A MEASURE OF THE BIAS IN THE MARCH CPS EARNINGS IMPUTATION SYSTEM

Edward J. Welniak and John F. Coder, Bureau of the Census

I. INTRODUCTION

Previous studies of the March CPS income imputation system have revealed a downward bias in procedures for imputing missing earnings values. Our research attempted to quantify this bias in more detail than previous studies. To quantify the bias, earnings information from a sample of fully reported respondents was deleted creating a sample of pseudo-nonrespondents. Earnings values were imputed to these cases and the imputed values compared to the originally reported values. These differences are then examined by detailed socio-economic group.

Our analysis has indicated some interesting patterns of bias in the March CPS imputation system. The largest biases tend to show up in areas which were minimally controlled or uncontrolled by the variables defining the statistical matching codes in the imputation's hot deck procedures. The imputation procedure has a downward bias of about 2 percent for both men and women. A detailed look at the biases show that part-time and less than full-year workers have relatively larger biases than full-time or year-round workers. The bias concerning class of worker and occupation of longest job were also significant. Underimputation of about 11 percent existed for self-employed professionals and 29 percent for self-employed managers. Overall, the impact on all self-employed was an underimputation of 16 percent.

II. A BRIEF DESCRIPTION OF THE CURRENT POPULATION SURVEY AND INCOME DATA COLLECTION

Each year since 1947 the Census Bureau has been compiling statistics on the annual income of persons and families in the United States. The Current Population Survey (CPS) has been the major source of annual statistics since that time.

The CPS is a monthly household survey designed to provide national estimates of employment, unemployment, and other related labor force statistics. In March, the labor force questions are supplemented with a group of questions covering work experience and income for the previous calendar year. In March 1979 (the reference month for the data used in this study) the survey consisted of about 56,000 interviewed households inhabited by about 125,000 persons aged 14 years old and over.

The work experience and income questions are asked for each person 14 years old and over. Interviews are conducted by Census Bureau interviewers. These interviews can be done either by personal visit or over the telephone. Proxy interviews are accepted from household members deemed capable of answering for other household members unavailable at the time of interview.

The March 1979 CPS questionnaire contained eleven separate income questions, each covering a specified income source or combination of income sources. The questions involved in this study are the 3 earnings questions covering: (1) wage or salary income, (2) nonfarm self-employment income, and (3) farm self-employment income. In these items, the receipt of earnings by type is determined and the amount recorded in cases of affirmative response to the receipt questions.

As a lead-in to the earnings questions, a series of questions covering work experience for the previous calendar year are also asked for each person 14 years old and over. Questions for this area include number of weeks worked, usual hours worked per week, number of employers, reasons for part-year employment, etc. In addition, the occupation, industry, and class of worker are collected for the job held longest during the previous calendar year. These three sets of information: (1) earnings, (2) work experience, and (3) longest job, are the major data items involved in this imputation study.

III. INCOME NONRESPONSE ON THE MARCH CPS

Nonresponse to sensitive questions such as income on household surveys which rely on voluntary cooperation is a longstanding problem with survey takers. The March CPS income supplement is no exception. In March 1979 the nonresponse rates for the earnings items being examined in this analysis are as follows:

Wage or salary.--A total of 11.4 percent of the sample persons were nonrespondents. Of the total nonrespondents, 38 percent indicated receipt but gave no amount. Of the total with no indication of receipt, 89 percent received an amount as a result of imputation. At the aggregate level, a total of about 18 percent of the total $1.053 billion of wages or salary was imputed. About 17 percent of all persons with wage or salary income had an imputed value.

Nonfarm self-employment.--About 7.7 percent of the sample persons were nonrespondents for nonfarm self-employment. Of these, 7.6 percent indicated receipt but gave no amount. At the aggregate level about 32 percent of the $88.6 billion aggregate was imputed. Twenty-six percent of the recipients were imputed.

Farm self-employment.--Farm self-employment had a nonresponse rate of 7.2 percent. Of these nonrespondents, 2.4 percent indicated recipiency but reported no amount. Nearly 22 percent of the farm income aggregate of $14.6 billion was imputed. These imputed amounts went to 21 percent of the farm income recipients.

IV. INCOME IMPUTATION PROCEDURES ON THE MARCH CPS

The March CPS work experience and income supplement data have been imputed for missing responses since 1962. Since that time, procedures have evolved with changing questions, procedures, and with the progression of computer capabilities.

