Imputation via Triangular Regression-Based Hot Deck[*][†]

Scott Susin
U.S. Census Bureau[‡]

Abstract

In principle, hot deck imputation methods preserve means and variances, and can also preserve covariances with other variables included in the allocation matrix. In practice, dimensionality problems arise quickly as predictive variables are added and allocation matrix cells become small, undermining the hot deck's theoretical advantages. Predictive-mean nearest-neighbor imputation avoids dimensionality problems, but can reduce the variance. A combination method is described: using the predicted values from a set of sequential, triangular regressions to form hot deck matrices. Triangularity allows the inclusion of predictive variables that are themselves subject to non-response. The method enables the rapid development of allocation schemes, eliminates dimensionality problems, and aids in predictor selection. The implementation of this method in American Housing Survey income data is described and evaluated.

Keywords: Imputation; Allocation; Predictive Mean; Sequential Regressions; Hot Deck

1 Introduction

This paper describes the income imputation system developed for the 2005 American Housing Survey (AHS), called a triangular regression-based hot deck. The method breaks little new theoretical ground. In fact, it could easily be improved on, albeit at some cost in additional computation and complexity. Nonetheless, the method represents a good compromise between current U.S. Census Bureau methods (the hot deck) and more sophisticated imputation schemes (such as chained regressions and multiple imputation).

The method has many desirable properties. It eliminates dimensionality problems, requires only weak assumptions about the distribution of the data, allows for the flexible imposition of logical constraints, and is robust to model misspecification. Compared to hot deck methods, the imputation scheme described reduces the amount of work required to develop an allocation scheme and provides the analyst guidance in creating the scheme. In practice, it is also better than a hot deck at imputing data that reproduces the relationships among variables (covariances) present in the reported data.

2 Hot Deck Methods

The hot deck allocation (or imputation) method is widely used at the U.S. Census Bureau and other statistical agencies. In this method, the analyst specifies an allocation matrix based on characteristics thought to predict the variable being allocated. For example, in an allocation matrix predicting earnings, one cell might consist of white renters, aged 18-25, with a high school education. When earnings are not reported, they are imputed from the reported earnings of the last observation processed (typically geographically close) that falls in the same allocation cell.

The hot deck method has several advantages. First, is its processing simplicity: the hot deck requires a single pass through the data, and can be easily implemented in statistical packages such as SAS that process a single observation at a time rather than holding a complete data set in memory. Of course, with the rapid development of computing technology, this processing simplicity has become less and less important, but still matters for very large datasets, such as the decennial census. Second, the hot deck preserves the distribution of the data. Recipient cases will have the same mean and variance as the donors. Importantly, the hot deck imposes no distributional assumptions on the data. Hence, it will also preserve other features of the data, such as heavy tails or heaping (when data are reported in rounded numbers).[1]

In principle, hot decks can also preserve the relationship between the allocated variable and other variables. In practice, however, dimensionality problems arise quickly, sharply limiting the number of variables that can be used. The American Community Survey (ACS) earnings allocation matrix, for example, uses 6 variables, with 2-20 categories in each, generating 3000 cells in the matrix.[2] Too many cells are undesirable, because small cells increase the usage of starting values, and increase the chance that a single donor will be used multiple times. The necessity for collapsing smaller cells, a process that requires considerable time and effort, is an important drawback of the hot deck method.

Dimensionality problems imply that many important variables are routinely omitted from allocation matrices. Economists have criticized the Census Bureau for not including education, a variable with strong theoretical foundations, in the Current Population Survey (CPS) earnings allocation matrix

[*] This paper is released to inform interested parties of ongoing research and to encourage discussion. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

[†] Citation: Susin, Scott (2006). "Imputation via Triangular Regression-Based Hot Deck," 2006 Proceedings of the American Statistical Association, Section on Survey Research Methods [CD-ROM], Alexandria, VA: American Statistical Association: pp-pp.

[‡] U.S. Census Bureau, Housing and Household Economic Statistics Division, 4700 Silver Hill Road, Washington, DC 20233-8500; 301-763-5880; scott.j.susin@census.gov.

