MULTIVARIATE ITEM IMPUTATION FOR THE 2000 CENSUS SHORT FORM

Yves Thibaudeau, US Census Bureau, Todd Williams, US Census Bureau, Tom Krenzke, Westat
Yves Thibaudeau, US Census Bureau, Statistical Research, FOB 4-3000, Washington DC 20233

KEY WORDS: Census Short Form, Hot-Deck Imputation, Multiple Regression

1. Introduction

The intent behind the paper is to expose a simple methodology for short form item imputation in the 2000 census. The short form records seven demographic items for each occupant of a housing unit (HU) and is delivered to all the HU's in the United States. We construct the methodology with two objectives in mind: to design a system that is adaptable to the wide spectrum of multivariate contingencies generated by the short form, and to build a system from commonly available of-the-shelf software components to keep the programming to a minimum.

We review existing imputation procedures in section 2. In section 3 and 4 we propose imputation strategies for the HU items and the person items respectively. We present examples to illustrate our technique both for the HU and person items.

2. Review of Some Existing Imputation Procedures

The specifications for the procedures for item imputation in the 1990 census are complex and lengthy (Treat, 1994). One of the most important component of the 1990 deterministic rules is the particular application of the hot-deck. The hot-deck in 1990 consists of replacing a missing item by the latest observed value for that item. This form of the hot-deck is used for tenure, race, and hispanic origin of the householder, and age, gender, and relationship of all persons. This seems reasonable but causes problems: Some of the stochastic properties of the items can be lost through a deterministic process of this type. We believe that the imputation process should reproduce the random quality of the missing items. We aim for a methodology that captures the elements of the probabilistic structure of the missing items still alive in the observations.

Little and Rubin (1987, p. 237), and Kalton (1981) discuss versions of the hot deck in a random set-up. Theirs involve the selection of items at random from a group of qualified donors. This approach restores the stochastic nature of the distribution of the missing items. The key to a faithful stochastic reproduction of the missing items is the selection of the group of donors. In most respects, the delineation of a complex structure of donors is just as tedious as model selection. The advantage is that it avoids parameter estimation. The drawback is that unobserved but legitimate values of the missing items can never be imputed. This approach is referred to as "model implicit".

We show in the subsequent sections how to generate imputations through a probabilistic model. Our technique attempts to circumvent the problems associated with deterministic imputation and implicit model imputation. We first center the attention on imputing the housing unit items on the short form.

3. Imputing Housing Unit Items for the 2000 Census.

In this section we lay down a strategy for imputing housing units (HU) items in the 2000 census. We focus on designing a realistic model for imputation. We give a specific example of the method.

3.1 A Log-Linear Model for Housing Unit Items

Log-linear models are particularly well suited for the analysis of the HU items since those are categorical, with the exception of the age of the householder which we impute in the next section. We turn to log-linear models in an attempt to simplify and at the same time generalize the use of hot-deck variables to impute missing items. Our model shall embed the discrimination power of the hot-deck as it was used in the deterministic imputation of the 1990 census, and the stochastic quality of model implicit imputation.

We integrate HU and hot deck items in the model. The HU items are: tenure, gender, race, and origin of the householder. The hot-deck variables are: tenure of the nearest preceding householder who reported the tenure item; tenure of the nearest following householder who reported the tenure item; race of the nearest preceding neighbor who reported the race item. The order of precedence is that of the census file and corresponds to geographical contiguity.

We expect the tenure of a given unit to often agree with either hot-deck tenures, and race of the householder to be significantly associated with the hot-deck race. There are potent interactions between the HU items. We sketch a tentative model that describes the relationships between all the variables.

We use a notation consistent with the command language of the procedure CATMOD from the SAS software.

$$n_{t,\,g,\,r,\,o,\,hd1,\,hd2,\,hd3}$$
$$\sim (T|G|R|O\ @\ 3)\ (R|HD3)\ (HD1|HD3)\ (T|HD1|HD2) \qquad (1)$$

The LHS of (1) is the count of HU's with demographic items represented by seven subscripts. Subscript t, g, r, o correspond to tenure, gender, race, and hispanic origin of the householder. Subscripts hd1, hd2, hd3 correspond to hot deck items: the tenures of the preceding and following HU's, and the race of the preceding householder. For the purpose of this analysis each item is binary: tenure is "owner" or "renter"; gender is "male" or "female"; race is "black" or "non-black"; hispanic origin is "hispanic" or "non-hispanic".

