THE SURVEY OF INCOME AND
PROGRAM PARTICIPATION

AN INVESTIGATION OF THE IMPUTATION OF MONTHLY EARNINGS FOR THE SURVEY OF INCOME AND PROGRAM PARTICIPATION USING REGRESSION MODELS

No. 15

V. J. Huggins and L. Weidman
Bureau of the Census

U. S. Department of Commerce BUREAU OF THE CENSUS

Survey of Income and Program Participation

An Investigation of
the Imputation of Monthly Earnings
for the Survey of Income and Program
Participation Using Regression Models

No. 8607 — 15

February 1986

Acknowledgement

This paper was prepared by Vicki J. Huggins and Lynn Weidman of the Statistical Research Division, Bureau of the Census.

Suggested Citation

Huggins, Vicki J. and Lynn Weidman. "An Investigation of the Imputation of Monthly Earnings for the Survey of Income and Program Participation Using Regression Models," Working Paper Series No. 8606. Washington, D.C.: U.S. Bureau of the Census, 1986.

TABLE OF CONTENTS

INTRODUCTION

This study using data from the first four waves of SIPP extends the previous ISDP study [1]. On each longitudinal record there are twelve months (3 waves) of data for several variables. In this phase we consider imputation of earnings for each of the four months of the second wave, assuming that all earnings for waves 1 and 3 are reported. A separate multiple regression model is fit for each month and months 5, 6, 7 and 8 are imputed in that order. This Is not a general study of imputation methods, but an investigation of the applicability of this particular approach. This work can easily be extended to consider other patterns of missing data. The study will be reported in several parts:

Preliminary analyses of some variables
Construction of files for model estimation and imputation
Estimation of models
Imputation Procedures (using estimated models)
Evaluation of imputation
Discussion of results
Suggestions for future research

Preliminary Analyses of Some Variables

The objective in preliminary analyses of the data was to gain knowledge about any peculiarities of the data, determine which variables to incorporate into our modeling of amounts of wages and salaries (referred to as earnings from hereon and denoted in formulas as EARN), and to ascertain how data should be separated to produce better models. Our previous research using ISDP data [1] suggested that splitting the records into subsets with similar variability should improve the overall quality of imputation using estimated models.

The SIPP data file used for data analyses, modeling and imputation contained 16,886 records. A record was available for all persons fifteen and older who had a job 1 record for the first three interviews and were a self or proxy interview for <u>all three</u> interviews. Rotation groups 1-3 interviews were from waves 1, 2 and 3 and rotation group 4's interviews were from waves 1,3 and 4. (For simplicity we will treat rotation group 4 as if it also was interviewed in waves 1, 2 and 3.) Approximately 10% of the file is from the nonwhite population and 20% of the file had at least one imputed earnings amount. The variables available for use in our models are: rotation group, usual hours worked, weeks with job 1, monthly earnings, pay frequency, weeks absent without pay for any job, age, sex, race and education.

Discussions with subject matter specialists indicated to us that changes in interview status (self, proxy) between waves may affect what was reported for earnings amounts. In order to examine this we looked at the change in reported earnings between the last month of a wave and the first month of the following wave. Two measures of change of earnings amount between waves were computed and their mean values examined for each of four groups identified by the possible combinations of interview status: (self, self) (self, proxy), (proxy, self) and (proxy, proxy). The measures of change used were:

(1) Ratio of earnings in the first month of a wave to the earnings in the last month of the preceding wave, i.e., EARN (5)/EARN (4) and EARN (9)/EARN (8).

(2) Difference between earnings in the first month of a wave and earnings in the last month of the preceding wave, i.e., EARN (5) - EARN (4) and EARN (9) - EARN (8).

The mean percentage difference computed for EARN (5)/EARN (4) in each of the groups defined by change in interview status ranged from 2.14 to 2.79 (See Table 1). The mean percentage difference computed for EARN (9)/ EARN (8) in each of the four groups ranged from 2.9 to 4.51. These measures do differ from group to group, but not in any detectable pattern. E.g., the group (proxy, proxy) has the largest mean percentage difference between EARN (8) and EARN (9), and the smallest between EARN (4) and EARN (5). Because of these results, interview status was not included in the models for earnings.

In previous modeling of amounts of wages and salaries using ISDP data, we determined that splitting the data into subgroups of records based upon the variability of the records' earnings responses should provide more accurate models on which to base imputation. To examine the variability of the data we first split the records into sixteen subsets based on every combination of rotation group (1-4) by sex (male, female) and race (white, nonwhite). Then we constructed 12 categories defined as follows: [0, .2], (.2, .5], (.5, .75],

TABLE 1

Average monthly difference in earnings between waves

interview statuses	m5-m4	m9-m8
11	-15.04	41.87
12	-89.53	36.38
21	110.11	44.17
22	10.90	87.18

Average montly ratio of earnings between waves

interview statuses	m5/m4	m9/m8
11	3.59	3.90
12	3.74	5.08
21	3.95	4.70
22	3.14	5.51

Key to interview statuses: 1st number is for first wave in calculation and 2nd number for second wave

1 = self interview

2 = proxy interview

Example: 12 for m9/m8 indicates self for m8 and proxy for m9

(.75, .9], (.9, 1], (1, 1.1], (1.1, 1.25], (1.25, 1.50], (1.50, 2], (2, 5], (5, 10], (10, B). Ratios of month-to-month earnings for a record, i.e., EARN (i+1)/EARN (i) for i=1,...11, were computed and counted according to which of the above twelve categories they fall in. Distributions of the maximum ratio that occurs on a record and the minimum ratio that occurs were examined. (See Tables 2 and 3).

