Proceedings of the Annual Meeting of the American Statistical Association, August 5-9, 2001

FLEXIBLE MATCHING IMPUTATION: COMBINING HOT-DECK IMPUTATION WITH MODEL-BASED METHODOLOGY

Todd R. Williams U. S. Bureau of the Census[*]
U. S. Bureau of the Census, Washington, D.C. 20233

Keywords: Hot-Deck Imputation, Regression, Flexible Matching

I. Introduction

A major concern of the Census Bureau is how to compensate for information that is missing for individuals interviewed by either the decennial census or by one of the many surveys the Bureau conducts. This missing information may be due to refusal to answer a particular question or that the information is not consistent with other data and; therefore, fails an edit check. The Census Bureau currently fills in a large part of the missing information using a hot-deck method of imputation. This method matches a data record containing missing data with a donor within the same data file that does not have missing data (Stiller. et al., 1998). A match is made using a set of variables whose values are both present and identical on both the imputed and the donor records. The missing data is then replaced with the data found on the donor record.

The purpose of our research is to test an alternative method of imputation in an attempt to find a method that better preserves multi-variable relationships. In our analysis, we test what we refer to as our flexible matching method of imputation (Long, 1992). Our flexible matching imputation procedure combines hot-deck imputation with model-based techniques. The procedure performs hot-deck imputation by matching a data record with missing values to a donor record. The match is made on a set of matching variables whose values on the missing data record are the same as those on the donor record. The matching variables are ranked from highest to lowest priority. If a match can not be made using all of the matching variables, the lowest priority variable is dropped and a match is attempted using the remaining matching variables. This process is repeated until either a match is made or all of the matching variables are dropped. Even though we find it to be extremely rare that a match can not be made with at least one of the matching variables, a default set of matching variables can be declared to ensure that there will be a match.

When matching on variables in a hot-deck imputation procedure, we find it sometimes difficult to make an exact match. This is especially true for continuous variables. The current hot-deck procedure handles this problem by coding the values of a variable into predefined categories and performing the match on the category values. The categories are based on subject matter constraints and the need to ensure that a donor is found. Our flexible matching imputation procedure automatically codes the continuous variable values into their corresponding deciles and performs the match based on the deciles. Both imputation procedures look for direct matches on categorical variables. When a match is not found, the current hot-deck procedure uses a set of cold-deck values to replace the missing values of a data record. With our flexible matching procedure, we drop the least important matching variable and continue to try to find a donor using the remaining matching variables. This process is repeated until a donor is found.

The current hot-deck imputation procedure tries to find a donor record that is closest in geographical location to that of the missing data record. With our flexible matching procedure, geographical location is not a constraint in the matching.

The model-based portion of the flexible matching imputation procedure determines the matching variables. For each variable that has missing values, a regression model is fitted to the set of data records that do not have any missing values. We refer to the set of nonmissing data records as the fully observed data. For missing continuous variables, a multivariate linear regression model is fitted. For each missing categorical variable, a polytomous logit model is fitted. A stepwise procedure is used to find the regression model that best predicts the values for the missing data. The first predictor variable placed in the model is the most important matching variable, the second predictor variable added to the model is the second most important matching variable, and so on for the third, fourth, and remaining predictor variables.

The set of matching variables used in the current hot-deck method are usually predetermined by subject matter and donor availability considerations. Because our flexible matching imputation procedure uses the best predictors of the missing variable found in the data, we will try to show that our flexible matching procedure is better at preserving multi-variable relationships that appear in the data.

II. Analysis

For our analysis, we use data collected by the 1999 American Community Survey (ACS). The Census Bureau developed the ACS to replace the long form of the

2010 Census. As opposed to the census long form which collects data every ten years, the ACS will collect demographic, housing, social and economic information every year for all states as well as all cities and other areas of 65,000 or more.