[1] The variance of the mean and other sample statistics calculated from the imputed data set will be understated, however (Rubin 1987)

[2] Occupation/class of worker (20 Categories) × Weeks worked (5) × Hours (3) × Age (5) × Sex (2)

(Lillard, Smith & Welch 1986). Education is not included in the ACS allocation matrix either, and it is easy to see why, since adding a four category education variable would expand the matrix to 12,000 cells! Many other variables strongly correlated with earnings are not included due to this dimensionality problem (for example, public assistance payments).

Omitting *any* variable correlated with the imputed variable is undesirable, since this will bias the correlation between the omitted and imputed variable towards zero. (see, e.g., Little (1988)). In regressions of earnings on education, for example, the coefficient on education will be biased, presumably towards zero, if education is omitted from the allocation matrix. Since the aim of statistical agencies is to produce general purpose allocations suitable for numerous different analyses, this is a strong argument for including as many variables as possible as allocation predictors, which is generally not possible with hot deck methods.

At the same time as it restricts the number of variables that can be used in the hot deck, the hot deck provides little guidance to the analyst creating the imputation matrix as to which variables should be used. Should education replace weeks worked in the earnings allocation matrix? Analysts typically rely on theoretical knowledge, knowledge of the literature, and intuition to make the choice. In principal, one could impute using several alternative matrices and examine summary statistics to compare them, or run auxiliary regressions to make such choices. However, this adds a substantial amount of additional work to what is already a cumbersome procedure.

3 Alternative Imputation Methods

Many alternatives to the hot deck have been proposed by statisticians. For a recent overview of various imputation methods, see Durrant (2005). The goal here is to find a method that is flexible (in the sense that it can be combined with a priori constraints on the data); simple to implement and justify to nonstatisticians; and that avoids placing parametric constraints on the data. Of course, the method should also reproduce the distribution of the reported data.

This study does not address methods for making valid inferences from statistics estimated from the final data set after imputation. Although it is possible to impute multiple values to each recipient case, in order to calculate standard errors that take into account the uncertainty due to imputation, that is not the focus here.

Multiple imputation, as proposed by Rubin (1987), entails much more than simply imputing several different values to each recipient. It also requires specifying a particular joint distribution of the data, typically multivariate normality. Methods which allow the distribution of the allocated variables to be determined by the data (nonparametrically), such as the hot deck, are often much more appealing. In the case of income data, multivariate normality is a particularly unattractive assumption, since the various components of income are clearly not distributed normally. Many income components, such as interest income, are constrained to be nonnegative, have a spike in the distribution at zero, and have heavy upper tails.[3] It must be noted, however, that multiple imputation does have the most fully developed body of theory.

Another alternative is predictive mean matching (Little 1988). This involves regressing the variable to be imputed on a vector of predictors (in the sample of donors with complete data). Next, predicted values are calculated for both the donors (with complete data) and recipients (with incomplete data). The donor with the closest predicted value to a particular recipient is chosen, and that donor's observed value (not the predicted value) is imputed to the recipient. This method is quite attractive, since it avoids the dimensionality problems discussed above and, by avoiding distributional assumptions, is likely to be robust to mispecification (Chen & Shao 2000). It also reproduces the distribution of the complete data. The drawback is that one case can easily be used as a donor multiple times. As as result, estimators calculated from the final data set may be inefficient (Durrant & Skinner 2006).

Finally, a number of authors have proposed using a series of equations to model the conditional distribution of each variable, in order to avoid multiple imputation's requirement of specifying a joint model of all the predictors and imputed variables. That is, instead of assuming, say, joint normality, one can specify a series of equations, mixing OLS regression with logit models, or anything else. This adds considerable flexibility to model nonnormal distributions as well as to impose logical consistency constraints, such as the fact that only homeowners can have a mortgage (Raghunathan, Lepkowski, van Hoewyk & Solenberger 2001, Buuren, Boshuizen & Knock 1999, Buuren, Brand, Groothuis-Oudshoorn & Rubin 2005). This study describes and tests a simple version of this type of imputation via sequential regression, combined with a version of the predictive mean matching approach.

4 Allocation Methodology

The problem here is impute a series of nine income variables (denoted $Y_i, i = 1 \ldots 9$): earnings, social security income, etc. Each of these variables may have missing data. The missing component of the data vectors are denoted Y^{miss}, and the reported components are Y^{obs}. These variables are to be imputed jointly, with the goal of preserving the covariance structure of the income components, as well as the relationship between Y_i and a set of predictor variables ($X_j, j = 1 \ldots J$) which do not have missing data. That is, the X_js have already been imputed using some simpler method.

