Implementing the HEDIS® Medicare Health Outcomes Survey

Applying Missing Data Imputation Methods to HOS Household Income Data

Prepared By:

Kazi Ahmed, PhD, Assistant Vice President, Analysis

Vivian Kong, MPH, Senior Health Care Analyst

Annie Chiu, BA, Health Care Analyst

Thomas Whiting, MPA, Senior Health Care Analyst

Prepared by the National Committee for Quality Assurance (NCQA)
for the Centers for Medicare and Medicaid Services (CMS)
Under Contract Number HHSM-500-2004-00015I-0001
OY2 Task 4.18b—Deliverable 411
April 22, 2009

The Centers for Medicare & Medicaid Services' Office of Research, Development, and Information (ORDI) strives to make information available to all. Nevertheless, portions of our files including charts, tables, and graphics may be difficult to read using assistive technology.

Persons with disabilities experiencing problems accessing portions of any file should contact ORDI through e-mail at ORDI_508_Compliance@cms.hhs.gov

IMPLEMENTING THE HEDIS
MEDICARE HEALTH OUTCOMES SURVEY

Imputation Analysis for HOS Income Data
Applying Missing Data Imputation Methods to HOS Household Income Data

TABLE OF CONTENTS

1.0 BACKGROUND

The Medicare Health Outcomes Survey (HOS), formerly known as the "Health of Seniors Project," is an annual survey of a randomly selected Medicare population in managed care settings, including Medicare recipients who are disabled and under 65 years of age. HOS was developed in 1997, in response to the fast-growing number of Medicare beneficiaries receiving health care through Medicare Advantage Organizations (MAO). In 2006, CMS implemented the Medicare HOS 2.0 for MAOs, which evaluates physical and mental health status using the Veterans RAND 12-Item Health Survey (VR-12).

HOS data are collected annually by NCQA-Certified HOS Survey Vendors, who administer the survey using a mixed mail/telephone protocol. For each cohort, data are collected at Baseline and at Follow-Up (two years after Baseline). For the Baseline survey, HOS uses a randomly selected sample size of 1,200 members from each MAO that reports HOS. To reduce burden on survey respondents, members who were sampled for and who returned a completed survey* the previous year are excluded from sampling. All eligible members are sampled for plans with fewer than 1,200 members.

***Note:** A **completed survey** is a survey that can be used to calculate physical or mental health summary scores.*

After the fielding cycle is complete, HOS survey vendors submit data to NCQA for a rigorous validation and cleaning process. In general, Baseline response rates have been 64% or higher; the response rate for Cohort 9 Baseline, the data to be used in this study, was 66.8%. Cohort 9 Baseline data, collected in 2006, comprised members from 203 MAOs.

Like most surveys, HOS data are affected by missing observations. While these are distributed across most variables, self-reported income consistently experiences a high rate of missing data. For example, HOS 2.0 questionnaire item 64 (Q64) asks respondents to report their annual household income. In the Cohort 9 dataset, the rate of missing observations for the income variable is more than 20%. Since income is an important variable for researchers, it is important to pay attention to the missing values and address the problem accordingly.

CMS wanted to provide researchers with strategies for imputing (i.e., filling in) missing income information. In consultation with CMS, NCQA addressed the issue of missing household income data in the HOS data set by designing and implementing a study to analyze missing income values and impute the missing values with annual income data from an external data source. In October 2008, NCQA completed a study and Dr. Joachim Bruess, NCQA Director of Analysis, and Myriam Bikah, NCQA Senior Heath Care Analyst, drafted a report—*Implementing the HEDIS Medicare Health Outcomes Survey: Imputation Analysis for HOS Income Data*—for submission to CMS. CMS asked NCQA to undertake a subsequent study to select and apply several missing data imputation methods—from simple to complex—employing statistical techniques that use existing information from the data set. This report is a result of that effort.

The Bruess and Bikah (2008) report addressed missing income information in the HOS Baseline data set by applying external imputation methods that use publicly available income information from Census 2000. Externally available zip code level income data were used to impute missing income data for the 65-and-older population. Since the HOS income variable is a group variable with various income categories, Census income data were grouped accordingly before imputation. Refer to Bruess and Bikah (2008) for methodological details.