For our evaluation, we want to choose a variable that has a higher percentage of missing cases in relation to other variables, one that is applicable to a large portion of the data, and one that will show strong relationships to other data variables. We find that the wages and salary variable supplied on the 1999 ACS is a good candidate. From this point on, we will refer to this variable as wages.

In our analysis, we want to compare the imputed wages values derived from the current hot-deck procedure performed by the Census Bureau with the imputed values derived from our flexible matching procedure. We perform this by comparing the mean wages taken from both the hot-deck and flexible matching imputed data to the those of the fully observed data.

To further the comparison, we also derive imputed wages values directly from fitted multiple regression models. The models essentially have the same predictor variables as those determined to be matching variables by the flexible matching procedure. In an attempt to obtain the best fitted models for the both the flexible matching and the direct model procedures, we fit models that predict the cubed roots of the wages. This transformation of wages to their cubed roots increased the fit of the model by decreasing the effect that extremely large wages had on the model. We also include two variable interaction terms in the models to also increase the fit.

The set of data that we use in our analysis is all reference persons who worked within the last twelve months taken from the entire 1999 ACS. The total number of records is 78,851 in which 9,667 or 12.3% are missing a value for wages. We divide this data into three sets based on tenure: 1) reference persons who are owners with monthly mortgage payments, 2) reference persons who are owners with no mortgage payments, and 3) reference persons who are renters. We break the data into these three categories because each category contains information that is unique.

We use variables, such as tenure, in our imputation that have not yet been imputed in the current hot-deck procedure at the time that wages are being imputed. Our argument for using variables that may have missing values is that these variables have proven to be important in determining values for the missing wages. By leaving out the variable, the information from the records that have values for the variable is totally ignored. In the case of tenure, 0.12% of the records with a missing value for wages is also missing tenure. By not using tenure, the information for 99.88% of the records is not being used. Those reference persons who do not fall into one of the three categories or have a missing value for tenure are ignored in our analysis.

Table 1 gives the matching variables by rank that are used by our flexible matching procedure for imputing wages for owners who have a mortgage. The reason the table shows three sets of matching variables is that two of the variables used as matching variables are not given a value for missing data until wages have been imputed. These two variables are monthly mortgage payment and property value of the house. As can be seen, they are important matching variables when present. We let the flexible matching procedure find the default number of matching variables which is five. Matching variable names that are preceded by an asterisk indicate that these variables are also used in the current hot-deck imputation procedure. The number in parenthesis at the end of each column title is the number of data records needing imputation for wages in that category. Table 2 shows the same information for owners who do not have a mortgage and Table 3 shows the information for renters.

Table 1. Matching Variables for Owners with a Monthly Mortgage Payments

Rank	**Monthly Mortgage Payment Present (2,748)**	**Monthly Mortgage Payment Missing (514)**	**Monthly Mortgage Payment and Property Value Missing (352)**
1	* Self-employed	Property Value	* Self-employed
2	Monthly Mortgage Payment	* Self-employed	Level of Education
3	* # of Weeks Worked per Year	* # of Hours Worked per Week	* # of Weeks Worked per Year
4	Level of Education	* # of Weeks Worked per Year	* # of Hours Worked per Week
5	* # of Hours Worked per Week	Level of Education	* Sex of Person

Table 2. Matching Variables for Owners with No Mortgage Payments

Rank	Property Value Present (1,327)	Property Value Missing (783)
1	* Self-employed	* Self-employed
2	* # of Hours Worked per Week	* # of Hours Worked per Week
3	Property Value	Level of Education
4	* # of Weeks Worked per Year	* # of Weeks Worked per Year
5	Level of Education	*Sex of Person

Table 3. Matching Variables for Renters

Rank	Monthly Rent Payment Present (3,223)	Monthly Rent Payment Missing (335)
1	* Self-employed	* Self-employed
2	* # of Weeks Worked per Year	* # of Weeks Worked per Year
3	Monthly Rent Payment	Level of Education
4	* # of Hours Worked per Week	* # of Hours Worked per Week
5	Level of Education	* Age of Person