The expression on the RHS of (1) represents a summary of the configuration of the expected value of the count in terms of the interactions between the variables. Each set of parentheses contains information on the interdependency structure. The first set of parentheses indicates that all 3-way interactions between tenure, gender, race, and origin are integrated in the model. The subsequent sets of parentheses on the RHS represent the 2-way interactions between race and preceding race, the 2-way interactions between preceding tenure and preceding race, and the 3-way interactions involving tenure, preceding tenure, and following tenure. Due to the hierarchical structure, interactions of order less than those explicitly included are always included.

3.2 Item Imputation for a California DO or the Powerful Discrimination Functions of Race and Hispanic Origin.

We use model (1) to produce HU imputations for Los Angeles District Office (DO) 3205. Then compare our results for the imputations of 1990. Essentially, when its parameters are estimated, model (1) functions as a ratio estimator to impute the missing HU items.

We use the CATMOD procedure of SAS (M step), coupled with conditional expectation (E step), in a simple version of the EM algorithm to estimate the parameters of model (1). A new set of parameter is estimated for each tract. A tract is a geographical unit containing approximately 1500 housing units or 4000 persons. The distribution of the missing items conditional on the contingencies is multinomial and easy to simulate given the value of the parameters.

When the contingencies define cells of HU's that are homogeneous with respect to the nonresponse mechanisms our approach is most effective. Of course in practice it is impossible to ensure homogeneity. But we construct cells that are homogenous with respect to observable variables. In table 1 we examine four contingencies defined by the race of the preceding neighbor and the hispanic origin of the householder. The numbers and proportions of blacks for the four resulting contingencies are given when race is observed and when race is imputed under the 1990 imputation and under our multivariate method.

We center the attention on the last contingency of table 1 (hispanic householder, black neighbor). Under the 1990 imputation strategy hispanics with missing race and a black neighbor are designated blacks disproportionately. Almost 80% of the hispanics of known origin but unknown race and preceded by a black neighbor are imputed as blacks. That is ten times the observed rate. Our multivariate strategy allows for a strong interaction between the origin and race items in (1), and leads to a proportion of designated blacks consistent with the observed rate. Table 2 Gives the total numbers of black hispanic householders. In 1990 two-thirds of these householders are in fact a creation of the imputation methodology. Our multivariate approach guards against these inconsistencies.

4. Imputing Person Items for the 2000 Census

This stage of our method replaces the missing data that can differ for each person within a HU. For this stage, an estimation procedure that fits logistic and multiple regression models to the non-missing data is proposed. The logistic regression models are used to predict response values for a person's missing relationship to the householder. The multiple regression models are used to predict values for a person's missing age. The models were developed based on our comparing the significance of the predictor variables and the goodness-of-fit of each model. By using this modeling approach, we hope to produce replacement data that maintains the same relationships as found in the non-missing data.

4.1 Predicting Missing Relationship

Excluding the householder and the householder's spouse which are determined in editing, there are ten types or levels of response for the relationship variable that require imputation. For each of the ten levels, we find the probability of the missing value equaling that level for each person with a missing relationship value. We determine these probabilities by fitting a logistic regression model to the non-missing data where the response variable is the multinomial relationship variable. For the model, the response function f_i for level i is

$$f_i = \ln\left(\frac{p_i}{p_{10}}\right) = \beta_{i0} + \beta_{i1}x_1 + \ldots + \beta_{ik}x_k$$

for $i = 1,2, \ldots, 9$, where k is the number of predictor variables. The probabilities associated with the levels of response are for $i = 1,2, \ldots, 9$

$$p_i = \frac{\exp(f_i)}{1 + \sum_{m=1}^{9} \exp(f_m)} \text{ and } p_{10} = \frac{1}{1 + \sum_{m=1}^{9} \exp(f_m)}$$

The most significant predictor variable for predicting a person's relationship to the householder is the difference in age between the person whose relationship is missing and the householder. Unfortunately, the age of either person may also be missing, which leads us to develop three models. Our first model is used when both ages are present and the most important predictor variable is the difference in the ages. Our second model is used when the age of the person is present, but the age of the householder is missing. In this model the age of the person is the most important predictor variable. Our third model is used when the person's age is missing. This model includes the mean number of householder children within a HU for the tract and the mean householder age for the tract as predictor variables.