In general, imputing data by using external information can be an approach fraught with limitations. While it is a viable option, it is also time consuming and requires additional resources. Because of these constraints, external imputation may not be the best method for many researchers planning to use the HOS data set. Though income data may be available from external sources, external data for other missing information may not be readily available (or might not ever be available); thus, it is important to turn to other methodological options that impute missing data using information available in the data set. Fortunately, there are many such methods available to analysts. Parametric and nonparametric imputation methods are now routinely available in many commonly used statistical software programs. This report provides a brief summary of some selected methods, applies them to the HOS data set for missing income values and compares and report results. It should be noted here that the income values imputed are not the actually dollars but the discrete values representing the income categories presented in the HOS from low to high. Values 1-9 were used to represent the nine HOS income categories. For example, the value "1" represented the HOS income group of "less $5,000," the value "2" represented the income group of "$5,000 to $9,999" and so on (see Table 2 for Household Annual Income categories).

2.0 INTRODUCTION TO MISSING DATA HANDLING

2.1 Missingness

Before discussing the application of missing data imputation methods, we must address the problem of missing data in data sets. Most generated data are missing a certain amount of information on one or more variables in the data set; the rate of missing data can range from less than 1%, to 50% or more. But no matter how large or how small this rate is, it can cause bias in parameter estimation or lead to inefficient analyses.

Bias and inefficiency depend on missing data—whether this occurs by chance or by design, there are a number of causes. Data may be missing because of participants' refusal to respond to certain questions on a survey; for example, respondents with high incomes may not want to disclose their income or someone receiving drug treatment may not want to disclose whether they have been arrested for drug use. Data may be missing because participants do not have the information to answer a question; for example, most respondents would not have an answer to, "How much money did you spend on gasoline last year?"

Data may be missing because participants overlooked questions during self-administration of the survey. Nonresponse is also likely to occur if a question does not apply to a participant; for example, asking, "How did you vote in the presidential election?" will generate missing values if participants are not eligible to vote or did not vote. Nonresponse may also result from information being intentionally deleted by a data collector to protect confidentiality.

In longitudinal studies, data may be missing because a participant may be deceased or has moved away from the address on file, or because of problems extracting data from the database.

Whatever the reason for missing data, most data sets are not complete, and data analysis must be run with missing data on one or more variables. Depending on the nature of the missing data, statistical analysis of data with missing values may yield biased estimates. In some cases, certain multivariate statistical analyses, such as factor analysis, discriminant analysis or structural equation modeling (SEM), cannot be performed unless the cases with missing values are either deleted or imputed. If cases are deleted, the loss of sample size may limit the use of these multivariate analytical procedures.

2.2 Missing Data Pattern

Selection of data imputation analysis methods also depends on missing data patterns. Little and Rubin discussed the following patterns: univariate missing data, multivariate missing data, monotone missing data, haphazard missing data, file matching and latent variable missing data. **Univariate missing data** occurs when data on only one variable in the data set is missing; **multivariate missing data** occurs when a subset of respondents does not answer a set of questions or items; **monotone missing data** occurs when data on subsequent waves of panel surveys are missing because of panel member attrition; **haphazard missing data** does not follow a particular pattern; data are missing because of **file matching** issues; data are missing simply because of the "unobserved," latent nature of the variable as seen in the factor analysis method. Missing data patterns are displayed in the tables below.

Table 1 Missing Data Patterns

1. Univariate				
Y1	*Y2*	*Y3*	*Y4*	*Y5*
2	5	1	23	1
5	4	1	22	0
4	3	1	22	1
6	2	0	24	1
3	4	0	21	
2	3	1	22	
2	2	1	23	
4	1	0	25	
5	4	1	21	

2. Multivariate				
Y1	*Y2*	*Y3*	*Y4*	*Y5*
2	5	1	2	3
5	4	2	2	3
4	3	1	3	2
6	2	3	1	2
3	4			
2	3			
2	2			
4	1			
5	4			

3. Monotone				
Y1	*Y2*	*Y3*	*Y4*	*Y5*
2	5	1	23	1
5	4	1	22	0
4	3	1	22	1
6	2	0	24	1
3	4	0	21	0
2	3	1	22	
2	2	1		
4	1			
5				

4. General				
Y1	*Y2*	*Y3*	*Y4*	*Y5*
2	5	1	23	1
5	4	1	22	0
4	3	1	22	1
6	2	0	24	1
3	4	0	21	
2	3	1	22	
2	2	1	23	
4	1	0	25	
5	4	1	21	

5. Latent Variable				
Y1	*Y2*			
2				
5				
4				
6				
3				
2				
2				
4				
5				

The HOS Cohort 9 Baseline data set provides 116,098 observations and 177 variables. The missing data pattern in this data set may be characterized as "multivariate and general." The monotone missing data pattern also existed when the Follow-Up was collected two years after the Baseline.