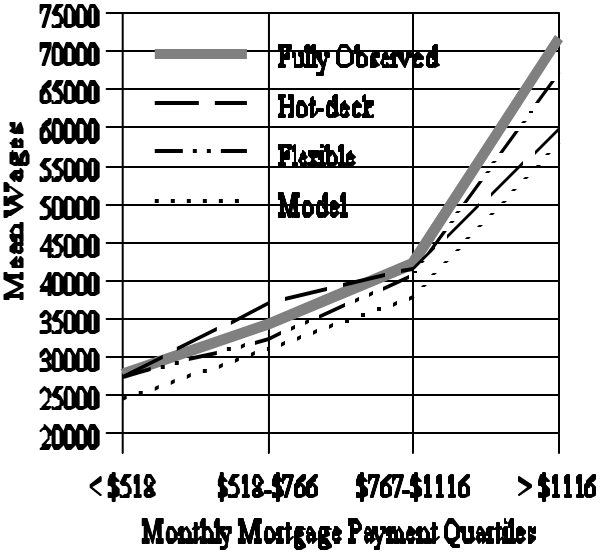
Other variables that are not picked as matching variables by our flexible matching procedure are mode of data collection (mail, CATI or CAPI), number of persons in the household, marital status, ethnicity, race, citizenship, employment status, and MSA status. One variable that is used by the current hot-deck method for matching is the occupation of the person. This variable has twenty levels to it and is very cumbersome when trying to use it in fitting a model. We solve this problem by making, for each level of occupation, a separate indicator variable and introducing all of the indicator variables along with other possible predictor variables into a stepwise multiple regression procedure for each of the categories listed in the tables. We find that only two levels make the top ten list of predictor variables for any given category. We include these two variables from each category as possible matching variables in our flexible matching procedure and, as is shown in the tables, none of the occupation indicator variables are found to be one of the top five matching variables.

III. Results

First, we will look at the relationship between the wages of owners and their monthly mortgage payments. Figure 1 shows the mean wages for the fully observed data (Fully Observed) and the imputed data derived from the current hot-deck procedure used by the Census Bureau (Hot-deck), our flexible matching imputation procedure (Flexible), and directly from the fitted regression models (Model). Using the fully observed data, we calculate quartiles for the monthly mortgage payment values. For each quartile, we plot the mean wage taken from the fully observed data and the imputed data from each method. We connect the points by either solid or broken lines in order to highlight the changes from one quartile to another. The only significant points are those on the vertical quartile lines.

In Figure 1, we see that the means of the imputed data from our flexible matching procedure come closest to the fully observed data means. What we find surprising is that, even though the means from the imputed data derived directly from the models change from one quartile to another in a pattern comparable to the fully observed data means, they are consistently lower than the fully observed data means. If we have a good-fitting model, we would hope to impute data that are close to that of the fully observed data in terms of measurements such as the mean values we have in Figure 1. We believe that one possible problem is the fit of the model. When

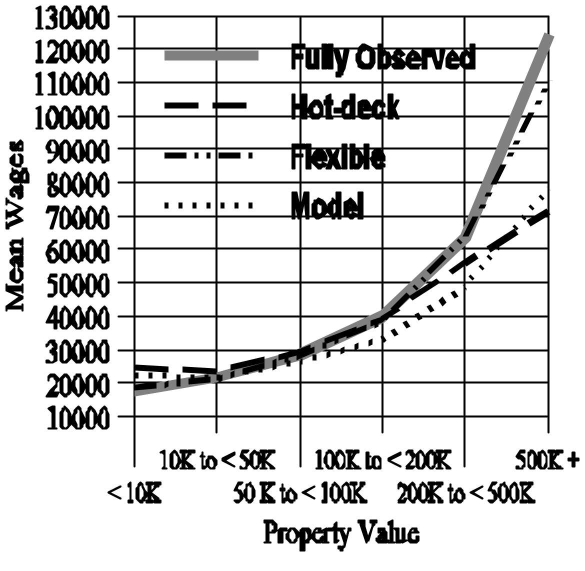
Figure 1. Mean Wages of Owners with Monthly Mortgage Payments