The imputation method is a simple version of the chained equations method. Because the income components have a substantial fraction of zeros, and are mostly constrained to be nonnegative (with the exception of self-employment or business income), the imputation proceeds in two steps. First, we impute indicators for the receipt of each type of income (denoted $D_i, i = 1 \ldots 9$). Next, amounts (Y_i^{miss} are imputed for the cases with $D_i^{miss} = 1$. An example of a receipt variables is employment, with earnings being the corresponding amount

[3] Schafer & Olsen (1999) discuss alternatives to multiple imputation for this case.

variable.

The method can be summarized as follows:

1. Estimate a set of nine regressions predicting D in the sample of completely reported data. These regressions are sequential and triangular (see below).[4]
2. Split the observed data into hot deck cells using the predicted values from each regression in (1). Choose cutpoints that put approximately 500 observations in each cell, creating one hotdeck per income type.[5]
3. Apply the regression coefficients from (1) and the cutpoints in (2) to the cases with missing data, thus assigning each piece of missing data in each case to a hot deck cell.
4. Impute missing data using the nine hot decks in the usual manner.

In order to preserve the covariance matrix of the income amounts, it would be desirable to use other income variables as predictors. However, since all the income variables contain missing values, there is a problem of circular dependence: social security cannot be imputed until earnings are computed, and vice versa. To overcome this problem, we specify a set of sequential, triangular equations:

$$\begin{aligned}
D_1^{obs} &= f(X^{obs}) \\
D_2^{obs} &= f(X^{obs}, D_1^{obs}) \\
D_3^{obs} &= f(X^{obs}, D_1^{obs}, D_2^{obs}) \\
&\dots \\
D_9^{obs} &= f(X^{obs}, D_1^{obs}, D_2^{obs}, \dots, D_8^{obs}).
\end{aligned}$$

Table 1a shows the variables in X for the receipt equations. These equations are estimated using OLS regression in the set of cases with completely reported income data. Receipt of the first income type is imputed using only the X variables. Receipt of the second income type is imputed using X plus the first income type, and so on. The equations are ordered by the R^2 from the regression of D_i on the X variables.

Table 1a: Explanatory Variables in Typical Receipt Regression Model

Working (2) X Sex (2) X Race (2)
Sex (2) * Age (5)
Working (2) X Tenure (2) X Housing Cost (4)
Working (2) X Family Type (3) X Relat. to hhdr (5)
Working (2) X Kids (4) X Relat. to hhdr (5)
Working (2) X Sex (2) X Education (4)
Working (2) X Citizenship (2)
Receipt indicators (0 to 8 dummies)

Then, for each income type, the reported data is split into a series of hot deck cells based on the predicted values from each regression, with the cutpoints chosen to put about 500 observations in each cell. For example, the 500 cases with the highest predicted earnings go into the first hot deck cell. The next 500 cases go into the second cell, and so on. The process is repeated for the other income types, each of which has its own hot deck.

Finally, imputation proceeds in the usual hot deck manner. The same coefficients (from the regression on the observed data) are applied to the nonreporters (cases with missing data). The missing values are filled from the donor case that is within the matching hot deck cell and is most recent in sort order (typically a case close in physical proximity).

This method of triangular equations is recommended by Raghunathan et al. (2001) (who call them "sequential" equations). An alternative would be to impute a set of starting values that don't condition on the other income types (perhaps using only X), and then iterate a number of times though $D_i = f(X, D_{-1})$, where D_{-1} are all the income receipt indicators other than D_i. We don't use this more sophisticated alternative because it would require multiple passes through the data, or the use of specialize software.

Having imputed the receipt indicators, the amount variables are imputed next. The allocation proceeds in a similar manner, except that we can now condition on the receipt variables, and the regressions are confined to those with positive amounts ($Y_i^{obs} > 0$). Specifically, a set of equations are estimated as follows:

$$\begin{aligned}
Y_1^{obs} &= f(X^{obs}, D^{obs}) \\
Y_2^{obs} &= f(X^{obs}, Y_1^{obs}, D^{obs}) \\
Y_3^{obs} &= f(X^{obs}, Y_1^{obs}, Y_2^{obs}, D^{obs}) \\
&\dots \\
Y_9^{obs} &= f(X^{obs}, Y_1^{obs}, Y_2^{obs}, \dots, Y_8^{obs}, D^{obs}).
\end{aligned}$$

As with the receipt variables, these equations are used to create a set of nine hot decks. The X variables for the amount equations are shown in Table 1b.