2.3 Assumptions of Missingness

The properties that distinguish missing data methods are defined by the nature of dependencies in mechanisms or the processes that lead to missing data. Such mechanisms are rooted in the assumption that missing data are missing completely at random (MCAR) or missing at random (MAR). If missing data are not MCAR or MAR, then they are missing systematically or are missing not at random (MNAR).

When missing data do not depend on values of a variable—for example, Y, missing or observed—then data are MCAR. This assumption is defined as:

$$P(\mathbf{R}|Y, \mathbf{X}) = P(\mathbf{R}|Y, \mathbf{X}^{obs}, \mathbf{X}^{mis}) = P(\mathbf{R}|\varphi)$$

where **R** is response indicators (i.e., $\mathbf{R}_j = 1$ if the j^{th} element of **X** is observed, and equals 0 otherwise) governed by parameter φ.

This assumption states that missingness is not related to any factor, known or unknown, in the study that generate the data (Horton and Kleinman, 2007); it does not mean that the pattern itself is random, but that missingness does not depend on data values.

When missing data for all variables are MCAR, there is no need to consider imputing missing values because the sample may be considered a simple random subsample of the original set of observations (Allison, 2002). When a subgroup of respondents refuses to provide income data, the refusal could still occur in the sample as MCAR.

Missing data are considered MAR when variable Y (income) is dependent on variable X (age), but within each X category there is no relationship between those with missing Y and those without. This assumption is defined as:

$$P(\mathbf{R}|Y, \mathbf{X}) = P(\mathbf{R}|Y, \mathbf{X}^{obs,} \varphi).$$

It states that missingness depends only on observed quantities, which may include outcomes and predictors. Data are MAR after controlling for missingness due to observed quantities (Horton and Kleinman, 2007). According to Collins, Schafer and Kam (2001), the MAR assumption may be made plausible by including a relatively rich set of predictors in the model. Others (Moon et al., 2006) have noted that including outcome information improves estimation of missing predictors. It may be noted that the assumption of MCAR can be formally tested against the alternate hypothesis of MAR (Little, 1988; Diggle et al., 2002).

Another assumption to consider is ignorability. MAR data could be non-ignorable when the parameters that describe the missing data process are unrelated to the parameters to be estimated. In other words, there is no need to model the missing data mechanism as part of the estimation process. According to Allison, MAR and ignorability are equivalent conditions, not distinct, and so there is no need to model the missing data mechanism when data are MAR. However, in rare situations when they are not equivalent, even model-based methods that assume ignorability work well.

When data are not MAR, the missing data mechanism is said to be non-ignorable (NINR [non-ignorable nonresponse] or MNAR). In that situation, the missing data mechanism must be modeled to obtain unbiased estimates of the parameters of interest. It should be noted that because the values of the missing data are unknown, the MAR condition is impossible to test. NINR models allow assessing the sensitivity of results to deviations from MAR missingness (Carpentar, Kenward and Vansteelandt, 2006). Unless additional information is available, it is impossible to test whether the MAR condition is present (Little and Rubin, 1987).

2.4 Approaches to Handling Missing Data

There are many different methods to handle missing data. Such methods may be grouped into the following non-mutually-exclusive categories (Little and Rubin, 2002), which range from a simple approach of no imputation to various probability-distribution–generated, model-based multiple imputation procedures, as discussed below.

1. Procedure Based on Completely Recorded Units

This procedure excludes all observations with missing values in any one set of variables used in analysis, and is commonly known as the **listwise deletion method**. A lot of information may be lost when this method is used. For example, if a data set has 30 variables and each variable has a 5% chance of missing data, applying the listwise deletion method will reduce the sample by 73%, leaving only 23% of the observed data values for analysis. If the chance of missing data per variable increases to 10%, then only 5% of the complete-case observations will be retained (100% – 10% = 90%, or 0.9; 0.9^{30} = 0.042391, or 4.2%; 4.2%/0.9 = 4.66%, or 5%).

In a data set with a sample of 1,000 observations, 5% translates to only 50 complete-case observations available for analysis.