obtaining imputed values directly from a fitted regression model, it is important to have the model fit the data as well as possible. In fitting multiple regression models in order to predict the cubed root of wages, the best R-squared value we are able to obtain is 0.6 and the worst is 0.55. This means that for any given model at least forty percent of the variation in the data is not explained by the model. It is even worse if we do not transform wages to their cubed roots. We feel that this unexplained variation at least partially explains the results we are seeing in Figure 1 for the data imputed directly from the models.

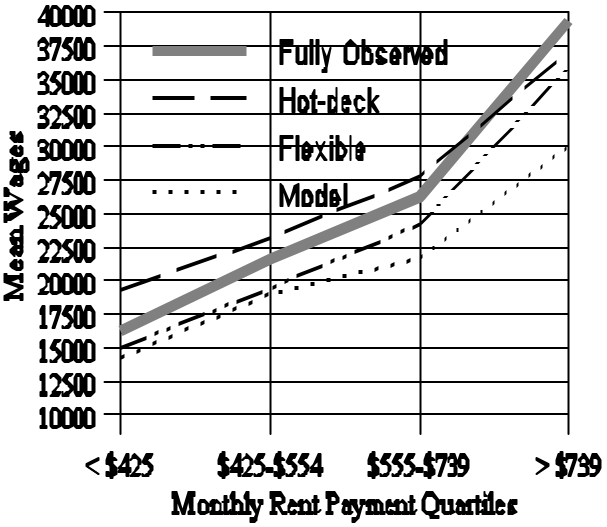
Now let us look at the relationship between the wages of owners and the property value of their houses. This comparison includes both owners with and without monthly mortgage payments. In Figure 2, we show the mean wages of the owners by six property value categories. As in Figure 1, we connect the points to highlight the changes. In this figure, we see that the means of the wages imputed using our flexible matching procedure are very close to the fully observed data means. Again, we do not seem to be nearly as accurate with the wages imputed directly from the models.

Figure 2. Mean Wages of Owners by Property Value



Finally, we look at the relationship between the wages of renters and their monthly rent payments. Figure 3 displays the mean wages for renters by monthly rent payment quartiles. In this figure we find that, unlike the comparison for owners, the means of the imputed data from our flexible matching procedure are not closest to the means of the fully observed data. We see that the flexible matching imputation means change from one quartile to another at increments comparable to the fully observed data means, but they are consistently lower. We can explain this by showing that there is an underlying factor that is captured by our flexible matching procedure.

Figure 3. Mean Wages of Renters by Monthly Rent Payments

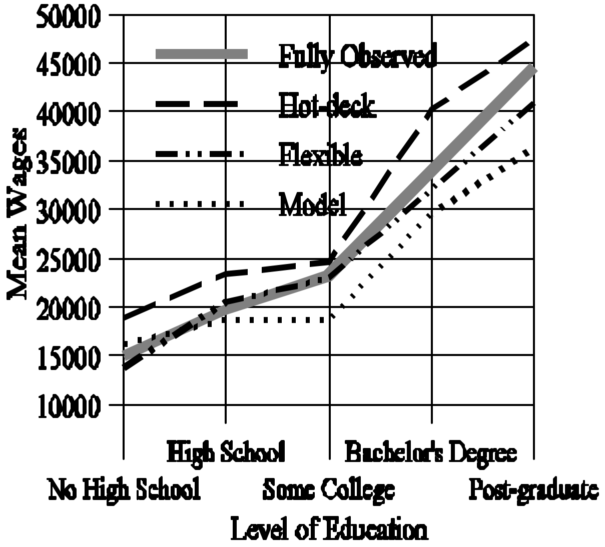


Looking back at Table 3, we see that the level of education obtained by the person is considered an important factor in predicting wages and this variable is not used in the current hot-deck procedure. Figure 4 gives the mean wages of renters by their level of education. The five levels of education displayed are 1) did not graduate from high school, 2) high school graduate but did not attend college, 3) attended college but did not receive a bachelor's degree, 4) received a bachelor's degree only, 5) received a post-graduate degree. We find in Figure 4 that the means of the wages imputed by our flexible matching procedure are closest to the fully observed data means for most of the education levels.