The highest minimum ratio that occurs for any case is in the range (.9, 1] and the largest number of minimum ratios occur in (.75, .9]. The next largest number of minimum ratios occurs in [0, .2] which is largely a result of setting ratios to zero when zero earnings are reported for consecutive months. Ignoring this problem, the next largest number of minimum ratios occurs in (.5, .75]. For nonwhite females, the number of minimum ratios in (.75, .9] and (.5, .75] are approximately equal whereas all other cells have more minimum ratios in (.75, .9] than (.5, .75]. The distribution of maximum and minimum ratios suggests that the variability of the records based on earnings is spread across the twelve defined intervals enough to warrant subsetting the data records into more homogeneous groups that can be modeled separately. Only 2.5% of all the records have a maximum ratio greater than 5 which indicates that the problem of evaluating the worst cases and their overall effect on imputation is reduced in scope. It obviously makes sense to treat this extremely variable group differently from the rest of the population.

Similar crosstabulations of the variables WKS-ABSENT, the number of weeks absent from any job, NONRECEIPT, whether or not wages and salaries were received and WKS-W-JOB1, the number of weeks with job 1, were produced to determine their distributions, their relationships to each other and whether WKS-ABSENT and WKS-W-JOB1 would be useful in modeling NONRECEIPT of earnings.

Table 4 gives the total number of NONRECEIPT responses on the file for each rotation group-race-sex cell and suggests that there are enough NONRECEIPT responses to model; i.e., 25% of the file has a report of nonreceipt of earnings for at least one month. Some fraction, say 10%, of the white population with nonreceipt responses can be used to model nonreceipt of the white population and the entire set of nonwhite records with nonreceipt responses used to model for the nonwhite population. The remaining portion of the white population with nonreceipt responses can be used for imputation and subsequent evaluation. Also indicated in the distribution of NONRECEIPT is the fact that modeling by rotation group may not be necessary as frequency counts remain consistent across rotation groups.

TABLE 2

Distribution of Maximum Ratios

			upper bound of interval								
rotation group	sex	race	.9	1.0	1.1	1.25	1.5	2.0	5.0	10.0	∞
1.	male	white	6	98	197	632	526	306	247	36	21
		non-white	0	11	16	52	42	33	18	3	0
	female	white	20	102	179	482	391	304	246	33	18
		non-white	0	7	22	40	49	40	32	6	4
2.	male	white	5	121	188	587	514	334	265	36	12
		non-white	0	9	17	69	48	28	21	1	1
	female	white	19	108	145	439	430	271	239	28	12
		non-white	0	16	18	50	47	41	32	6	0
3.	male	white	8	104	197	654	474	319	270	37	14
		non-white	1	11	10	61	43	27	18	0	1
	female	white	9	93	201	505	385	274	240	39	4
		non-white	1	14	28	74	60	26	30	4	0
4.	male	white	13	104	217	603	539	324	291	38	9
		non-white	0	6	16	65	48	33	19	3	0
	female	white	12	92	170	475	446	262	245	20	11
		non-white	0	19	19	48	54	37	22	4	0

TABLE 3

Distribution of Minimum Ratios

upper bound of interval

rotation group	sex	race	.2	.5	.75	.9	
1.	male	white	452	168	494	630	325
		non-white	40	13	31	71	20
	female	white	513	143	387	458	274
		non-white	55	16	54	49	26
2.	male	white	438	187	471	658	308
		non-white	35	20	43	70	26
	female	white	532	139	356	446	218
		non-white	49	30	40	58	33
3.	male	white	447	193	517	632	288
		non-white	26	17	41	65	23
	female	white	518	134	380	435	293
		non-white	64	28	56	56	33
4.	male	white	392	181	474	734	357
		non-white	30	13	49	76	22
	female	white	450	137	391	518	237
		non-white	44	17	50	65	27

TABLE 4

Total number of NONRECEIPT respondents by
Rotation group - sex - race

		white	non-white
Rotation group 1	male	419	38
	female	493	46
Rotation group 2	male	410	34
	female	508	46
Rotation group 3	male	407	25
	female	500	61
Rotation group 4	male	370	27
	female	431	41

WKSABSENT refers to any job the person may have held, not just job 1 to which all the other variables refer. An indicator in the data that points out when WKSABSENT is relevant to job 1 could be of some use, however this information is not available on this particular data file. It is also difficult to gauge how valuable a variable WKSABSENT is because of the low percentage of weeks absent reported. Since it refers to any job, not just job 1, the percentage should be at least as high as the percent of nonreceipt reported for job 1, but it is not. This seems to indicate that WKSABSENT would probably not be very helpful in modeling NONRECEIPT of earnings.

The distribution of WKS-W-JOB1 shows 35 cases when persons supposedly with job 1 reported zero weeks with job 1. 20% of the cases were with job 1 for less than a full twelve months. WKS-W-JOB1 could be helpful in modeling NONRECEIPT or earnings amounts.