Table 1b: Explanatory Variables in Typical Amount Regression Model

Marital Status (4)
Race (2)
Sex (2) X Age (5)
Tenure (2) X Housing cost (4)
Family Type (3) X Relationship to householder (5)
Kids (4) X Relationship to housholder (5)
Sex (2) X Education (4)
Working (2)
Income Receipt Indicators (8 dummies)
Income Amounts (0 to 8 amounts)

The previous set of equations is used only for those with for those with two or more missing amounts. Since it is common for only one amount to be missing, these cases are treated

[4] Although logits or probits would be more efficient, OLS regression is used here. OLS is still consistent and I prefer its functional form assumptions (the additivity of indicator coefficients) in this context.

[5] Putting 500 donors in each cell is basically an arbitrary number, chosen to limit the reuse of donor cases.

separately. For them, a different set of nine equations are estimated:

$$Y_i^{obs} = f(X^{obs}, D^{obs}, Y_{-1}^{obs}), Y_i^{obs} > 0$$

and these cases get their own set of nine hot decks.[6]

Cases with no reported income data (neither amounts nor receipts) are treated separately. For them, we aim to preserve the covariance matrix not by conditioning on the other income variables (since there aren't any variables to condition on), but by jointly allocating all the income variables, so that a recipient receives all its data from a single donor. For these cases, a traditional hot deck is created, since there is no obvious way to combine a set of nine separate equations and hot decks into a single matrix.

5 Comparing Reported and Imputed Data

Testing the imputations of the receipt variables is straightforward, as is testing the imputations of amount variables conditional on receipt. In both cases, donors are simply compared to recipients. However, for some purposes, we would also like to compare the amount variables without conditioning on receipt, that is, filling in zeros for those who do not receive a particular type of income. In particular, this kind of information is needed to compare the covariance matrices of the imputed and reported data. In the case of amounts-with-zeros, the right comparison is not obvious, so this section discusses all three comparisons in detail.

Consider a simple example with only one income type, which will be called "wages" for concreteness. Wage information is collected using two questions: "Are you employed?" and "What are your wages?" (asked of the employed). There are three possible patterns of nonresponse: (1) Answer employment and answer wages (if employed), (2) Answer employment, refuse wages, and (3) Refuse employment. In the last case, those who do not answer the employment question are not asked the wage question. Notice that the employment rate for group 2 is 100 percent, since only the employed get the opportunity to refuse the wage question.

Employment is imputed using as donors all who answered the employment question (groups 1 and 2) as donors for group 3 who refused to answer the question. Let the employment rate for the donors (groups 1 and 2 combined) be E. Assuming that there are no systematic differences between the donors and recipients, the expected value of group 3's employment rate is also E. Hence, to test whether the allocations are working, the donors should be compared to the recipients, within each cell of the imputation matrix.

Wages are imputed only for the employed (others have zero wages). In this case the donors are the employed members of group 1, who answered both questions. The recipients are the employed members of groups 2 and 3. Let average wages for the employed members of group 1 (the donors) be W. Then, again assuming no systematic relationship between wages and the reporting of wages (no systematic differences between donors and recipients), the expected value of wages in groups 2 and 3 combined is also W. So, again, the allocations can be tested by comparing the donors to the recipients.

In order to test the employment and wage imputations jointly, we can fill in zero wages for the non-employed and compare average wages in groups 1 and 2 (those answering the employment question) to average wages in group 3 (those not answering employment and having both employment and wages imputed. These two quantities should be equal. The employment rate of groups 1 and 2 is E, as assumed above, which implies that the employment rate of group 3 is also E. Assuming that there are no systematic differences among the employed members of the three groups, the wages of the employed are all equal in expectation to W (the average wages in group 1, conditional on employment). Including the nonemployed, expected wages in group 1 and 2 combined are EW, and the expected wages of group 3 are the same.

