COMPARISON AND EVALUATION OF ALTERNATIVE ICM IMPUTATION METHODS

Suzanne M. Dorinski, Rita J. Petroni, Michael Ikeda, and Rajendra P. Singh[1]
Suzanne M. Dorinski, Statistical Research Division, U.S. Bureau of the Census, Washington, DC 20233

Key Words: conditional distributions, hot-deck, flexible matching, imputation

Introduction

To produce Dual System Estimates for the 1990 Census, the Census Bureau imputed missing items based on conditional distributions or from previous records using a hot-deck approach. For the 1995 Census Test, the Bureau primarily used flexible matching imputation to impute values for the Integrated Coverage Measurement (ICM) samples. The ICM samples are used to calculate Census coverage rates among various demographic subgroups. This paper compares and evaluates the two methods as a first step in selecting an imputation method for Census 2000 ICM samples. We use complete records from one of the 1990 Census evaluation samples and simulate item nonresponse by dropping items.

Final results indicate that in general the method used in the 1990 Census produces results which are more consistent with the reported data.

The next section describes the 1990 and 1995 imputation methods. The succeeding section details the simulation done for this study. The results from both methods of imputation and comparisons of the two methods follow. The final section presents conclusions.

Background

The item nonresponse imputation method used in the 1990 Census is a hot-deck imputation procedure that fills in values for the missing data. Certain information about other household members is used in the hot-deck procedure when such information is available. When information on other household members is unavailable, the hot-deck procedure imputes values based on either a previous household with reported values or the distribution of reported values in the entire file. Tenure is imputed first, followed by race, Hispanic origin, sex, and age.

When information on other household members is unavailable, the tenure, race and Hispanic origin imputations use values from a previous reporting household. When tenure is not reported for the household, tenure is imputed from the reported tenure of the previous household with a similar structure. Missing race is imputed from the race distribution of other household members when race is reported for at least one household member. When race is not reported for any household member, missing race is imputed from the race distribution of individuals in the previous housing unit with reported race. Missing Hispanic origin is imputed from the Hispanic origin distribution of other household members when the Hispanic origin is reported for at least one other household member. When Hispanic origin is not reported for any household member, missing Hispanic origin is imputed from the Hispanic origin distribution of individuals in the previous housing unit with reported Hispanic origin.

The sex imputation uses the distributions shown in Table 1, depending on the characteristics of the person with missing sex. We generate a random number between 0 and 1, then use the random number to assign a sex from the cumulative distribution. The age imputation uses the distributions shown in Table 2, depending on the characteristics of the person with missing age. As with the sex imputation, we generate a random number between 0 and 1, then use the random number to assign an age from the cumulative distribution. For more information on the 1990 imputation method, see Diffendal and Belin (1991).

The flexible matching imputation used in the 1995 Census Test performs hot-deck imputation by finding matching variables and using the variables to match an incomplete record with a complete record. Matching variables are found for each possible combination of missing variables that exist on the incomplete records. Within each combination, separate matching variables are determined and ranked by order of importance for the missing continuous variables and each individual missing categorical variable. Multivariate linear regression models are used in finding and ranking the matching variables for the missing continuous variables and polytomous logit models are used for the missing categorical variables. For more information on the

[1] This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

flexible matching imputation, see Williams (1995a and 1995b).

Once the matching variables are found, an attempt is made to match the incomplete record to a complete record, using the values of the matching variables first for the missing continuous variables, and then each individual missing categorical variable. If a match is found, the appropriate missing variables are replaced with the variable values on the complete record. If a match is not found after a full cycle through the complete records, the least important matching variable is dropped and an attempt at finding a match is resumed using the remaining matching variables. Imputed values for the missing continuous variables are taken from the same complete record, while imputed values for each missing categorical variable are taken from different complete records.

The flexible matching imputation software produces an output file that lists the matching variables for each missing variable. The software also allows users to add additional matching variables. The missing variables are imputed in a predetermined order. For this study, tenure was imputed first, then age, sex, Hispanic origin, and race.

A FORTRAN program that prepares the file for the flexible matching imputation software performs the tenure imputation and also does the imputation of sex for spouses. As in the 1990 Census method, the tenure imputation is a hot-deck based on the value reported by a previous household, however not by type of structure. If one spouse's sex is missing, it is imputed to be the opposite sex of the reporting spouse. However, if both the householder and the spouse do not report sex, the householder is imputed as male and the spouse is imputed as female. For more information, see Ikeda and Petroni (1996).

Simulation Methodology

We used the 1990 E-sample data file for this study. The 1990 E-sample was part of the Census evaluation sample used to determine how many individuals were correctly enumerated. The items imputed in both the 1990 Census and the 1995 Census Test are tenure, race, Hispanic origin, sex and age. Thus we focus on these items when evaluating the methods. However, the 1990 Census imputation method allows hot-decking based on missing values for relationship and marital status, so we included those items when simulating missing data. The 1995 Census Test did not include group quarters, so we excluded group quarters records from the 1990 E-sample data file for this study.