Even with the potential of losing information, this method is simple and easy to perform and is less likely to yield biased and inefficient estimates when the sample size is large and the missingness is MCAR. However, when the sample size is small, the overall proportion of missing cases is large (more than 5 percent) and the missing data are not MCAR, but MAR or NMAR, then the complete-case sample size could be reduced considerably and the estimates derived with this small sample will be biased (large standard errors) and inefficient due to information loss, especially when making an inference for a subpopulation. Most software programs use this as the default method of handling missing data.

In this study, the results obtained from a listwise deletion method are used and compared against results from several missing data imputation methods described below.

Another non-imputation method is the **pairwise available-case method,** which applies when multivariate data analysis is involved. This method uses observation pairs that have non-missing values and is appropriate for estimating covariance and correlation matrices. Pairwise available-case estimates recover some of the information in partially recorded units that is lost by complete-case analysis. Under MCAR, data yield consistent estimates of covariance and correlation, but collectively, the estimates have severe limiting deficiencies in practical problems.

2. Weighting Adjustment Procedure

The **weighting adjustment procedure** is an extension of the complete-case analysis. Its aim is to reduce bias in an estimate obtained from a complete-case analysis (listwise deletion) by differentially weighting the complete cases to adjust for bias—similar to applying weight to units in a probability sample in order to make inferences about a finite population. This procedure involves dividing the sample into a number of weighting classes based on variables observed for respondents and nonrespondents. Weights are then calculated for each weighting class and applied to responding units in the class when a parameter (e.g., mean) is estimated.

A common variant of weighting adjustment methods is the **propensity score method.** Though weighted complete-case estimators are often easy to compute, computing their standard errors is complex. Computing difficulties can be overcome with special software programs that can handle methods based on Tyler Series expansions, balanced repeated replication or jack-knifing (i.e., re-sampling original data). We will report results from missing data imputation using the propensity score method.

3. Imputation-Based Procedure

While complete-case and available-case analyses do not use cases with missing values, there is an array of methods that handle missing data problems by imputing the values of missing variables. These methods include single imputation of values for each variable with missing values and imputing more than one value (also called **multiple imputation**), which allows an assessment of imputation uncertainty (Little and Rubin, 2002).

Imputation is a general and flexible approach to missing data issues. According to Little and Rubin (2002), "Imputations are means, or draw from a predictive distribution of the missing values, and require a method of creating a predictive distribution for the imputation based on the observed data (p 59)." This distribution can be generated by either applying a formal statistical model (e.g., multivariate normal) with explicit assumptions, or by applying an algorithm, which implies an underlying model.

Little and Rubin talk about two generic approaches to generating the predictive distribution: explicit and implicit modeling. Explicit approaches are based on explicit statistical models (e.g., multivariate normal). Implicit approaches focus on algorithms of imputing under implied underlying models. Explicit modeling methods are mean imputation, regression imputation and stochastic regression imputation. Implicit modeling methods are hot deck imputation, substitution, cold deck imputation and composite methods.

Explicit approach

Unconditional mean imputation substitutes missing values with means from responding units in the sample. **Conditional mean imputation** imputes the means of certain groups or weighting classes. **Regression imputation** replaces missing values with predicted values using information from observed or imputed variables. Mean imputation can be regarded as a special case of regression imputation, where predictor variables are dummy-indicator variables for cells within which the means are imputed. **Stochastic regression imputation** is similar to regression imputation: in addition to replacing the missing values with predicted values, residuals are drawn to assess the degree of uncertainty in the predicted value. Herzog and Rubin (1984) describe a

two-stage procedure that uses stochastic regression for both normal (normal linear regression with normally distributed residuals) and binary (as in logistic regression) outcomes.

Implicit approach

Hot deck imputation involves replacing the missing value with values drawn from similar responding cases. Variants of hot deck include simple random sampling with replacement, within adjustment cells, nearest neighbor and sequential ordered by a covariate. Substitution replaces non responding units with alternative units not selected in the initial sample. This method is used during fieldwork; during analysis, it should be treated as a case with imputed values. **Cold deck imputation** replaces a variable's missing value with a constant value acquired from an external source. **Composite methods** combine two or more methods; for example, hot deck and regression imputation can be combined by calculating predicted means from a regression and adding a residual randomly chosen from the empirical residuals to the predicted value when forming values for imputation.