Note that groups 1 and 2 cannot be considered "donors," since those in group 2 had wages imputed. All members of group 3 are recipients, but the recipients in group 2 are combined with the donors in group 1 for the sake of this comparison. Although this comparison may seem a little odd, the discussion in the previous paragraph has shown that it is a valid one (the average wages of groups 1 and 2 combined should equal the average wages of group 3).

We might naively consider comparing group 1 (complete responders) to groups 2 and 3 (all with either wages or employment allocated). Although this seems like a natural comparison, these two groups do not have the same expected wages. Expected wages in groups 2 and 3 combined are the weighted average of group 2 with 100 percent employment and group 3 with E employment, or $QW + (1 - Q)EW$, for some $Q \in [0, 1]$. Let the employment rate in group 1 be E_1. We know that $E_1 \leq E$. That is, group 1 must have an employment rate no greater than the rate for groups 1 and 2 combined, since $E_2 = 1$ (the employment rate for group 2 is 100 percent). Clearly, $E_1W < QW + (1 - Q)EW$, since $E_1W < W$ and $E_1W < EW$. So the naive comparison is invalid. By including group 2 (with 100 percent employment) among the recipients, we inflate that groups average wages relative to the the complete responders.

6 Results

6.1 Means and Standard Deviations

The method described above is used on the American Housing Survey (AHS), starting with the 2005 survey. The AHS is a nationally representative survey which includes a battery of income questions quite similar to the U. S. Census Bureau's American Community Survey (ACS).[7]

Table 2 compares the means and standard deviations of the reported data and the imputed data in the AHS, labeled re-

[6] In principle a separate set of nine equations and hot decks could be created for each pattern of missing data. However, this would require 1023 regressions and hot decks. This is one reason for using a set of triangular equations.

[7] For a description of these data sets see
http://www.census.gov/hhes/www/housing/ahs/ahs.html and
http://www.census.gov/acs/www/.

Table 2: Reported and Imputed Means and Standard Deviations

	Reporters: Unadjusted				Nonreporters[a]: Unadjusted			
	Mean	Std Dev	Std Error	N	Mean	Std Dev	Std Error	N
SS	9,311	5,729	59	9,461	9,487	6,357	112	3,202
Wages	37,797	43,033	229	35,351	37,281	49,517	501	9,756
Retirement	15,632	18,877	279	4,566	14,370	15,232	411	1,375
Interest, dividends, rental	11,468	41,754	572	5,330	13,352	45,769	1,019	2,019
SSI	5,064	4,265	118	1,306	4,268	3,760	243	240
Welfare	2,786	3,030	117	670	2,099	2,107	175	145
Workers' Comp.	8,629	9,323	233	1,604	9,741	11,720	655	320
Self-emp.	31,747	62,176	1,037	3,593	31,232	70,202	2,005	1,226
Other	6,754	12,348	276	2,000	5,704	6,923	407	290

NOTE: [a]. Nonreporters had the amount indicated in the table imputed but also reported at least one receipt (complete nonreporters are excluded). Zeros and edited data are excluded.

Table 3: Regressions of Log Annual Earnings. Coefficients (SEs in parentheses)

	2004 ACS		2003 AHS		2005 AHS	
	Reported	Imputed	Reported	Imputed	Reported	Imputed
Intercept	7.18	7.56	7.24	7.97	7.21	7.28
	(0.01)	(0.02)	(0.04)	(0.07)	(0.03)	(0.06)
Years of Education	0.146	0.109	0.148	0.084	0.148	0.134
	(0.001)	(0.001)	(0.002)	(0.005)	(0.002)	(0.004)
Experience	0.107	0.103	0.097	0.086	0.106	0.104
	(0.000)	(0.001)	(0.001)	(0.002)	(0.001)	(0.002)
Experience squared X 1000	-1.860	-1.630	-1.680	-1.380	-1.880	-1.770
	(0.007)	(0.016)	(0.024)	(0.044)	(0.027)	(0.047)
Female	-0.483	-0.430	-0.493	-0.392	-0.469	-0.470
	(0.003)	(0.008)	(0.012)	(0.025)	(0.011)	(0.022)
Black	-0.122	-0.106	-0.096	-0.141	-0.052	-0.064
	(0.005)	(0.010)	(0.019)	(0.038)	(0.019)	(0.031)
Hispanic	-0.033	-0.032	0.031	-0.161	0.052	0.017
	(0.005)	(0.011)	(0.019)	(0.041)	(0.017)	(0.035)
N	534,320	89,732	42,771	16,664	36,296	11,574