To simulate nonresponse, we looked at response patterns for given subgroups on the whole file. For age, sex, race, Hispanic origin, relationship and marital status, there are 64 response patterns, ranging from an individual providing all six data items to an individual refusing to provide all six data items. The subgroups we used are based on Hispanic origin, race, sex and age categories. The Hispanic origin and race classifications include the following 8 categories:

White nonHispanic, other
Black nonHispanic
American Indian nonHispanic
Asian nonHispanic
Black Hispanic
Hispanic not Black, not Asian, not American Indian
Hispanic American Indian and
Hispanic Asian.

Combining the Hispanic origin and race classifications with the sex and age (0-17, 18-29, 30-54 and 55+) categories produces 64 subgroups. Hence Black Hispanic males between 18 and 29 years of age were treated differently than Asian nonHispanic females between 30 and 54 years of age.

We then deleted any record missing at least one of the six variables. The resulting file with no imputed data became the base file for our study. We randomly assigned response patterns by subgroup to individuals on our base file, based on the frequency of the response pattern occurring for that subgroup on the whole file. Once the response patterns were assigned, we were able to blank out the corresponding variables and thus simulate missingness.

Race, Hispanic origin, sex, age, relationship and marital status are all person-level variables on the Census questionnaire. Different members of a household may have different response patterns for those variables. Tenure is a household-level variable and thus should be the same for every member of the household. For that reason, simulation of missing tenure was handled separately from the other variables. We calculated frequencies of missing tenure by household size on the whole file, then applied those frequencies to households on our base file.

Results

To evaluate the two imputation methods, one could compare two measures of success: the number of correct imputations each method produces, or the "closeness" of the marginal distributions produced by each imputation method to the reported marginal distributions. The

number of correct imputations is a micro-level measure, while the "closeness" of the marginal distribution to the reported distribution is a macro-level measure. The focus of the ICM samples is to produce accurate results at aggregate levels so that differential undercounts may be examined. Thus we will first compare the "closeness" of the marginal distributions to evaluate the methods. Secondarily, we will then look at the number of correct imputations each method produces.

Table 3 shows the results from both methods for imputation of tenure. Since tenure was imputed at the household level, the numbers in the table represent households, not persons. The 1990 method and the 1995 method produce approximately the same marginal distributions. However, the 1990 method produces more correct imputations.

The results for imputation of race are shown in Table 4. The 1995 method imputes race using Hispanic origin and strata (a racial and Hispanic origin code for the composition of the sampled areas) as matching variables. We chose to use household identifier as an additional matching variable. The household identifier variable is used to try to find a match among other persons in the household. The 1990 method, which uses either the reported racial distribution of the person's household or the racial distribution of a previously reporting household, produces marginals which are closer to the reported marginals as well as more correct imputations.

Table 5 shows the results from both methods for imputation of Hispanic origin. The 1995 method uses household size and strata as matching variables. We chose to use race (if available) and household identifier as additional matching variables. The 1990 method uses either the reported Hispanic origin of the person's household or the Hispanic origin distribution of a previously reporting household. The 1995 method produces marginals which are slightly closer to the reported marginals. However, the 1990 method produces more correct imputations. The 1995 method has 57.5% more persons incorrectly imputed as not Hispanic.

The results for imputation of sex are shown in Table 6. The 1995 method uses Hispanic origin and relationship to person one as matching variables when Hispanic origin is available and household size in place of Hispanic origin when Hispanic origin is missing. The 1990 method uses the reported sex distribution of the entire file by relationship to person one. The 1990 method produces marginals which are closer to the reported marginals. However, there is room for improvement of both methods. The methods are mixed in terms of which gives more correct imputations.

Table 7 shows the results from both methods for imputation of age. The 1995 method uses relationship to person one, tenure, type of mail return (a code indicating short form or long form, Spanish questionnaire or not, mailed return or not), strata and sex as matching variables when sex is not missing, and uses type of structure in place of sex when sex is missing. The 1995 method is imputing too many persons in the 0-9, 10-19, and 30-44 age categories, while not enough in the 20-29, 45-64, and 65+ categories. The 1990 method, which uses the reported age distribution of the entire file by household size, produces marginals which are closer to the reported marginals as well as more correct imputations.

Conclusions

Overall, the 1990 method of imputation performs better for the characteristics studied. However, for Hispanic origin, the flexible matching imputation produces slightly better marginals, but fewer correct imputations.