As a result of previous work completed using ISDP data, it appears that very little extra information is actually gained from incorporating other variables on the file into the modeling of NONRECEIPT and EARN and this seems to be basically verified here in our preliminary data findings. Therefore, the longitudinal information of NONRECEIPT and

EARN must be exploited. We have done this for EARN by splitting the data into similar subgroups defined by month-to-month variability. Age and education are the only other variables used.

In addition to examining the variability of the records across all twelve months of responses, we needed to know how earnings differed between waves on a record. Table 5 gives the percent of records that increase and decrease in earnings between waves 1 and 2 (months 4 and 5) and waves 2 and 3 (months 8 and 9). The percent of records that increase is approximately the same in the transition from wave 1 to 2 as that from wave 2 to 3. Similarly, the percent of-decrease of EARN for both wave transitions is equal. However, the percent of records that increase for both wave transitions is about 50% but the percent of records that decrease for both is only about 35%. Also, the average decrease between waves is a bit larger than the average increase.

TABLE 5

Changes in Earnings Between Waves

	WAVES 1 & 2	WAVES 2 & 3
% of records that increase in earnings	51.77	50.27
% of records that decrease in earnings	35.64	35.19
Average increase	$300.84	$279.15
Average decrease	$325.00	$332.00

We then went one step farther in splitting the data to model EARN. We removed records that had constant values within waves to later analyze using the Table 5 information about increases and decreases between waves. (Results of this analysis are not presented in this report.) We also excluded all records that contained at least one imputed value for EARN 3,419 records had at least one impute. We created subfiles based on rotation group, sex, race and month-to-month earnings ratios.

Construction of Files for Model Estimation and Imputation

From our work with the ISDP, we know that different data distributions occured for different race by sex combinations. Therefore, our first separation of the data was into three groups: white males, white females and non-whites. There were not enough non-whites on this file to allow them to be separated into groups of males and females. The second separation was by the variability in monthly earnings for the 12 months on a record. This was also suggested by our work on ISDP. There are people whose month-to-month changes in earnings follow different patterns: some are fairly constant, some are variable but not unusually so and others jump around quite a bit. Thus we want to separate the data into at least three variability groups before working with it.

There are two ways which we considered for measuring variability. One is the average variance within waves and the other is the maximum and minimum ratios between consecutive months. The first measure is based on absolute differences in earnings and the second on relative differences. For this first set of imputations, we used the max/min ratios for our definition of variability. If there were enough records in a rotation group x race x sex combination we separated it into three groups by (i) the value of the maximum or minimum ratio between adjacent months if there are no zero earnings or (ii) maximum reported earnings if there are some zeros.

Group 1 - records with max and min ratios between .75 and 1.25 and records with some 0 earnings, $100 reported maximum

Group 2 - records with min ratio between .2001 and .75 or max ratio between 1.2501 and 5, not in group 3 and records with some 0 earnings, $300 reported maximum

Group 3 - records with min ratio less than .2 or max ratio greater than 5, and all remaining records with some 0 earnings

When this has been completed, we have the counts in each of the files for rotation groups 1 and 4 as given in Table 6. We will refer to the records in these three groups as consistent, semi-consistent and erratic, respectively.

TABLE 6

	White male		White female		non-white	
	R1	R4	R1	R4	R1	R4
Consistent	313	280	250	182	65	58
Semi-consistent	647	681	553	566	145	144
Erratic	378	315	431	370	100	75

Ri = rotation group i

Regression models to be used for imputation are to be estimated and evaluated for each of these 18 files. It is desirable to evaluate the imputation performance of a model on a different set of data than is used for model estimation. If there were "enough" records on a given file it was separated into two further files - an estimation file and an imputation file. A model was obtained using the estimation file and earnings for the months of wave 2 were imputed onto each record of the imputation file. Both the reported and imputed values are then available on this file for evaluating the success of the imputation procedure. If there are too few records for this separation, we must both estimate and impute using all the available records. This was the case for the non-white files. For the others, if 400 or more records were available, then we used 200 for estimation, otherwise we used about one-half of the records for estimation.

The estimation files were randomly generated from the available records using the sampling procedure in SPSS. (Results of model fitting and imputation evaluation indicate that a more careful selection of imputation files is needed. The samples should be stratified according to the distribution of max and min ratios in the population.)

Estimation of Models

Although the feasibility of modeling NONRECEIPT was discussed in the analysis section, we have only looked at earnings in this study. Of all the available data files, we will present and discuss results only for white males in rotation groups 1 and 4. The reason for this is the large amount of time required to produce the graphical displays used for comparison and evaluation. Looking at two different rotation groups should indicate the value of future work in this area.

The earnings in a given month are modeled as a multiple linear regression on four sets of variables.

Set 1: Monthly earnings in the four preceding months and the four following months (if available) (Since we are looking at the case where all months of wave 2 are missing, any following month that is in wave 2 is not available.)

Set 2: Same as set 1 except that earnings are divided by weeks with job 1 minus weeks absent from any job. If this difference is 0, the ratio is set to 0.

Set 3: Age divided into five categories: (15,24) (25,34) (35,54) (55,64) (65+)

Set 4: Education divided into three categories: (0,11) (12,14) (15+) years completed

The models are fit using the statistical package GLIM. First sets 1 and 2 are fit and terms that test as being insignificant are removed. Then age and education are fit, each being retained if it reduces the error sum of squares by a substantial amount. The resultant models are given in Appendix A. (Before determining final models the residuals should be adjusted for outliers and perhaps these cases should be removed and the models re-estimated.)