For a person with a missing relationship value, we derive the probability p_i associated with each response level i of the relationship variable from the appropriate model. We impute a value for relationship by randomly selecting the value from the multinomial distribution with parameters $p_1, p_2, \ldots, p_{10}$.

4.2 Predicting Missing Age

We fit four multiple regression models to the complete data in order to replace missing age values with predicted values. Within a HU, we impute the missing age of the householder before any other missing age. The first two of our four models are used for predicting the age of the householder when the age of another person in the HU is available and when there is no other age available. The third model is used for predicting the age of a child or stepchild of the householder and the fourth is used for predicting the age of any other member in the HU. The general form of a multiple regression model for predicting age is

$$AGE = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \quad ,$$

where k is the number of predictor variables.

For predicting the age of the householder, the age of another person in the HU is a strong predictor variable. Only one person's age is used in our model and this person is determined by the following order: spouse, oldest child, youngest parent , unmarried partner, first listed roommate with a non-missing age and, oldest grandchild. The age of a person is used only if there does not exist a person higher in the order or the ages of all persons higher in the order are missing. When none of these ages is available, our second model is used to predict the age.

The most significant predictor variable in our third model, which predicts the age of a child or stepchild of the householder, and our fourth model, which predicts the age of all other persons, is the age of the householder. To ensure that the ages differ when there is more than one child belonging to the householder, we create a predictor variable specifically for the third model. This variable provides the order in which the child is listed within the HU in relation to the other children. It is a significant predictor of the child's age and different ages are provided for children with different rankings.

Once a predicted value from the appropriate model is found for a person's missing age, we add random error to the value. We accomplish this by randomly selecting a residual from the distribution of the residuals obtained by fitting the model, where the residual is the observed value minus the predicted value. This randomly selected residual is added to the predicted value to produce the replacement value for the missing age.

4.3 Results Using the 1990 Census Data

We performed the new imputation method for imputing a person's relationship and age using 1990 Census data from the Los Angeles DO 3205. For the imputed HUs, we compared estimates derived using the old hot deck method and estimates from the new model fitting method. These estimates were also compared with estimates derived from the complete data HUs. A complete data HU is a HU where person item imputation was not needed. We found similar results to those given below using data from a Sacramento, Ca. DO and a Bergen County, NJ DO.

For comparing the imputed values of a person's relationship to the householder, the percentage of HUs containing persons with a certain type of relationship are displayed in Figure 1 for HUs that contain and do not contain a spouse of the householder. The three categories shown are HUs with 1) children only, 2) other relatives that can also include children, but not include nonrelatives, and 3) nonrelatives. Figure 1 shows that the two imputation methods provide almost identical results.

To compare imputed ages between the old and new methods, we calculated the mean age for the householder, spouse of the householder and oldest child of the householder for persons with imputed ages and for persons from the complete data HUs. The following table gives the means for the entire DO. This table shows that the means for the imputed ages produced by the new imputation method are lower than those produced by the hot deck method.

Mean age of	Complete Data	Old Imputes	New Imputes
Householder	48.4	46.8	43.9
Spouse	44.2	46.2	41.5
Oldest Child	12.5	12.2	11.7

More dramatic differences in the mean ages between the two imputation methods can be seen when viewed at the tract level. Figure 2 displays the mean age of the householder by tract. This figure shows that for most of the tracts the mean age is lower for householders whose ages are imputed by the new method. The same results can be seen with the mean ages of the spouse of the householder (not shown).

Because the new imputation method fits most of the multiple regression models to the complete data with a high degree of accuracy, we feel that the lower average predicted values of age may represent a downward bias associated with missing data HUs that is not captured by the old hot deck method. It appears that an overall change in the mean age would more likely be seen for an individual tract than for the entire DO. For both imputed relationships and ages, we intend to make further comparisons within tenure and race categories and to use data from other DOs to determine the consistency of the findings.

5. Conclusion

In the paper we present a two-step approach to impute short form items in 2000. The central theme of our strategy is the preservation of multivariate relationships throughout the imputation process. We also stress computing simplicity and portability in as much as these qualities apply to the SAS software. We showed that in the case of the Los Angeles DO 3205 our system naturally adjusts to the local contingencies to avoid the pitfall of designating a disproportionate number of householders as blacks, as was done in 1990. There may well be other, perhaps more subtle, pitfalls of this nature. We recommend that any imputation strategy in 2000 be based on a multivariate approach.