Table 4: Selected Correlations Between Income Components. Correlation Coefficients with the Largest Differences Between Reported and Nonreported

	Wages/ Social Security	Wages/ Retirement	Wages/ Self-employment	Wages/ Interest	Social Security/ Retirement
AHS					
Reporters	-0.47	-0.26	-0.15	-0.06	0.43
Nonreporters	-0.56	-0.31	-0.07	-0.11	0.44
Difference	-0.09	-0.05	0.08	-0.05	0.01
ACS					
Reporters	-0.47	-0.29	-0.21	-0.10	0.44
Nonreporters	-0.37	-0.16	-0.03	0.00	0.36
Difference	0.10	0.13	0.18	0.10	-0.08

porters and nonreporters. Reporters are those who answered all the income amount questions. Nonreporters did not respond to at least one amount question. Those who answered no income questions at all (neither receipt nor amount) are excluded, because this group was imputed using a different method (a traditional hot deck).

In general the means and standard deviations in the imputed data are fairly close to the data from reporters. Four of the differences in means are statistically significant at the 5 percent level, but none are larger than $2,000 (interest: $11,468 vs. $13,352) or 25 percent (welfare: $2,786 vs. $2,099) different. Of course, we would expect the means and SDs to be different if the characteristics of reporters and nonreporters differ systematically. Hence, the top panel suggests that any systematic difference in response rates is relatively minor.

6.2 Wage Regression

An important goal of an imputation system is to preserve the relationship between the allocated variables and other variables in the data set. As a check of this, Table 3 estimates basic earnings regressions using the 2003 AHS, the 2004 ACS, and the 2005 AHS. Comparing the regressions on the reported data, many of the coefficients are remarkably close in the three data sets: the constant, education, experience (age - education - 6), experience squared, and female. Only the coefficients on Black and Hispanic differ appreciably.

The main comparison is between the regressions in the reported and imputed samples. The focus is on the education coefficient, the "return to education," which is the target of much attention from labor economists. In the 2004 ACS, the coefficient on education is 0.146 in the reported data but only 0.109 in the imputed data. The difference is statistically significant, and substantively large. This result is not surprising, since education is not included in the ACS allocation matrix. As noted above, it is impractical to include more than a limited number of variables in a traditional hot deck such as the ACS uses. The 2003 AHS, which also uses a hot deck that lacks an education variable, shows an even bigger difference. The return to education falls from 0.148 in the reported data to 0.084 in the imputed data. In the 2005 AHS, which uses the simple chained equations method discussed above, the coefficients are 0.148 (reported) and 0.134 (imputed), the closest of the three, although still showing some signs of bias.

The female coefficient shows a similar pattern, which is a little surprising since all three allocation methods use a sex variable (although the 2003 AHS combined female and minority into a single variable). The 2004 ACS imputed female dummy is 0.05 lower than the dummy in the reported regression, 0.10 lower in the 2003 AHS, but in the 2005 AHS the coefficient is almost exactly the same in both the imputed and reported data. Again we see that the triangular equations methods comes closest to reproducing the relationships in the original data, because it takes into account many more variables in determining the imputations.

6.3 Covariance Matrix

In both data sets, the covariance matrices for reporters and nonreporters are fairly similar.[8] Also noteworthy is the similarity of the correlation matrices of the reporters across data sets, with the exception of correlations with the interest variable.[9] In almost every case, the difference between reporters and nonreporters is smaller in the AHS than in the ACS.

Reporters are defined as those who answered all the receipt questions. They are compared with nonreporters, who had at least one receipt imputed. As discussed above, this is the correct comparison, rather than a naive comparison of donors versus recipients. In addition, cases with any edited data and complete non-reporters are excluded, as discussed previously.

Table 4 summarizes the correlation matrices, displaying the correlations for the five pairs of variables with the largest differences between the reporters and nonreporters (in either data

[8] The complete correlation matrices are available from the author or from www.huduser.org.