The present imputation system was developed with two major goals in mind. First, we wanted to reduce nonresponse bias in the income statistics. Second, we wanted to maintain, wherever possible, the interrelationships between items in instances of multiple item imputations. To achieve these goals, we developed a rather complex statistical matching procedure for imputing nonresponse. The

procedure is essentially a "hot deck" system whereby nonrespondents and donors are matched on detailed socio-economic characteristics using a sorting and merging procedure rather than a hot deck matrix. Once a matching donor is located, values reported by the donor are assigned or imputed to the nonrespondent's record. In these instances, all missing information is imputed from the same donor in order to preserve the interrelationships between variables. A brief step-by-step account of the imputation procedure is as follows:

1) Respondents and nonrespondents are separated into their respective groups. The respondent group represents the donors in the imputation model.

2) Several matching records are created for each nonrespondent and donor. Each record defines an unique set of characteristics to be used in matching, each with less strict requirements.

3) Donor and nonrespondent records are sorted based on the detailed matching key.

4) The donor and nonrespondent files are merged and the donor matching at the most detailed group of socio-economic variables is selected as the donor for that specific nonrespondent.

5) Information from the donor is linked to the nonrespondent's record and the substitution of missing data takes place.

The variables used to define matching respondents and donors are quite detailed at the highest levels. For example, Table 1 shows the variables used to match in cases where the only piece of missing information is one or more of the three different earnings sources. In this particular situation, the variables defining the hot deck are comprised of about 6 billion unique combinations. This kind of detail allows matching of individuals very much alike in these measured characteristics and helps reduce imputation bias.

As mentioned earlier, we allow for several matches at various levels of detail. This is required because we cannot expect to find a matching donor at the most detailed level, given the approximately 66,000 donors and 16,000 nonrespondents. The final or lowest level must be established with variables of sufficient generality to assure finding a match and yet preserve certain consistencies which cannot be violated. In the example given above and in Table 1 there are in effect 12 possible levels of match, the final one based simply on age, sex, and educational attainment.

V. METHODOLOGY FOR MEASURING BIAS

Even though the hierarchical matching has worked well, it is not without shortcomings. Previous studies have shown a downward bias in imputing missing earnings values. What we have attempted to do is look at this bias in more detail than in the past. The methodology used to examine the imputation bias was based on the simulation of actual nonresponse. This was done by selecting as a base those persons with fully reported earnings data and creating a subset of pseudo-nonrespondents whose missing data could be imputed. These imputed values could then be evaluated with respect to the originally reported values. The pattern of nonresponse was based on the actual nonresponse found on the March 1979 CPS.

The steps taken in the procedure are as follows: First, nonresponse rates were tabulated from the March 1979 CPS by detailed socio-economic groups shown in Table 2. Second, a data file was created which contained households with fully reported earnings, work experience, and longest job information. Third, pseudo-nonrespondents were created from these fully reported households by deleting appropriate earnings, work experience, and longest job data based on the true nonresponse rates for their socio-economic group. A record of their location was kept so they could be found later for comparison purposes. Next, the pseudo-nonrespondents were run through the March imputation system to have their deleted information imputed. Finally, analysis was done by comparing the imputed earnings values of the pseudo-nonrespondents with their original earnings values.

VI. SUMMARY OF BIASES BY SELECTED CHARACTERISTICS

A summary of the imputation biases by selected characteristics is shown in Table 3. A measure we used to indicate degree of bias is the ratio of the imputed to the reported mean for the characteristics shown.

Overall, the mean imputed earnings value of $10,725 was 2.4 percent below the reported mean earnings of $10,991. The imputed mean for men was $13,890, 2.5 percent below the reported mean of $14,251. For women, the imputed mean was $6,222, about 2.0 percent below the reported mean of $6,353. Although overall the imputed and reported means are very similar, some characteristics when related to the level of earnings indicate some serious biases.

Weeks Worked.--Of the selected characteristics shown in Table 3, weeks worked exhibit some of the most serious bias problems. The under 13 weeks group has imputed amounts which, on average, are much larger than the reported average earnings for that group. For the under 13 weeks group, the ratio of imputed to reported earnings was 2.94. Weeks worked categories between 13 and 49 weeks also exhibited upward imputation bias whereas the largest group, 50 to 52 weeks group, showed a significant downward bias of about 6.2 percent. These problems largely held true for both men and women. For the 50 to 52 weeks group the downward bias for women was about 9.6 percent, 5.2 percent for men.

Usual Hours Worked Per Week.--The pattern of imputation bias for usual hours worked per week is similar to that for weeks worked. The lowest hours worked category of "under 10" had an imputed mean of $2,170 compared to the reported mean of $1,003. The ratio of imputed to reported means for men and women was 1.92 and 2.39 respectively in the "under 10 hours" category. In the highest hours worked category, "45 hours or more," there appears to be a significant underimputation problem. For men the ratio of means, imputed to reported, was .88 whereas the comparable figure for women was .83.

