pay. The continuous variables were the wage rate for hourly paid jobs and total monthly earnings for each of four reporting months in a single wave. Each of the three categorical and five continuous items (wage rate plus four monthly earnings) can be reported for each of three waves in the 1984 SIPP Panel. The merged data set contains longitudinal Wage and Salary records for a total of 23,005 job reports: 19,223 reports for individuals' first jobs; 2978 for the second jobs; 684 for the third jobs; and 120 for the fourth jobs. To simplify the presentation, results from only the first job are given.

Table 1 presents counts of item responses, both total and missing, by wave for the Job 1 Wage and Salary variables of interest. Among these variables, the item missing data rates for the categorical items are very small, ranging from a low of 0.16% item missing data for the Wave 1 employer category variable to a high of 2.45% for the Wave 3 frequency of pay question. The percentages of item missing data among earnings items are higher, particularly at the first wave of data collection: 9.37% item missing data for reports of Job 1 monthly earnings in Wave 1 of the 1984 SIPP Panel. However, item missing data rates for Job 1 monthly earnings drop substantially at Waves 2 (2.97%) and 3 (3.16%).[1] At 9.7%, the Wave 1 item missing data rate for hourly wage reports is also relatively high but, unlike Job 1 monthly earnings, the missing data rate for this variable rises slightly at Waves 2 and 3.

Table 1

Item Nonresponse in the 1984 SIPP Wage and Salary Data

Job 1 Variable	Wave	Item Responses	Imputed Values		Percent of missing values for which longitudinal imputation is possible
			Number	%	
Occupation	1	17,110	90	.53	56.6
	2	15,766	101	.64	85.1
	3	15,196	85	.56	78.8
Employer Category	1	17,110	62	.36	70.7
	2	15,766	44	.28	86.4
	3	15,196	25	.16	76.0
Pay Frequency	1	17,110	316	1.85	56.9
	2	15,766	362	2.30	75.1
	3	15,196	373	2.45	82.5
Monthly Earnings*	1	68,440	6,410	9.37	68.9
	2	63,064	1,875	2.97	41.1
	3	60,784	1,918	3.16	61.1
Hourly Wage	1	10258	993	9.70	39.4
	2	9476	1103	11.60	54.5
	3	9141	1038	11.40	57.2

*Item response totals are 4 monthly responses for each wave.

It is important to know not only the rate at which responses are missing but also what proportion of these missing values can be imputed longitudinally. LDS imputation is possible only when the missing item has actually been observed at a preceding or succeeding wave. Table 1 also indicates the extent to which missing items can be imputed longitudinally. Among the categorical variables, the percentage of missing responses which can be imputed by direct substitution from another wave ranges from 56.5% to 86.4%. Similarly, longitudinal imputation of missing data on the earnings items appears promising. For example, almost 69% of the missing values for Job 1 monthly earning at Wave 1 could be imputed using the LDS procedure.

V. Comparison of the CSHD and LDS Imputation Methods

Once the LDS imputations were made, the effect of the CSHD and the LDS imputations on distributions of the categorical and continuous variables of interest could be examined. In this section simple frequency distributions and distributions of change in individual reports from one wave to the next are compared between CHSD and LDS imputed values for each of the variables of interest. *The LDS method examined here uses the original CSHD imputed value whenever a substitute value was not available on an alternate wave. Thus, results for the LDS method will incorporate a proportion of missing value cases which were imputed by the secondary CSHD technique.*

Due to the very low rates of item missing data, CSHD and LDS imputations should not be expected to have widely differing effects on the overall frequency distributions of the categorical variables. For Job 1 Wave 1 employer category and frequency of pay, there is no difference in the distributions whether the CSHD or the LDS method is used to impute for missing data. Similar results were observed for Waves 2 and 3 and for other categorical type variables.

The categorical variables are essentially job descriptors, and given that the job was not changed, their values should not be expected to change significantly from one wave to the next. However, Table 2 indicates that, even in instances where no imputation is involved, a wave to wave change in response value for these variables can occur in as many as 20% of cases. It is difficult to say what proportion of this observed change is real, as opposed to a reflection of response error or coding inconsistency.