6. References

Kalton, G. (1981). *Compensating for Missing Data.* Ann Arbor: Survey Research Center, University of Michigan.

Little, R.J., Rubin, D.B. (1987). *Statistical Analysis with Missing Data.* Wiley Ed.

Treat, J.B. (1994). *Summary of the 1990 Census Imputation Procedures For the 100% Population and Housing Items.* DSSD REX Memorandum Series BB-11. US Bureau of the Census.

Disclaimer

This paper reports 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.

Table 1. Percentages of Observed and Designated African Americans among Householders by Observed Hispanic Origin of the Householder and Race of the Preceding Neighbor - California DO 3205

Hispanic Origin of the Householder and Race of the Preceding Neighbor	% Observed Blacks among Householders with Known Race	% Designated Blacks among Householders with Imputed Race under the 1990 Census	% Designated Blacks among Householder with Imputed Race under Multivariate Imputation	Difference between the Numbers of Designated Blacks under the 1990 Census and under Multivariate Imputation
Non-Hisp. Hsldr and Non-Black Neighbor	71.4 %	24.7 %	68.0 %	- 84
Non-Hisp. Hsldr and Black Neighbor	93.2 %	93.9 %	96.0%	-16
Hispanic Hsldr and Non-Black Neighbor	1.5 %	11.3 %	0.9 %	291
Hispanic Hsldr and Black Neighbor	**7.5 %**	**77.8 %**	**6.8 %**	**1198**
Total	67.0%	43.2%	17.2 %	1389

Table 2. Numbers of Observed and Designated Hispanic African American Householders under the 1990 Census and under Multivariate Imputation - California DO 3205

Number of Observed Hispanic Black Householders (Percentage among Householders with Observed Race & Hispanic Origin)	Number of Designated Hispanic Black Householders under the 1990 Census	Number of Designated Hispanic Black Householders under Multivariate Imputation	Total Number of Hispanic Black Householders under the 1990 Census (Percentage among All Householders)	Total Number of Hispanic Black Householders under Multivariate Imputation (Percentage among All Householders)
998 (0.95 %)	2385	425	3383 (2.50 %)	1423 (1.05%)

Figure 1

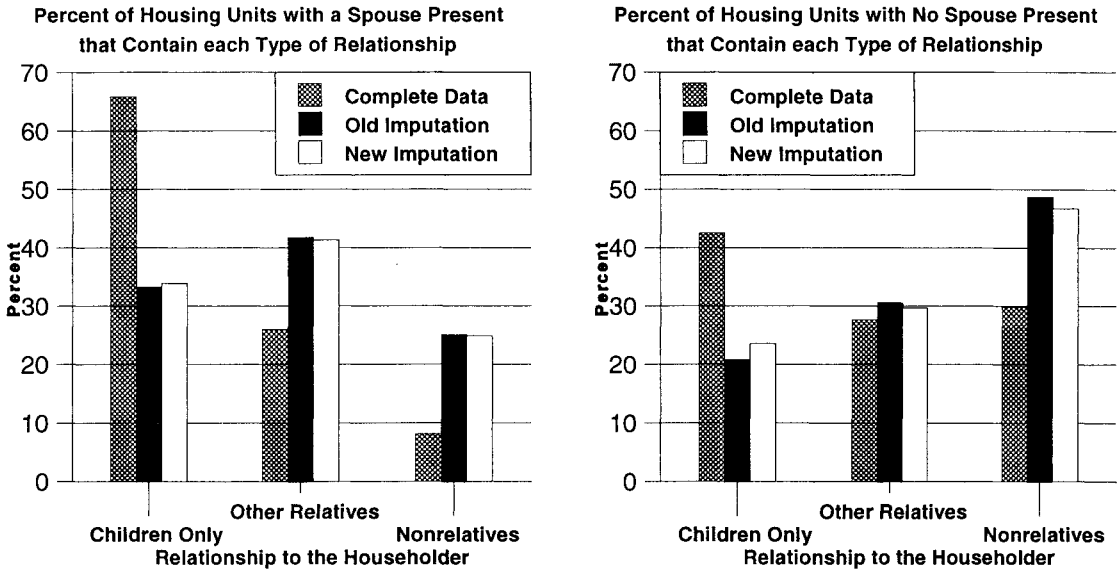

Figure 2