Imputation Procedures

For each record in an imputation file we will impute earnings sequentially for months 5 through 8. Values imputed for previous months will be used for determining an impute for the current month. Let $\underline{b}$ be the vector of parameter estimates for a particular month and model, X be the design matrix used to calculate $\underline{b}$, and $\hat{\sigma}^2$ be the mean square error. Let $\underline{x}_o$ be the vector of regressor variables corresponding to $\underline{b}$ for a given record. We have imputed monthly earnings values using two methods.

mean imputation: Impute $y_m = \underline{x}_o' \, \underline{b}$

random imputation: Select a normal (0,1) random deviate z.

$$\text{Impute } y_r = \underline{x}_o' \, \underline{b} + z \left[\underline{x}_o' \, (X'X)^{-} \underline{x}_o \hat{\sigma}^2 + \hat{\sigma}^2 \right]^{1/2} .$$

The expected value of the impute for each of these methods is $\underline{x}_o' \, \beta$, where β is the true population parameter vector. The difference between these methods is that mean imputation imputes the same value for y_m each time a record has the regressor variables $\underline{x}_o$, but random imputation imputes a value y_r which depends on the z generated. The reason for using random imputation is to better represent the variability from the observed population. There are two types of variability here: variance among earnings in a given month and variance across the four months of wave 2 for each record. The particular random component used here is not necessarily the one to be used in practice, but it does give average variances across wave 2 imputes that are within 20% of the observed average variance across waves 1 and 3 for four of the five imputed files. These same average variances for mean imputation are roughly 20% to 40% of the observed.

It is not our purpose to compare these two methods to determine which is better. The evaluation of mean imputation will be discussed and the differences between the results of the two methods as seen in the tables and figures in the appendices will be pointed out. Evaluation procedures are discussed in the next section.

Evaluation of Imputation

There are different means of evaluating the success of an imputation procedure. Which ones are most appropriate are determined by the goals of the imputation process. Is it important to duplicate earnings patterns for the whole wave on a record, or just for each month, or for the difference between successive months, or for the difference between other pairs of months? Do we want to look at the marginal distributions of imputed patterns or their joint distribution with observed patterns?

Most of the measures and distributions we have computed use combined data from the four months of wave 2. When looking at these results one must remember that the results come from four separate estimated models. If just one of these is bad it can cause the overall imputes to look worse than they actually would if further work was done to improve the bad model. For example, there may be some outliers that should be removed before estimating a model for a given month. A more thorough examination of the connection between modeling and imputation needs to be carried out before we can have full confidence in our evaluation procedures. This should be kept in mind as one looks at the comparison of observed and imputed values.

There are two properties of the imputations that we want to examine: closeness to the observed data and variability compared to the observed data. Closeness is a function of how well the imputed values maintain the distributions of the observed data. Variability is a function of the month-to-month change in the observed data and the size of random errors used in random imputation. The definition of variability we are using here is the estimated variance for the four observations in a wave. There are many statistics that can be used to measure the closeness of imputed and observed values. Those that we have computed are:

(M1) Difference between observed and imputed values;

(M2) Percentage difference between observed and imputed values;

(M3) Difference in ratios of consecutive months' earnings between observed and imputed values;

(M4) Percentage difference in ratios of consecutive months' earnings between observed and imputed values.

In Appendix B we give the mean and the square root of the mean squared values of each of these measures, where sums are taken over all the imputed months on all the records in the indicated group. If c_i is one of the measures M1-M4 for a given month on a given record, then $\sum c_i/N$ is the mean and $(\sum c_i^2/N)^{1/2}$ is the square root of the mean squared measure. N is the number of terms in the summation. (See Table B.1) Also given are histograms of observed and imputed values as well as scatterplots of observed vs. imputed values. The imputations are supposed to capture the overall distribution of the missing values, not necessarily the record by record distribution, so we want the histograms of the imputed values to match those of the observed. The scatterplots of imputed vs. observed show how close the imputed and observed values are for individual cases. The measures of closeness are based on these individual comparisons of imputed and observed values for each month in wave 2.

For the comparison of variability we look at wave variances via histograms and scatterplots. Let v_2 be the variance for wave 2. Appendix C contains histograms of v_2

for observed and imputed data. Random imputation is supposed to more truly represent the population variation than mean imputation and if true, this should show up in these histograms.

When looking at these comparisons we note that no adjustment for unusually large c_i's has been made. These values are included in the $(\sum c_i^2/N)^{1/2}$ computations and a single very large value will distort this statistic. This should be kept in mind when reading Table B.1 and the discussion in the next section.

Discussion of Results

A. Closeness of Imputed to Observed Values for Wave 2

The statistics $(\sum c_i^2/N)^{1/2} = \overline{\text{dev}}$ for M1 - M4 given in Table B.1 are measures of the average closeness of each imputed value to its corresponding observed value. Probably the most revealing measure is $\overline{\text{dev}}$ for M2. For consistent groups and MI the measure is .196 and .104, and for the semi-consistent groups it is .365 and .440. These latter values may be unacceptable and .196 seems large for the consistent group while .104 seems good. However, we don't know how well competing procedures can perform. In all cases but two (M3 and M4 for Rotation group 1 - consistent) this measure is less for mean imputation (MI) than for random imputation (RI).