If cases where one or both values have been imputed are compared to cases without imputations the CSHD imputations lead to a significant reduction in wave to wave response consistency. On the other hand, the LDS imputation method produces a high level of cross-wave consistency for these job descriptors. In fact, one might view the LDS method as overriding the observed natural variation in responses and thereby forcing an artifically high level of wave to wave consistency.

The drop in cross-wave consistency for the job records with CSHD imputed values is so large that it suggests that the level of agreement across waves might be explained by "chance" alone. In the case of the hot-deck method, a discrete response category is modeled as an ANOVA-type function of a series of categorical factors (e.g., hot deck variables such as age, sex, race, education). If the model is weak, the probability of a correct imputation degenerates to the multinomial probability of agreement between the true value and a "random" imputation. The greater the number of response categories and the more uniform the odds across categories, the more difficult it is to impute the correct (or matching) value. The Wage and Salary categorical variables for which the CSHD imputation results in high wave to wave consistency do in fact have either few categories or highly unequal odds across categories.

Table 2

Wave to Wave Consistency in Categorical Variable Values Under the CSHD and LDS Imputation Methods

Job 1 Variable	Wave Comparison	No Imputation		One or Both Waves Imputed		
		n	% Agreement	n	CSHD % Agreement	LDS % Agreement
Frequency of Pay	1 to 2	14,079	81.3	475	45.8	89.3
	2 to 3	13,111	79.8	478	46.8	94.7
Employer Category	1 to 2	14,477	95.6	77	77.9	97.4
	2 to 3	13,546	95.4	43	69.7	100.0
Occupation Code	1 to 2	14,425	78.4	129	26.4	72.1
	2 to 3	13,470	78.8	119	19.3	78.8

For example, the employer category variable with six response categories has as the largest category "private company" with 82% of the cases. By simply imputing the code value for this largest category to each missing item, we might expect to be correct about 82% of the time. For this variable, even a random imputation of respondents' values will, in expectation, impute a matching value 69% of the time. In Table 2, a two-wave comparison involving CSHD imputations for this variable shows 78% agreement from Wave 1 to 2 and 70% agreement from Wave 2 to 3.

Although the small sample sizes and limited set of variables prevent us from drawing any firm conclusions, the data suggest that the CSHD imputation of these job descriptors provides only small increases in accuracy relative to what we might expect by chance alone.

Considering the continuous variables, Tables 3 and 4 compare characteristics of the earnings variables after CSHD and LDS imputations have been made for item missing data. Table 3 compares the sum of up to four monthly Job 1 earnings values for each wave of data collection; Job 1 hourly wage rates are compared in Table 4. The upper panel of each table presents findings for all cases, both those with nonmissing data and those where a missing amount has been imputed. The lower panel of each table presents only those cases where one or more component earnings amounts have been imputed.[2]

Table 3

Imputation of Job 1 Earnings. Comparison of Sample Earnings Distributions After CSHD and LDS Imputation for Item Missing Data

Wave	Statistic	Imputation Method	
		CSHD	LDS
All Job 1 Reports			
Wave 1	Mean	$4,796	$4,750
(n=16,895)	Std.Dev.	4,199	4,140
	Skewness	1.94	1.94
	Kurtosis	6.95	7.00
Wave 2	Mean	$5,041	$5,051
(n=15,569)	Std.Dev.	4,222	4,239
	Skewness	1.94	1.96
	Kurtosis	6.85	7.03
Wave 3	Mean	$5,128	$5,142
(n=14,994)	Std.Dev.	4,260	4,294
	Skewness	1.88	1.93
	Kurtosis	6.54	6.88
All Job 1 Reports With One or More Imputed Amounts			
Wave 1	Mean	$5,510	$5,225
(n=2,688)	Std.Dev.	4,491	4,174
	Skewness	2.32	1.70
	Kurtosis	8.23	9.38
Wave 2	Mean	$7,576	$7,896
(n=485)	Std.Dev.	6,115	6,343
	Skewness	1.70	1.78
	Skewness	3.60	3.86
Wave 3	Mean	$7,447	$7,859
(n=494)	Std.Dev.	5,943	6,485
	Skewness	1.73	1.82
	Kurtosis	3.62	3.94

The basic and not unexpected result found in Tables 3 and 4 is that even with item missing data rates of almost 10% at Wave 1, the choice of CSHD or LDS imputation appears to have only a small effect on the statistics examined.