Now let us look at the distribution of education levels of renters for the fully observed data and the imputed data. Figure 5 shows us that the percentage of persons with no more than a high school education for the imputed data is higher (52%) than that of the fully observed data (34%).

Getting back to Figure 3, we reason that the means of the wages imputed by our flexible matching procedure are consistently lower than the fully observed

Figure 4. Mean Wages of Renters by Level of Education



data means because there is a higher percentage of persons without any college education for the imputed data. As can be seen in Figure 4, persons with a lower level of education have, on the average, lower wages. We can show that our reasoning is correct by adjusting the fully observed data so that the percentage of persons for each education level is equal to that of the imputed data. For each person in the fully observed data set with education level *m*, we give an adjustment weight that is equal to the proportion of persons found with education level *m* for the imputed data divided by the proportion of persons found with education level *m* for the fully observed data. Figure 6 provides the same information as Figure 3 with the adjusted fully observed data means for wages displayed. We can clearly see that, when the proportion of persons for each level of education are the same between the fully observed and the imputed data, the means of the wages imputed by our flexible matching procedure are very close to the fully observed data means. The reason for the consistently lower means shown in Figure 3 is because the proportion of persons within each education level are not the same.

Figure 6. Mean Wages of Renters Monthly Rent Payments

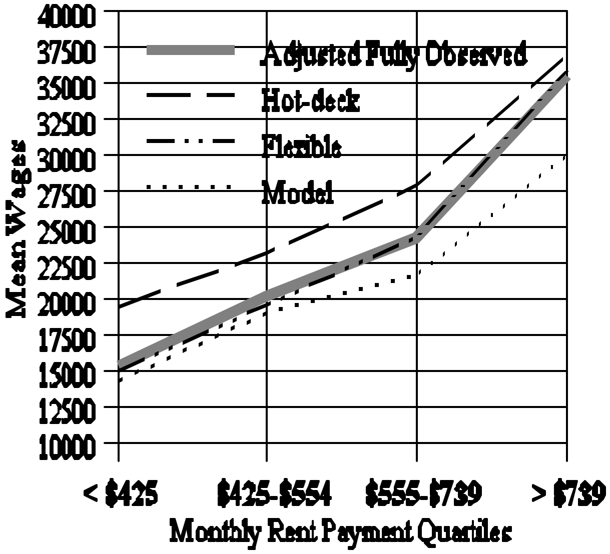
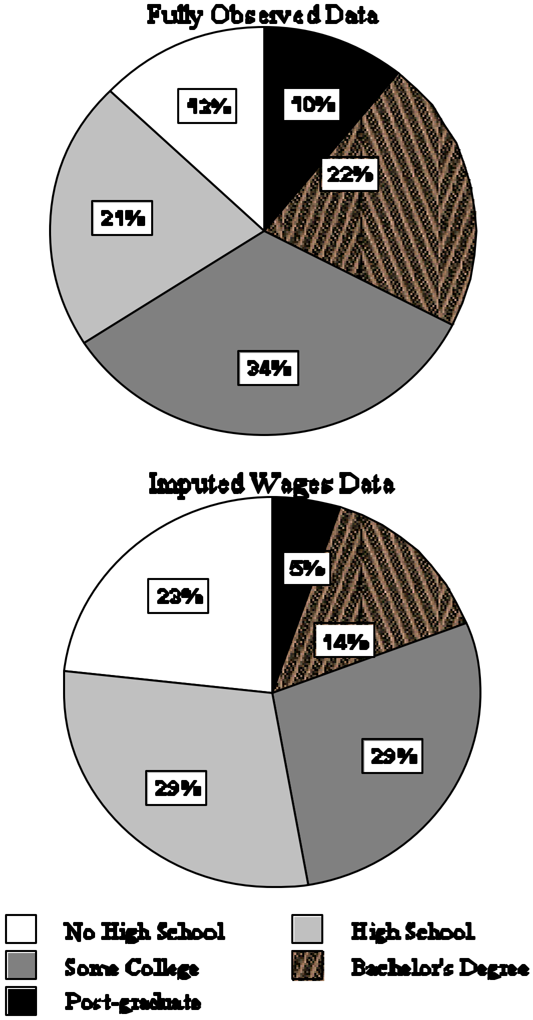