Occupation and Class of Worker of Longest Job.--The breakdowns by occupation and class of worker groups help to identify several problem areas within the imputation system. It appears that imputation for self-employed has a significant downward bias in the imputation system.

Overall, the ratio of imputed to reported means for self-employed was .84. For self-employed professionals this ratio was .89. For self-employed managers the ratio was .71. The mean imputed value for private household workers was $1,631 compared to a reported value of $1,218. This accounts for an overimputation of about 34 percent.

Education.--For the education groups shown in Table 3, it appears that the "less than 8 years" group has the most serious imputation problems. For this group we have an indication of overimputation. The mean imputed value was $8,152 about 17 percent higher than the reported mean of $6,970.

Age.--The ratios of imputed to reported mean earnings by age group indicates a situation of overimputation for the 14 to 19 year age group. Other age groups showed slight underimputation with the 20 to 24 and the 55 to 64 year groups having the largest underimputation.

Region.--Of the four regions, the Northeast is the only one that shows overimputation. The ratio of the imputed mean of $11,636 to reported mean of $11,162 was 1.04. The West, on the other hand, had a ratio of .92 and was the region with the most underimputation. The same pattern appeared for the males. Females varied slightly with the North Central being the region of overimputation (ratio of 1.01).

Residence.--Metro-nonmetro areas showed very little variation as shown in Table 3.

VII. ADJUSTING FOR BIAS

The second phase of this project will be to make use of these data to create imputation adjustment factors. These factors will then be used to "correct" actual imputed values on the March 1979 CPS. After this correction process, we will examine changes in the income and poverty statistics caused by the adjustments. We are presently developing the computer programs needed to complete this second step.

VIII. CONCLUSIONS

An examination of the results, shown in Table 3, have indicated some problem areas in the imputation system. These problems, in most cases, can be explained by reviewing the variables used to match nonrespondents and donors. The significant overimputation for persons working less than 26 weeks is probably due to matching at levels where little control was available for weeks worked. Similar problems with usual hours worked per week can be traced to matching variables which, at the most detail, distinguish only between part-time (under 35 hours per week) and full-time workers. Underimputation for the 45 hours per week or more group can be attributed to the same problems, insufficient control in the matching variables.

An adjustment procedure is being developed to try to reduce this bias, but it will be no real substitute for modifying the matching variables themselves. Each comparison shown in Table 3 does not reflect problems associated with that variable alone, but the combined effect across all variables. Because of this, any ad hoc adjustment can only provide some idea of the impact of an improved set of matching variables.

It should also be noted that the data used in this analysis was the result of one replication of the pseudo-nonrespondent technique. Ideally, the experiment should be repeated several times to obtain more reliable estimates of imputation bias; however, executing the March CPS income imputation program is very expensive and will probably prevent us from this necessity. We hope that in the near future we will be able to use what we have learned to modify the matching variables and improve the imputation accuracy.

TABLE 1.--VARIABLES USED IN DONOR-NONRESPONDENT MATCHING PROCEDURES OF THE MARCH CPS INCOME IMPUTATION PROCEDURE[1/]

Sex

- Male
- Female

Age

- 14 to 18
- 19 to 24
- 25 to 34
- 35 to 44
- 45 to 54
- 55 to 64
- 65 years and over

Race and Spanish Origin

- White, excluding Spanish
- Black, including Black Spanish
- White Spanish

Relationship to Family Head

- Head
- Wife of head
- Other relative of head
- Unrelated individual

Labor Force Status of Spouse

- Spouse in the labor force
- Spouse not in the labor force
- Not married spouse present

Weeks Worked Last Year

- Less than 13 weeks
- 13 to 26 weeks
- 27 to 39 weeks
- 40 to 49 weeks
- 50 to 52 weeks

TABLE 1.--CONTINUED

Educational Attainment

Less than 12 years
12 years
13 to 15 years
16 years
17 or more years

Class of Worker of Longest Job

Private wage or salary workers
Federal worker
State or local government worker
Self-employed
Without pay in family business

Region

Northeast
North Central
South
West

Full-time Part-time Status

Full-time (35+ hours per week)
Part-time (less than 35 hours per week)

Occupation of Longest Job

440 3-digit 1970 Census Occupations

Residence

Farm
Metro area, 1,000,000+
Metro area, under 1,000,000
Nonmetropolitan, nonfarm

Earning Recipiency Pattern

7 possible combinations of the 3 YES/NO's for: 1) wages or salaries, 2) nonfarm self-employment income, and 3) farm income

1/ Variables listed here are for highest level of matching for nonrespondents missing one or more earnings amounts only.