Further comparisons of imputed and observed values are presented graphically in Appendix B. The histograms represent the marginal distributions of observed, MI and RI values for the four months of wave 2 combined and show the following:

consistent records: Distributions of MI are very close to those for observed values.

semi-consistent records: Distributions of MI very close to the observed but imputes have too many values in (0, 100).

erratic records: MI has too few values at 0, but is otherwise very close to the observed.

The scatterplots represent the bivariate distributions of observed and imputed values. We do not expect the points to lie too close to a line with slope 1, but we can see how the imputes vary about different observed values. Because we divided the records according to variability, this shows up in the horizontal spread of the imputed values about each observed value.

consistent records: MI lies closely about the slope 1 line.

semi-consistent: Much more variable than for consistent records.

erratic records: Lots of variability. Note particularly the spread of imputes along the horizontal axis for observed values of 0.

Overall, the marginal distributions of MI appear close to those of the observed. The lack of imputes near zero for the erratic records points out the inability of MI to match the observed variability. RI has marginal distributions close to those of MI, with too many imputes near zero. The scatterplots for RI become increasingly spread out compared to MI as the variability of records increases, just as expected.

B. Monthly Comparisons

Recall that the months on a record are imputed successively with past imputes being used in computing the current month impute. This means that for a given month the closeness of the distribution of imputes to the distribution of observed values depends on this closeness for the previous months' imputes. (Imputes use a model based on the observed distributions of monthly values. If the imputed values for a month have a different distribution than the one observed, then they will also contribute to making the current month imputations differ from the observed.) Therefore, we want to look at how well each individual month is being imputed. How well does the distribution of imputes for a given month follow the observed distribution? Another concern is about the relationship of imputed values to observed values in waves 1 and 3. How similar are imputed and observed distributions for the differences in earnings between each month in wave 2 and month 4 and between each month in wave 2 and month 9?

Histograms of the distributions of monthly differences for MI and RI are presented in Appendix D. (Histograms for monthly values are not given because they look like the histograms in Appendix B for wave 2. There is no major change in their shape from month to month. Statements made about these distributions in the previous section are also applicable for each month.) If the histograms for an imputation procedure and observed values are very similar, this procedure gives us some of the information we want to have. There are many more distributions of this type that can also be examined, e.g. the joint distribution of month 4 values and the difference between months 7 and 4. However, which distributions are important depend on the goals of imputation.

In most cases the mean monthly differences are more regative for MI than for the observed values. This negative bias is apparent in the histograms for monthly differences. The MI histograms usually appear to be negatively shifted one interval compared to those for observed values and become less close to their shape for less consistent records. (Months 4 and 9 have essentially the same distribution of earnings. Thus the difference between any month in wave 2 and month 4 is the same as the difference between that month and month 9, and we don't need to look at these latter differences. Also, the absolute value of the mean of the difference between the imputed and observed value in any month of wave 2 is the same as the absolute value of the mean of the differences between that month and month 4 or month 9.) These monthly differences are distributed much more uniformly across several intervals for RI. This is due to the random component spreading out values from the mean imputations.

C. Variability of Imputed Values

Appendix C contains histograms of the variability of wave 2 values for observed, MI and RI. If e_i = earnings in month i of wave 2, then for a given record var2 $= \sum_{i=1}^{4} (e_i^2 - \bar{e})^2/3$, where $\bar{e} = \sum_{i=1}^{4} e_i/4$. The reason for looking at var2 is to see if the month-to-month variation is similar for imputed and observed values. This variability might help us to choose between two methods that were similar in measures of closeness and distribution.

Theoretically we expect MI to give variances that are much smaller than and RI to give variances that are closer to the observed variances. This is because the estimated models fit the average change from one month to the next and MI imputes average changes, even for cases with "extreme" month-to-month variation. RI adds a random component that attempts to mimic the variation of month-to-month changes about the average change, and occasionally imputes "extreme" variation.

The histograms present detailed distributions of var2. MI behaves as expected with too many cases having small variances. RI compares more closely with the observed, but always has too few cases that are too small for the plotting range. Instead, RI has a too spread out concentration in the pictured range. Table C.1 shows that the mean var2 for RI is usually close to the observed.

D. Summary

The histograms for individual months and combined months of wave 2 all have very similar distributions for MI. These distributions have shape close to those for observed values. MI does differ from the observed near zero for records not in the consistent group. For monthly differences the distributions are much more concentrated and show a small negative shift for MI compared to the observed. This negative bias also can be seen from the mean differences in Tables D.1. The distributions of variances show that MI severely underimputes the wave 2 variance. What these results show is that the distribution of MI values for each month is very similar to the observed distribution, but the patterns of variation for individual records across wave 2 are not.

The purpose of using RI is to improve this variation pattern. At the same time we hope that the monthly distributions will be maintained. Table C.1 shows that in all cases but one the mean var2 for RI is within 20% of the observed mean var2. The histograms in Appendix C show that the distributions of var2 are similar with RI not being as peaked near zero as the observed. The monthly histograms for RI are similar to those for MI except for an abundance of zero imputes, but the monthly difference histograms are not as peaked. In almost every case the measures of closeness M1-M4 are smaller for MI than for RI.

On the whole it seems as though MI and RI both do a fairly good imputation job for the marginal distribution of monthly values. There are ways to improve these procedures that will be proposed in the next section. When an improved regression procedure is determined it should prove very satisfactory with respect to most or all of the evaluation measures discussed here. Any other proposed imputation procedure should be compared to it before making a decision on the use of the proposed one.