4. Model-Based Procedure

Model-based procedures include a broad class of procedures generated by defining a model for the observed data. Inferences are based on the likelihood or posterior distribution under that model, with parameters estimated by procedures such as maximum likelihood. According to Little and Rubin (2002), this approach provides flexibility and the empirical means to display and evaluate the model assumptions. It also generates estimates of variance that take data incompleteness into account.

Multiple imputation (MI) replaces each missing value by a set of two or more imputed values. The set of values is ordered to impute values and form multiple (for example, five) completed data sets. Standard complete-data methods are used to analyze each data set. When imputations in these multiple data sets are based on draws from predictive distribution of the missing values under a specific model, the complete-data inferences can be combined to form one inference that properly reflects uncertainty resulting from nonresponse under that model. If the imputations are from two or more models of nonresponse, the combined inferences under the models can be contrasted across models to display the sensitivity of inference to models for nonresponse. This step is useful when non-ignorable nonresponse is being imputed.

3.0 APPLYING MISSING DATA IMPUTATION METHODS TO THE MISSING HOS HOUSEHOLD INCOME DATA

The Medicare HOS questionnaire has four major components: 1.) the VR-12, which is the core component; 2.) questions that gather information for case-mix and risk-adjustment; 3.) questions added by CMS that gather information required by the 1997 Balanced Budget Act; 4.) and questions that collect results for select HEDIS Effectiveness of Care measures. The survey was constructed to satisfy minimum psychometric standards necessary for group comparison. Items are scored and summarized into a physical component summary (PCS) and a mental component summary (MCS).

Baseline data from 2006 (Cohort 9), with 116,098 observations, were used for this study. Descriptive analysis showed that there were 23,405 observations with missing values for the annual household income question (20.16% of the total sample). For the purpose of this study, the following seven variables were used.

1. *General Health Status.* "In general, compared to other people your age, would you say that your health is: excellent, very good, good, fair, poor?"
2. *Gender.* "Are you male or female?"
3. *Age.* In years.
4. *Marital Status.* "What is your current marital status? Married, divorced, separated, widowed, never married."
5. *Home Ownership Status.* "Is the house or apartment you currently live in: owned or being bought by you, owned or being bought by someone in your family other than you, rented for money, not owned and one in which you live without payment of rent, none of the above?"
6. *Education.* "What is the highest grade or level of school that you have completed? 8th grade or less, some high school, but did not graduate, high school graduate or GED, some college or 2 year degree, 4 year college graduate, more than a year college degree."

7. *Annual Household Income.* "Which of the following categories best represents the combined income for all family members in your household for the past 12 months? Less than $5,000, $5,000–$9,999, $10,000–$19,999, $20,000–$29,999, $30,000–$39,999, $40,000–$49,999, $50,000–$79,999, $80,000–$99,999, $100,000 or more, Don't Know.,"

Table 2 presents the demographic characteristics of Cohort 9 respondents in the data set used in this study. The descriptive statistics provide a general background of the sample population for which we are attempting to impute the missing income values.

The variables described above were selected for imputation. Five of the seven variables were also used to test a multiple regression model (described below) with complete-case imputed income data sets using six different methods of imputation.

To display the actual value imputed for missing household income variable by different methods, we also include the first 30 observations of the income variables in the 6 data sets. In these first 30 cases, there were 8 missing values. Imputed values from these various methods are displayed in the table, as are the computed means and their respective standard errors for each imputed data set generated by the data imputation methods used in this study. The 30 observations and the descriptive statistics are displayed in Table 3.

A regression model, similar to the one used by Bruess and Bikah in their external income data imputation study, was tested to compare the results of imputation with different methods. Results from the regression analysis are displayed in Tables 4–12.

A multiple linear regression model was used to compare results. The general health status question ("In general, compared to other people your age, would you say that your health is… Excellent, Very good, Good, Fair, Poor") was the dependent variable, with age, education, gender and household income as independent variables. The linear regression model was used so results could be compared with the regression model-testing results from the Bruess and Bikah report, using external source income data. The PROC REG procedure in SAS was used to test the multiple linear model fitted by the ordinary least-squares method (OLS).

Regression analysis is the analysis of the relationship between one variable and a set of variables. The relationship is expressed as an equation that predicts a response variable (also called a dependent variable or criterion) from a function of regressor variables (also called independent variables, predictors, explanatory variables, factors or carriers) and parameters. Parameters are adjusted so that a measure of fit is optimized. For example, the equation for the *i*th observation might be

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where y_i is the response variable, x_i is a regressor variable, β_0 and β_1 are unknown parameters to be estimated and ϵ_i is an error term.