The findings in Tables 3 and 4 indicate that univariate analyses of the SIPP Wage and Salary earnings data will not be greatly affected by the imputation methodology that is used. However, the data presented here give no indication of the effect these imputation methods have on univariate distributions for population subclasses or domains. Furthermore, the result for descriptive univariate statistics has no implicit generalization to bivariate and multivariate analyses.

One form of longitudinal analyses of SIPP data is to examine how and why individual income and earnings change over time. For this kind of analysis, information is needed on the effects of CSHD and LDS imputations on distributions of micro-level change in earnings. Table 5 presents the distribution of Wave 1 to 2 and Wave 2 to 3 changes in individual respondents' Job 1 earnings. Columns (3) and (4) compare the change distributions for all cases (actual and imputed) having a nonzero earnings amount at each wave. Column (5) restricts the change distribution to cases where two actual reports were obtained. Sample distributions of change involving actual-imputed and imputed-actual combinations of values are described in columns (6) – (9).

Over time, it is expected that the average wages and earnings of panel respondents should follow an increasing trend. Looking at Table 5, the overall distribution of change (columns 3 and 4) does show, as expected, a positive increment in Job 1 earnings between successive waves. For the Wave 1 to 2 change, the average amount of this increase is appreciably lower when the CSHD method is used to impute for item missing data. Examination of standard deviations and percentiles shows that CSHD imputation both increases the variability and elongates the tails of the sample distribution of wave to wave change in earnings. In fact, if the change computation is restricted to pairs with one CSHD imputed value and one actual response, the result is a distribution which is highly variable and has many extreme values. Because the number of extreme changes imputed by the CSHD method is so large, the distributional statistics – particularly the means – reported in Columns (6) – (9) should be viewed as highly unstable. These statistics are reported here primarily as evidence of the variability which CSHD imputation can introduce to longitudinal measures such as change in earnings.

Given that a zero change model is implicit in the direct substitution imputations used in this exercise, the LDS method should be expected to compress the wave to wave change distribution about the zero value. In comparing differences between actual and imputed values, columns (7) and (9) indicate that the LDS method of imputing averages of actual values for a missing earnings report results in changes which average just slightly greater than zero. (An exception occurs in the estimates of change between Wave 2 actual and Wave 3 imputed values.) The "compression" effect which the LDS method has on estimates of change is evident in a comparison of percentile statistics for the change distributions. For example, in the sample distribution of change between Wave 2 CSHD-imputed and Wave 3 actual values, the 5th and 95th percentiles are –$10,082 and $8,939. For cases where

Table 4

Imputation of Job 1 Hourly Wage Rates. Comparison of Sample Income Distributions After CSHD and LDS Imputation for Item Missing Data

Wave	Statistic	Imputation Method	
		CSHD	LDS
All Job 1 Reports			
Wave 1	Mean	$6.58	$6.60
(n=10,456)	Std.Dev.	3.67	3.67
	Skewness	2.23	2.21
	Kurtosis	16.27	16.27
Wave 2	Mean	$6.73	$6.76
(n=9,416)	Std.Dev.	3.84	3.84
	Skewness	3.41	3.38
	Kurtosis	37.74	37.96
Wave 3	Mean	$6.75	$6.76
(n=9,078)	Std.Dev.	3.92	3.79
	Skewness	4.23	3.33
	Kurtosis	61.33	42.96
All Job 1 Hourly Wage Reports With One or More Imputed Amounts			
Wave 1	Mean	$7.23	$7.34
(n=993)	Std.Dev.	3.81	3.73
	Skewness	1.40	1.25
	Kurtosis	2.29	1.73
Wave 2	Mean	$7.18	$7.39
(n=1,103)	Std.Dev.	3.99	3.97
	Skewness	2.24	2.09
	Skewness	13.23	12.46
Wave 3	Mean	$7.47	$7.57
(n=1,038)	Std.Dev.	4.96	3.98
	Skewness	6.32	2.33
	Kurtosis	83.97	15.50

Table 5

Wave to Wave Change in Job 1 Earnings. Comparison of Sample Distributions Under CSHD and LDS Imputation Methods