Suggestions for Future Research

No effort was made to "fine tune" the models used in this study, so some improvements in the imputations could be made by making such an effort. This would include (1) improvement of the sampling method used to derive an estimation file and (2) adjustment for outliers in model fitting. By doing (1) we could assure that the distribution of observations in the estimation file follows closely the distribution of the file as a whole. Using (2) we can reduce the effect of influential observations in overall fitting. The combination of these procedures will produce a model that better fits the distribution of observed values.

When looking at the observed distribution of earnings one sees that they are not normally distributed (as assumed for our methods) but have positive tails. The model fitting and the random variables generated for RI may both be improved by transforming earnings so that they have more nearly a normal distribution. Estimation and imputation can then be performed on the transformed values and the results retransformed, with bias correction, to the original scale. This may reduce the negative bias in monthly differences for MI and some of the excess spread in monthly differences for RI.

A more significant improvement would probably be made by treating the four monthly earnings in wave 2 as having a multivariate distribution. Then a single multivariate regression model would be estimated rather than four separate univariate regressions. Such a model would better represent the simultaneous relationships among monthly earnings than do our current models. RI would then use a random vector for representing this. The general imputation situation is that a single month is missing from a wave. This situation would require a univariate model for each month, and that is why we have used them rather than multivariate models. By looking at month 5 we can get an idea of the success of univariate models for a single month. However, missing waves occur much more often than missing months and the multivariate approach should be examined.

Four areas for future work have been suggested: sampling improvement, outlier adjustment, data transformation and multivariate modeling. The expected returns in pursuing these areas differ but one question remains--how are the resulting imputations to be evaluated? Several ways of evaluating imputation have been presented and discussed. An effort should be made to determine which are the most important ones for SIPP before further work is carried out.

Reference

[1] Huggins, V. and Weidman, L. An Investigation of Model-Based Imputation Procedures Using Data from the Income Survey Development Program, July 1985.

APPENDIX A

Estimated Models - White Males

Table A.1

Consistent - Rotation Group 1

Month 5

Estimate	S.E.	Parameter
.092	.078	me1
.23	.15	me2
3.37	.56	me3
-.79	.18	me4
.46	.043	me9
-1.07	.55	ne2
-15.09	2.75	ne3
3.62	.63	ne4
-30.99	42.45	age1(1)
-102.3	38.72	age1(2)
-40.77	40.96	age1(3)
-83.00	41.53	age1(4)
-288.6	161.6	age1(5)
100.7	35.33	ed(2)
89.29	35.97	ed(3)

MSE = .1798 x 10^5

Month 6

Estimate	S.E.	Parameter
.30	.14	me2
4.88	.88	me3
-.61	.18	me4
.11	.079	me5
.46	.043	me9
-.85	.45	ne2
-24.62	4.38	ne3
3.82	.67	ne4
-31.67	42.42	age1(1)
-102.2	38.73	age1(2)
-40.68	40.97	age1(3)
-82.95	41.55	age1(4)
-288.1	161.7	age1(5)
100.5	35.35	ed(2)
89.29	35.99	ed(3)

MSE = .1800 x 10^5

Month 7

Estimate	S.E.	Parameter
2.05	.83	me3
-.50	.24	me4
.96	.13	me5
-.21	.074	me6
.43	.072	me10
-.30	.10	me11
-10.79	4.11	ne3
2.97	.96	ne4
72.35	57.14	age1(1)
57.51	54.11	age1(2)
84.71	56.23	age1(3)
105.01	61.07	age1(4)
191.70	74.19	age1(5)
-62.09	49.08	ed(2)
-71.31	49.34	ed(3)

MSE = .3337 x 10^5

Month 8

Estimate	S.E.	Parameter
-.22	.046	me4
.82	.054	me5
-.055	.039	me6
.14	.045	me9
.059	.042	me11
1.02	.16	ne4

MSE = .7389 x 10^4

Table A.3

Semi-Consistent Rotation Group 1

Month 5

Estimate	S.E.	Parameter
.48	.24	me1
.71	.34	me2
.83	.30	me4
.90	.25	me9
-2.61	1.12	ne1
-2.30	1.34	ne2
.54	.34	ne3
-2.50	1.24	ne4
-1.71	1.25	ne9

MSE = .1783 x 10^6

Month 6

Estimate	S.E.	Parameter
.32	.11	me4
1.03	.35	me5
.59	.25	me9
.23	.050	me10
-.79	.36	ne4
-3.10	1.37	ne5
-1.14	1.22	ne9
-217.42	100.81	age1(1)
-129.13	99.12	age1(2)
-334.11	125.16	age1(3)
-85.45	100.56	age1(4)
81.26	293.05	age1(5)
90.05	93.98	ed(2)
199.66	99.19	ed(3)

MSE = .1713 x 10^6

Month 7

Estimate	S.E.	Parameter
-193.91	83.28	mean
.35	.27	me3
-1.07	.46	me4
.99	.66	me5
.55	.11	me6
-.12	.14	me9
-2.07	1.37	ne3
4.52	1.90	ne4
-2.13	2.66	ne5
-.45	.29	ne10
1.89	.46	ne11