Table 2 Demographic Characteristics and General Health Status of Cohort 9 (2006) Respondents

Variables	%
Gender	
Male	40.69
Female	58.03
Missing	1.28
Age (Mean)	74 years
Education	
8th grade or less	10.74
Some high school, but did not graduate	15.59
High school graduate or GED	36.74
Some college or 2-year college	20.84
4-year college graduate	6.93
More than a 4-year college degree	7.04
Missing	2.12
Marital Status	
Married	53.64
Divorced	10.47
Separated	1.12
Widowed	28.89
Never married	4.14
Missing	1.73
Household Annual Income	
Less than $5,000	3.26
$5,000–$9,999	7.63
$10,000–$19,999	24.02
$20,000–$29,999	17.63
$30,000–$39,999	10.5
$40,000–$49,999	6.39
$50,000–$79,999	6.71
$80,000–$99,999	1.65
$100,000 or more	1.95
Missing	20.16
Race	
White	85.72
Black or African American	9.09
Hispanic	1.82
Asian	1.44
American Indian or Alaskan Native/Native Hawaiian or other Pacific Islander	0.19
Other	1.66
Unknown	0.08

Variables	%
Home Ownership Status	
Owned or being bought by you	68.94
Owned or being bought by someone in your family other than you	6.95
Rented for money	14.88
Not owned and one in which you live without payment of rent	1.6
None of the above	4.49
Missing	3.14
General Health Status	
Excellent	11.38
Very Good	27.12
Good	33.64
Fair	19.87
Poor	6.35
Missing	1.64

3.1 Applied Imputation Methods

This section describes the imputation methods applied to the missing household income data, which are: 1) listwise deletion; 2) overall mean (non-conditional) imputation; 3.) conditional mean imputation, 4) hot deck imputation; 5) propensity scores method; 6) imputed by chained equation (ICE) imputation; 7) Markov chain Monte Carlo (MCMC) imputation. In addition to these methods, we also replicated the external imputation method developed by Bruess and Bikah (Bruess and Bikah, 2008) by applying the CENSUS income data to our data set. The imputed data sets (once with 2% inflation adjusted rate and again for 3% inflation adjusted rate) were then used to run the mean and regression model tests. The test results are presented in tables below for comparison.

Methods 2–4 are **single imputation methods** because only one complete-case data set is generated. Methods 5–7 are **multiple imputation (MI)** methods because they are used to generate multiple complete-case data sets. The default number of multiple complete-case data

sets is five, which can be changed to any desired number. There is no agreement as to the right number of data sets one should generate, but some argue that more completed data sets are better for characterizing the variability introduced into the results through the imputation process (Horton and Kleinman, 2007).

MI is a three-step process to estimate incomplete data regression models (Rubin 1976). First, using a regression model, plausible values for missing observations are computed that reflect uncertainty about the nonresponse model and are used to impute the missing values. The process is repeated multiple times to create the number of "completed" data sets. The second step involves analyzing each complete-case data set. The final step involves combining the results (e.g., regression coefficients) by computing the mean of the resulting statistics (e.g., intercepts and beta-values of a regression model).

Three methods were chosen to test and display the results of missing data imputations. Table 3 presents the first 30 cases of the complete-case data sets produced by the imputation methods discussed above. Table 3 also includes the mean and standard errors of the mean for each complete-case data set generated by the imputation methods. Multiple data sets were used to test a regression model; results are presented in Tables 4–12.

3.2 Listwise Deletion Approach

This is the simplest of all the approaches to handling missing data. It excludes all observations in the analyses with no valid values for the income data. The PROC UNIVARIATE SAS statement

was used to apply the listwise deletion method to our data set. Listwise deletion is the default method for handling missing data in SAS.

3.3 Mean Imputation

Two types of mean imputation were used. In the first, the overall mean was imputed to the missing cases in the data set for the income variable (i.e., values 1-9 representing the nine income categories and not actual dollars). The mean income category value of 4.13 generated from PROC UNIVARIATE in the listwise deletion method was imputed for missing income values.