TABLE 2.--CHARACTERISTICS USED IN DETERMINING RATES OF NONRESPONSE

I. Sex

A. Male
B. Female

II. Age

A. Under 25 years of age
B. 25 to 54 years old
C. 55 years old and over

III. Education

A. Less than 9 years
B. 9 to 12 years
C. 13 or more years

IV. Weeks Worked

A. Less than 26 weeks
B. 27 to 49 weeks
C. 50 or more weeks

V. Hours Worked Per Week

A. Less than 15 hours
B. 15 to 39 hours
C. 40 or more hours

VI. Occupation

A. Salaried Professional
B. Self-employed Professional
C. Salaried Managers
D. Self-employed Managers
E. Sales
F. Clerical
G. Operatives and Transportation
H. Craftsmen
I. Nonfarm Labor
J. Farm and Farm Managers
K. Farm Labor
L. Service
M. Private Household

VII. Nonresponse

A. Complete questionnaire
B. Missing earnings value
C. Missing earnings value and recipiency
D. Missing work experience
E. Missing work experience and recipiency
F. Missing work experience and job
G. Missing job
H. Missing job and recipiency
I. Missing everything

TABLE 3.—NUMBER IMPUTED, MEAN IMPUTED EARNINGS, MEAN REPORTED EARNINGS, AND RATIO OF PERSONS BY SELECTED CHARACTERISTICS

Selected Characteristics	Number of Cases (In Units)	Mean Earnings Reported	Mean Earnings Imputed	Ratio
ALL RACES BOTH SEXES				
Total	10,151	10,991	10,725	.976
AGE				
14 to 19	1,033	1,954	2,218	1.135
20 to 24	1,457	6,736	6,288	.934
25 to 34	2,617	11,627	11,620	.999
35 to 44	1,844	14,104	13,668	.969
45 to 54	1,454	14,473	14,035	.970
55 to 64	1,392	13,974	13,395	.959
65 years and over	354	7,933	7,772	.980
EDUCATION				
Less than 8 years	401	6,970	8,152	1.170
8 to 11 years	1,956	7,220	7,023	.973
12 years	3,984	10,192	10,021	.983
13 to 15 years	1,991	11,174	10,586	.947
16 years or more	1,819	17,481	16,965	.971
REGION				
Northeast	2,110	11,162	11,636	1.042
North Central	2,681	10,935	10,568	.966
South	2,921	10,495	10,376	.989
West	2,439	11,499	10,527	.915
RESIDENCE				
In Metro, Total	6,347	11,724	11,286	.963
Inside Central City	2,609	10,818	10,645	.984
Outside Central City	3,738	12,357	11,734	.950
Nonmetropolitan	3,804	9,768	9,788	1.002
CLASS OF WORKER				
Private	7,409	10,268	10,336	1.007
Federal	364	15,852	13,299	.839
State or Local	1,334	10,158	10,567	1.040
Self-employed, Incorporated	164	29,242	22,925	.784
Self-employed or Farm	824	13,721	11,526	.840
Without Pay	56	1,275	1,737	1.362
OCCUPATION				
Professional				
Salaried	1,272	13,786	13,789	1.000
Self-employed	152	28,875	25,568	.885
Managers				
Salaried	851	18,953	18,965	1.001
Self-employed	293	22,178	15,714	.709
Sales	736	10,866	10,963	1.009
Clerical	1,740	7,802	7,491	.960
Operatives	1,270	13,540	13,738	1.015
Crafts	1,477	10,091	9,965	.987
Nonfarm Labor	508	6,747	6,726	.997
Farm and Farm Managers	184	9,950	9,664	.971
Farm Labor	157	3,522	3,690	1.048
Service	1,356	5,416	5,460	1.008
Private Household	155	1,218	1,631	1.339
WEEKS WORKED				
Less than 13	715	896	2,632	2.936
13 to 26	946	2,703	3,737	1.383
27 to 39	744	6,231	7,191	1.154
40 to 49	942	8,872	9,172	1.034
50 and over	6,804	14,018	13,148	.938
HOURS WORKED				
Less than 10	286	1,003	2,170	2.164
10 to 19	516	1,995	3,289	1.648
20 to 34	1,189	3,726	4,489	1.203
35 to 39	723	8,982	9,427	1.050
40 to 44	5,219	11,773	11,735	.997
45 and over	2,218	17,082	14,950	.852