MSE = .3393 x 10^6

Month 8

Estimate	S.E.	Parameter
-91.50	43.76	mean
.27	.064	me5
.046	.038	me7
.20	.038	me10
2.14	.30	ne6
.36	.18	ne12

MSE = .1055 x 10^6

Table A.5

Erratic - Rotation Group 4

Month 5

Estimate	S.E.	Parameter
.16	.13	me1
-.23	.16	me2
.61	.22	me3
.32	.17	me4
.17	.05	me9
-1.10	.87	ne3
-.56	.61	ne4
-39.51	124.49	age1(1)
10.38	156.95	age1(2)
472.65	208.04	age1(3)
256.34	193.12	age1(4)
141.72	357.26	age1(5)
163.25	145.38	ed(1)
295.31	161.58	ed(2)

MSE = .4371 x 10^6

Month 6

Estimate	S.E.	Parameter
.43	.17	me2
.78	.18	me4
.22	.10	me5
-.24	.22	me10
-.84	.75	ne2
-1.63	.62	ne3
-1.40	.49	ne4
1.05	.64	ne9
.73	.75	ne10
-35.68	121.31	age1(1)
11.64	149.09	age1(2)
338.09	199.36	age1(3)
175.80	186.86	age1(4)
153.88	344.92	age1(5)
137.03	140.53	ed(2)
248.49	157.55	ed(3)

MSE = .4069 x 10^6

Month 7

Estimate	S.E.	Parameter
-.28	.11	me4
.83	.053	me6
.74	.32	ne3
1.19	.39	ne4
74.20	85.45	age1(1)
-166.86	105.93	age1(2)
232.87	144.63	age1(3)
37.58	179.53	age1(4)
1.12	135.86	age1(5)
89.82	97.99	ed(2)
-84.61	108.85	ed(3)

MSE = .2059 x 10^6

Month 8

Estimate	S.E.	Parameter
1.20	.13	me7
.29	.22	me11
-1.77	.34	ne5
-.83	.41	ne7
.52	.52	ne9
.93	.40	ne10
-1.15	.87	ne11
-202.60	113.79	age1(1)
126.90	134.31	age1(2)
-303.66	189.17	age1(3)
-25.51	252.46	age1(4)
-27.92	175.81	age1(5)
170.64	128.62	ed(2)
241.21	140.80	ed(3)

MSE = .3464 x 10^6

Appendix B

Comparisons of Observed and Imputed Values —

Four Months of Wave 2 Combined

Figures B.1-B.15 are available on request. Write to:

Daniel Kasprzyk, Special Assistant
Office of the Chief
Population Division, Rm. 2025-3
Bureau of the Census
Washington, D.C. 20233

Table B.1

DATA	Type of Impute	$C_i = x_i - x_i$		$C_i = \frac{x_i - x_i}{x_i}$		$C_i = r_i - r_i$		$C_i = \frac{r_i - r_i}{r_i}$	
		$\sqrt{\frac{\Sigma C_i^2}{N}}$	$\frac{\Sigma C_i}{N}$	$\sqrt{\frac{\Sigma C_i^2}{N}}$	$\frac{\Sigma C_i}{N}$	$\sqrt{\frac{\Sigma C_i^2}{N}}$	$\frac{\Sigma C_i}{N}$	$\sqrt{\frac{\Sigma C_i^2}{N}}$	$\frac{\Sigma C_i}{N}$
Consistent records from Rotation Group 1	Mean Impute	268.756	71.638	.196	.039	2.824	-.134	2.958	-.149
	Random Impute	304.383	49.933	.273	.015	1.959	-.137	1.959	-.146
Consistent records from Rotation Group 4	Mean Impute	223.565	27.132	.104	-.003	.122	.004	.118	-.009
	Random Impute	292.688	3.882	.170	-.022	.224	-.011	.227	-.026
Semi-consistent records from Rotation group 1	Mean Impute	888.980	6.438	.440	-.026	5.231	-.223	4.272	-.279
	Random Impute	1000.357	-64.703	.700	-.111	7.531	-1.092	7.590	-1.154
Semi-consistent records from Rotation group 4	Mean Impute	1,231.927	25.793	.365	-.031	1.533	-.099	1.521	-.156
	Random Impute	1,345.561	-27.847	.739	-.106	12.595	-2.113	13.399	-2.294
Erratic records from Rotation group 4	Mean Impute	1,064.113	97.082	43.370	-15.218	19.567	-1.833	19.528	-1.618
	Random Impute	1,380.645	-123.924	58.732	-19.327	28.503	-6.673	26.995	-5.197

Appendix C

Histograms of Wave 2 Variances

Figures C.1-C.9 are available on request. Write to:

Daniel Kasprzyk, Special Assistant
Office of the Chief
Population Division, Rm. 2025-3
Bureau of the Census
Washington, D.C. 20233

Table C.1

Means of Wave 2 Variances

Record Variability	Rotation Group	Observed	MI	RI
consistent	R1	22967	8099	27658
	R4	30122	8357	32685
semi-consistent	R1	263326	86822	219094
	R4	1274066	44998	276166
Erratic	R4	540614	103995	536371

Appendix D

Comparisons of Monthly Values

Figures D.1-D.18 are available on request. Write to:

Daniel Kasprzyk Special Assistant
Office of the Chief
Population Division, Rm. 2025-3
Bureau of the Census
Washington, D.C. 20233

Tables D.1

Mean Monthly Earnings

Table D.1.a.