Figure 5. Distribution of Renters by Level of Education



IV. Summary

We conducted this research with the idea of trying to find a method that will provide imputed values that maintain the multi-variable relationships found in the nonmissing data. The following summarizes our findings.

We find that our flexible matching imputation procedure is able to determine matching variables that have the strongest effect on finding values for missing wages. By fitting a group of multivariate linear regression models, our procedure is able to keep the variables that provide the most information in determining values for wages and leave out those that do not make a significant contribution. Because these variables are used as matching variables for finding donors within a hot-deck imputation setup, our procedure is able to come closest in regards to maintaining the multi-variable relationships involving these variables. We witness this in our analysis of the relationships between wages and monthly mortgage payments and between wages and property value for owners. We also see this in the relationships between wages, monthly rent payments and the level of education for renters.

It appears that the current hot-deck procedure will only use matching variables that do not contain missing values or whose missing values have already been replaced. As seen with the monthly payment variables and property value, this prevents some of the more important matching variables from being used in the hot-deck imputation procedure when the information is available. Our flexible matching procedure is designed to find sets of matching variables based on the presence of the variables. For example, in the case of renters we see that the monthly rent payment is an important matching variable when imputing for wages. Because it is possible for the monthly rent payment to also be missing, our procedure provides a set of matching variables to use when both wages and the monthly rent payment are missing. This allows us to match a missing data record to a donor record using the monthly rent payment variable when it is available. This helps us to maintain the strong relationship between the wages of renters and their monthly rent payments.

We also examine the imputed wages that are obtained directly from fitted multiple regression models. The distributions of the wages imputed directly from the models are not as close to the distributions for the fully observed data as we had hoped. The means of the imputed wages tend to be lower than expected. When imputing for ages of reference persons using 1990 Census data (Williams, 1998), we find that imputing values directly from fitted models can work well if two things happen. First, the model has to fit the data well. The regression models that we use for imputing wages for the 1999 ACS do not fit as well as some of the models we use for imputing ages for the 1990 Census. Second, a method of including the variation found in the fully observed data needs to be added. This can be partially accomplished if randomly selected residuals from the model are added to the imputed values. Since this makes imputing values directly from the fitted models more difficult, we avoided this in our analysis, but plan to include it in future comparisons.

A possible shortcoming of our comparisons involves comparing the estimates from two populations, the fully observed and the imputed, which may have different distributions because the imputed data is not missing at random. This is seen with education level for renters. Once the fully observed data is adjusted so that the distribution of the levels of education is that of the imputed data, a clearer picture of how well the imputation procedures perform can be made. A good way to overcome this problem is to simulate missing data at random, impute the missing data using each of the methods, and compare the results to the actual data. We plan to use this approach in future research.

References

Long, Kimberly (1992), "Flexible Matching Imputation in the American Housing Survey," in *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 221-226.

Stiller, John G. and Dalzell, Donald R. (1998), "Hot-deck Imputation with SAS Arrays and Macros for Large Surveys," in *proceedings of the Twenty-Third Annual SAS Users Group International Conference*, pp. 1378-1383.

Williams, Todd R. (1998), "Imputing Person Age for the 2000 Census Short Form: A Model-Based Approach," in *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 680-685.

* Disclaimer

This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.