[9] The fraction reporting the receipt of interest, dividends or rental income is higher in the ACS than in the AHS, perhaps driven by differences in the interview modes (predominantly mailed questionnaires in the ACS and personal interviews in the AHS).

set). The AHS imputation system always comes closer to reproducing the correlations than the ACS. The improvement is smallest for Social Security versus wages (a difference of 0.08 in the AHS and -0.11 in the ACS). It is largest for retirement income versus salary (0.06 in the AHS; -0.13 in the ACS) and retirement income versus Social Security (-0.01 in the AHS; 0.08 in the ACS).

7 Conclusion

This paper has described an income imputation system which uses a triangular sequence of regression equations to define a series of hot decks. Although this method is less efficient than some more sophisticated alternatives described in the literature (because only a single pass through the data is made, rather than iterating) it is simpler to implement. We showed that this imputation method is capable of reproducing the distribution of the reported data fairly closely, if not perfectly. More importantly, the method was able to reproduce the correlations of the imputed variables with each other, and with other variables in the data set.

A number of comparisons suggest that this method represents an improvement over methods currently used at the U.S. Census Bureau. First, the paper showed that there is substantial bias when a wage regression, a model commonly used by labor economists, is estimated in wage data imputed using conventional Census Bureau methods.[10] When this regression is estimated in the 2003 AHS and 2004 ACS, both of which use a conventional hot deck, the return to education and the female wage penalty are biased towards zero. This is unsurprising: the dimensionality problems inherent in traditional hot deck methods sharply limited the number of variables that can be included in the hot deck. Neither data set included education in its hot deck, and the 2003 AHS did not include a sex variable directly. Regression-based hot decks, however, place no such limits on the number of variables that can be included. When implemented in the 2005 AHS, this methods was able to almost eliminate the bias.

Both methods do fairly well in reproducing the correlations among the various income variables. The ACS's traditional hot deck method did much better than might be expected, given that no explicit attempt is made to reproduce the correlation structure. Apparently the variables included in the hot deck are strong enough predictors that the correlation matrix for nonreporters is fairly similar to that in the reported data. However, for almost all correlation coefficients, the regression-based hot deck used in the 2005 AHS came closer to reproducing the correlation matrix of the reporters.

It should be noted that imputing income is harder in the AHS than in the ACS. Since the AHS has a sample size less than a tenth that of the ACS, the hot deck is constrained to have fewer cells, limiting the ability to find close matches for nonreporters. In addition, the AHS has fewer variables available for predicting income than the ACS. In particular the AHS has few labor force variables, such as occupation. The ability of the hot deck based on a sequence of triangular regressions to do "more with less" represents an impressive achievement.

[10] Including cases with an imputed dependent variable is a poor practice which will generally result in biased estimates, but it is a common practice.

References

Buuren, S. V., Boshuizen, H. C. & Knock, D. L. (1999), 'Multiple imputation of missing blood pressure covariates in survival analysis', *Statistics in Medicine* **18**, 681–696.

Buuren, S. V., Brand, J. P. L., Groothuis-Oudshoorn, C. G. M. & Rubin, D. B. (2005), 'Fully conditional specification in multivariate imputation', *Journal of Statistical Computation and Simulation* . forthcoming.

Chen, J. & Shao, J. (2000), 'Nearest neighbor imputation for survey data', *Journal of Official Statistics* **16**(2), 113–131.

Durrant, G. B. (2005), 'Imputation methods for handling item-nonresponse in the social sciences: A methodological review.', *National Center for Research Methods Working Paper 002* . http://www.ncrm.ac.uk/publications/.

Durrant, G. B. & Skinner, C. (2006), 'Using missing data methods to correct for measurement error in a distribution function', *Survey Methodology* . forthcoming.

Lillard, L., Smith, J. P. & Welch, F. (1986), 'What do we really know about wages? the importance of nonreporting and census imputation', *Journal of Political Economy* **94**(3), 489–506.

Little, R. J. A. (1988), 'Missing-data adjustments in large surveys', *Journal of Business & Economic Statistics* **6**, 287–296.

Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J. & Solenberger, P. W. (2001), 'A multivariate technique for multiply imputing missing values using a sequence of regression models', *Survey Methodology* **27**, 85–95.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York.

Schafer, J. L. & Olsen, M. K. (1999), 'Modeling and imputation of semicontinuous survey variables', *Proceedings of the Federal Committee on Statistical Methodology Research Conference* . http://www.fcsm.gov/99papers/.