Change Estimate	Sample Distribution Statistic	All Data		Actual - Actual (No Imputation)	Actual - Imputed*		Imputed - Actual*	
		CSHD Imputation	LDS Imputation		CSHD Imputation	LDS Imputation	CSHD Imputation	LDS Imputation
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Wave 2 - Wave 1	Mean	$104.66	$165.43	$183.45	-$302.28	$24.65	-$948.75	$27.42
	Std.Dev.	2,561	2,243	2,205	2,902	1,679	6,913	1,091
	5th-%tile	-3,199	-2,766	-2,790	-4,490	-2,489	-12,640	-1,232
	95th-%tile	3,484	3,223	3,297	1,974	2,190	8,655	1,750
	(n)	(14,344)	(14,344)	(11,892)	(2,021)	(2,021)	(182)	(182)
Wave 3 - Wave 2	Mean	$139.97	$145.13	$148.69	-$214.76	$0.35	-$94.64	$124.09
	Std.Dev.	2,449	2,243	2,117	6,295	1,406	5,359	1,480
	5th-%tile	-2,974	-2,710	-2,659	-10,082	-1,917	-5,536	-1,709
	95th-%tile	3,400	3,199	3,149	8,939	1,766	-8,792	3,175
	(n)	(13,403)	(13,403)	(12,818)	(142)	(142)	(215)	(215)

*The CSHD imputations -- including default imputations under the LDS method -- produce a high degree of variability in the wave to wave change values. Therefore, the statistics reported in these columns are also highly variable and should be interpreted with caution.

earnings are not imputed at either wave, the 5th and 95th percentiles of the corresponding change distribution are −$2,659 and $3,149. The comparable percentiles for Wave 2 LDS-imputed to Wave 3 actual change are −$1,977 and $1,766.

VI. Concluding Remarks

Cross-sectional hot-deck (CSHD) imputation is a practical and timely method for imputing missing item values on the SIPP Wage and Salary record for an individual wave. However, the evidence presented here suggests that the CSHD method may perform only slightly better than chance at imputing the correct response to a missing categorical item from the wage and salary variable set. CSHD imputations for continuous wage and salary earnings variables do not appear to appreciably alter the distributions of these items. However, the impact on both cross-sectional and longitudinal multivariate distributions is larger.

Given the cross-wave patterns of item missing data observed in the 1984 SIPP Wage and Salary record, the use of longitudinal imputation methods appears to be warranted for SIPP longitudinal files. For categorical variables, the direct substitution method is a practical approach to cross-wave imputations of missing items. For the continuous variables such as Job 1 earnings, the empirical tests clearly demonstrate the desirability of longitudinal imputations for missing data on these items. The LDS method of longitudinal imputation understates change, but this may be preferred to the gross overstatement of change resulting from the use of the CSHD method.

References

McMillen, David B. and Daniel Kasprzyk (1985). "Item Nonresponse in the Survey of Income and Program Participation," *Proceedings of the Survey Research Methods Section. American Statistical Association*, 360–365.

Bureau of the Census (1985). *Survey of Income and Program Participation User's Guide*, 2nd edition. Washington: U. S. Department of Commerce.

Kalton, Graham (1985). "Handling Wave Nonresponse in Longitudinal Surveys," *Proceedings of the Bureau of the Census Research First Annual Conference*, pp. 453–461. Washington: U. S. Bureau of the Census.

Kalton, Graham and James M. Lepkowski (1982). "Cross-wave Item Imputation," in *Technical, Conceptual, and Administrative Lessons of the Income Survey Development Program (ISDP)*, Martin H. David, ed., pp. 155–170. Washington: Social Science Research Council.

Nelson, Dawn, David B. McMillen, and Daniel Kasprzyk (1985). "An Overview of the Survey of Income and Program Participation, Update 1," *SIPP Working Paper Series*, No. 8401. Washington: U. S. Bureau of the Census.

Footnotes

[1] A comparison of the total response counts across waves shows a decline in the number of cooperating respondents who hold Job 1, the sharpest drop occurring between Waves 1 and 2 of the panel. From one wave to the next, the change in the number of Job 1 reporters is a function of both panel attrition due to Type A (household) and Type Z (individual) nonresponse and responding individuals who no longer hold Job 1 at a later wave.

[2] Since earnings reports are taken for each month of the reference period it is possible to have actual and imputed values in the same wave. In such cases, the wave earnings totals will be the aggregate of actual and imputed monthly amounts.