Further analysis has shown that most of the errors in the 1990 sex imputation occur when imputing the sex of single householders. Most 2+ person households are married-couple households, with the husband being listed as the householder almost all the time. This causes the sex distribution of reporting householders to be predominately male. However, householders who are single tend to be female, so using the overall sex distribution of all reporting householders causes most single householders to be imputed as male. We suggest that the 1990 sex imputation method for householders be subdivided even further, with married householders being imputed based on the distribution of all reporting married householders, and single householders being imputed based on the distribution of all reporting single householders. (Dorinski, 1996).

References

Diffendal, Gregg and Tom Belin, (1991), STSD Decennial Census Memorandum Series #V-112, "Results of Procedures for Handling Noninterviews, Missing Characteristic Data, and Unresolved Enumeration Status in 1990 Census/Post-Enumeration Survey", July 1, 1991.

Dorinski, Suzanne M. (1996), Census Bureau Memorandum for Documentation, "Evaluation of Sex Imputation Methods used in 1990 PES", September 23, 1996.

Ikeda, Michael and Rita Petroni (1996), "Handling of Missing Data in the 1995 Integrated Coverage

Measurement Sample", to be presented at the 1996 Joint Statistical Meetings.

Williams, Todd R. (1995a), Census Bureau Memorandum for Documentation, "Methodology Used for the Modeling of Missing Variables in the Flexible Matching Imputation Software", July 14, 1995.

Williams, Todd R. (1995b), Census Bureau Memorandum for Documentation, "Using the Flexible Matching Imputation Software", July 17, 1995.

Acknowledgments

The authors wish to thank Neal Bross for computer support on this study, Carma Hogue and Bill Bell for their comments on this paper, and Lisa Mundy for her help in preparing the final version of this paper.

Table 1. Sex Imputation Cells for 1990 Method

Kind of person	Distribution used
One-person household	All reporting one-person households
Persons in 2+ person households	
Nonreporting spouse made consistent with reporting (or imputed householder) spouse	
Householder	All reporting householders
Other relationship stated	All persons in 2+ person households except householders, spouses, missing relationships
Relationship missing	All persons in 2+ person households except householders

Table 2. Age Imputation Cells for 1990 Method

Kind of person	Distribution used
One-person household	All reporting one-person households with same marital status
Persons in 2+ person households	Reporting individuals in households of 2+ persons with same relationship to householder, same age of householder, and same marital status.

Table 3. Tenure Imputation Results Under 1990 and 1995 Methods

	Imputed				
	Own		Rent		
Reported	1990	1995	1990	1995	Total
Own	2186	2130	470	526	2656
Rent	509	567	1135	1077	1644
Total	2695	2697	1605	1603	4300

Table 4. Race Imputation Results Under 1990 and 1995 Methods

	Imputed								
	White, Eskimo, Aleut, Other		Black		Asian Pacific Islander		American Indian		
Reported	1990	1995	1990	1995	1990	1995	1990	1995	Total
White, Eskimo, Aleut, Other	7226	7210	73	71	30	46	21	23	7350
Black	79	203	1605	1462	5	16	1	9	1690
Asian Pacific Islander	99	123	6	7	386	362	2	1	493
American Indian	52	66	8	4	5	0	197	192	262
Total	7456	7602	1692	1544	426	424	221	225	9795

Table 5. Hispanic Origin Imputation Results Under 1990 and 1995 Methods

	Imputed				
	Not Hispanic		Hispanic		
Reported	1990	1995	1990	1995	Total
Not Hispanic	30193	30130	438	501	30631
Hispanic	290	457	1592	1425	1882
Total	30483	30587	2030	1926	32513

Table 6. Sex Imputation Results Under 1990 and 1995 Methods

	Imputed				
	Male		Female		
Reported	1990	1995	1990	1995	Total
Male	1097	1118	601	580	1698
Female	760	810	1146	1096	1906
Total	1857	1928	1747	1676	3604

Table 7. Age Imputation Results Under 1990 and 1995 Methods

	Imputed					
	0 - 9		10 - 19		20 - 29	
Reported	1990	1995	1990	1995	1990	1995
0 - 9	517	388	267	294	67	132
10 - 19	316	303	294	233	155	138
20 - 29	160	227	212	233	495	364
30 - 44	50	135	71	143	418	362
45 - 64	17	35	18	54	128	193
65 +	14	30	5	27	86	106
Total	1074	1118	867	984	1349	1295

Table 7. Age Imputation Results Under 1990 and 1995 Methods (continued)

	30 - 44		45 - 64		65 +		
Reported	1990	1995	1990	1995	1990	1995	Total
0 - 9	34	54	8	22	10	13	104
10 - 19	43	97	7	31	7	20	114
20 - 29	382	411	152	159	86	93	1240
30 - 44	986	776	429	491	218	265	3266
45 - 64	512	558	582	433	279	263	2746
65 +	241	367	296	270	378	220	1830
Total	2198	2263	1474	1406	978	874	7940