Consistent Rotation Group 1

Month	Observed (0)	Mean Imputation	Random Imputation	0-MI	0-RI
5	2050	1982	2005	68	45
6	2127	1998	2009	129	118
7	2129	2093	2125	36	4
8	2083	2018	2052	65	31
		mean absolute difference		74.5	49.5

Table D.1.b

Consistent Rotation Group 4

Month	Observed (0)	Mean Imputation	Random Imputation	0-MI	0-RI
5	1833	1820	1843	13	-10
6	1910	1850	1851	60	59
7	1986	1976	2012	10	-26
8	1847	1822	1855	25	-8
		mean absolute difference		27.0	25.75

Table D.1.c

Semi-Consistent Rotation Group 1

	Observed	Mean	Random		
Month	(0)	Imputation	Imputation	0-MI	0-RI
5	1769	1804	1859	-35	-90
6	1879	1789	1846	90	33
7	1968	2054	2138	-86	-170
8	1861	1804	1893	57	-32
		mean absolute difference		67.0	81.25

Table D.1.d

Semi-Consistent Rotation Group 4

	Observed	Mean	Random		
Month	(0)	Imputation	Imputation	0-MI	0-RI
5	1878	1848	1900	30	-22
6	1905	1805	1838	100	67
7	1947	1995	2053	-48	-106
8	1846	1826	1898	20	52
		mean absolute difference		49.5	61.75

Table D.1.e.

Erratic Rotation Group 4

	Observed	Mean	Random		
Month	(0)	Imputation	Imputation	0-MI	0-RI
5	1222	1047	1223	175	-1
6	1165	1033	1186	132	-21
7	1188	1095	1332	93	-144
8	1121	1133	1451	-12	-330
		mean absolute difference		103.0	124.0

Table A.6

Longitudinal Information Used in Estimated Models

Month 5

Group	Month 1	2	3	4	9
c1	x	x	x	x	x
c4	x	x	x	x	x
s1	x	x	x	x	x
s4	x	x			x
e4	x	x	x	x	x

Month 6

Group	Month 2	3	4	5	9	10
c1	x	x	x	x	x	
c4	x	x	x		x	x
s1			x	x	x	x
s4	x	x		x	x	x
e4	x	x	x	x	x	x

Month 7

Group	Month 3	4	5	6	9	10	11
c1	x	x	x	x		x	x
c4		x	x	x	x	x	x
s1	x	x	x	x	x	x	x
s4	x		x	x	x	x	x
e4	x	x		x			

Month 8

Group	Month 4	5	6	7	9	10	11	12
c1	x	x	x		x		x	
c4		x	x				x	x
s1		x	x	x		x		x
s4			x		x	x		
e4		x		x	x	x	x	

c = consistent	1 = rotation group 1
s = semi-consistent	4 = rotation group 4
e = erratic	

Table A.4

Semi-Consistent Rotation Group 4

Month 5

Estimate	S.E.	Parameter
.43	.25	me2
1.93	.22	ne1
-1.42	.93	ne2
1.40	.34	ne9
.51	.26	ne10
-143.53	90.42	age1(1)
121.63	102.92	age1(2)
75.51	112.04	age1(3)
-21 84	105.16	age1(4)
-8.80	285.16	age1(5)

MSE = .2407 x 10^6

Month 6

Estimate	S.E.	Parameter
.48	.055	me2
.36	.23	me3
.30	.26	me5
-1.32	1.03	ne3
-1.34	1.01	ne5
1.52	.35	ne9
.46	.27	ne10
-139.02	106.22	age1(1)
111.70	116.80	age1(2)
88.82	123.67	age1(3)
-10.60	115.35	age1(4)
-37.59	286.28	age1(5)
-39.64	96.42	ed(2)
85.79	105.24	ed(3)

MSE = .2406 x 10^6

Month 7

Estimate	S.E.	Parameter
.56	.34	me3
.18	.12	me5
.54	.14	me6
.22	.11	me10
-.11	.085	me11
-1.69	1.60	ne3
1.04	.56	ne9
-259.74	81.01	age1(1)
-389.53	146.80	age1(2)
-382.10	162.24	age1(3)
-410.25	168.55	age1(4)
-24.44	176.18	age1(5)

MSE = .4873 x 10^6

Month 8

Estimate	S.E.	Parameter
.67	.10	me6
.20	.12	me9
.47	.41	me10

MSE = .6588 x 10^6

Table A.2

Consistent - Rotation Group 4

Month 5

Estimate	S.E.	Parameter
.11	.08	mel
.30	.11	me2
.30	.09	me3
.078	.088	me4
.21	.047	me9

MSE = .2000 x 10^5

Month 6

Estimate	S.E.	Parameter
.15	.061	me2
.33	.10	me3
.59	.26	me4
.15	.057	me9
.085	.046	me10
-1.20	.97	ne4

MSE = .1961 x 10^5

Month 7

Estimate	S.E.	Parameter
62.78	47.87	mean
.29	.077	me4
1.05	.13	me5
-.42	.085	me6
.36	.092	me9
.15	.076	me10
20.99	44.83	ed(2)
-73.04	47.31	ed(3)

MSE = .3176 x 10^5

Month 8

Estimate	S.E.	Parameter
.85	.046	me5
.059	.036	me6
.084	.035	me11
.27	.091	me12
-1.30	.40	ne12

MSE = .5672 x 10^4