To impute the conditional mean, we first computed a variable called *newgrp,* which was a combination of various categories in three demographic variables (i.e., age, gender, education) deemed useful for this purpose. A collapsed age variable (≤ 64 = 0; 65–69 = 1; 70–74 = 2, 75–79 = 3, ≥ 80 = 4) was used. The categories of age (5), education (6) and gender (2) formed 60 categories (5x6x2). PROC UNIVARIATE in SAS was used to compute means for each of the 60 categories in the *newgrp* variable. The means were then imputed into the income variable.

Instead of manipulating the original income variable, we made a copy of the original variable, called *reinc*, which was then imputed. Tests were performed using this variable.

3.4 Hot Deck Imputation

Both SAS (macro) and SOLAS software programs were used to apply the hot deck method to impute the missing income categories. The SOLAS 3.20 is designed specifically for analysis of datasets with missing observations. SOLAS provides four distinct methods (group means; hot-deck imputation; last value carried forward; predicted mean imputation) by which analysts can perform single imputations and two distinct methods (predictive model based method; propensity score based method) for performing MI. SOLAS was used for hot-deck imputation and the propensity score based method (described below).

SOLAS generated imputed values by selecting from respondents who were similar with respect to a set of auxiliary variables (age, education, gender, marital status) associated with the income variable. Age and gender were complete with no missing values. Marital status and education had 2,003 and 2,459 missing cases, respectively, and were excluded from the hot deck imputation. SOLAS started the hot deck process by first sorting the data set with these variables. The process used the following rules: when more than one matching respondent was found, a respondent was randomly selected from the matching respondents; when no matching respondent was found, a random overall imputation was performed.

Applying the hot deck method on SOLAS was time-consuming. The HOS data set is a SAS data base created by SAS version 9.1 with the file extension *.sas7bdat*. Because SOLAS cannot handle SAS data sets created by SAS versions higher than 6, the SAS data set (with extension *.sas7bdat*) was first converted to an older version SAS database (version 6 with extension *.sd2*)

using the StatTransfer program. Because the HOS data set was so large, SOLAS took a long time to open the data base and more than six hours to compute the missing values with hot deck. ***Note:*** *To be released in 2009, the 64-bit version of SOLAS will significantly reduce delays in handling large data sets.*

The process generated a separate file with imputed data that had to the be converted back to *.sas7bdat* for testing (i.e., print the first 30 observations of data set with imputed values, means and standard errors and regression model testing).

In addition to the SOLAS hot deck imputation, we used a SAS macro developed by Lawrence Altmayer, of the U.S. Census Bureau, to apply the hot deck imputation method that implements an algorithm using data from other observations in the data set. Under this method, the imputation process is accomplished in two steps. First, two SAS macrovariables are created to store the number of observations per group (we used the HOS Contract Number variable in the data set) and the group number, then the hot deck imputation was applied to fill the missing values.

To accomplish this, a temporary flag array was created so that the new income variable could be retained. For a given group, the macro works first backward and then forward to impute a new value based on corresponding data from previous or subsequent observations in the group. Apart from imputing a two-digit value developed to identify the income categories for HOS, the main difference employed in customizing Altmayer's code was sorting the imputation dataset by zip code. The actual sort procedure is, in ascending order, by CTRLNUM (HOS Contract Number),

STATE, CITY and ZIP. Our decision to sort by zip code was based on the likelihood that similar income categories would live in comparable geographies.

Note: *Details of the steps in the SAS macro can be found in* Hot-Deck Imputation: A Simple DATA Step Approach *by Lawrence Altmayer (2002). The SAS macro we adapted appears in the Appendix.*

3.5 Propensity Score Method

We used SOLAS 3.20 to apply the propensity score MI method to impute missing values. The SOLAS **propensity score method** is a system in which an implicit model approach based on propensity scores and approximate Bayesian bootstrap is used to generate imputation. The propensity score is the estimated probability that a particular data element is missing. The missing data are filled in by sampling from cases that have a similar propensity to be missing. The underlying assumption about propensity score MI is that the nonresponse of an imputation variable can be explained by a set of covariates using a logistic regression model. The MIs are independent repetitions from a posterior predictive distribution for the missing data, given the observed data.

Variables are imputed from left to right through the data set, so values imputed for one variable can be used in the prediction model for a missing value occurring in the variable to the right. The system creates a temporary variable that will be used as the dependent variable in a logistic regression model. It is a response indicator and will equal 0 for every case in the imputation variable that is missing, and 1 otherwise. The independent variables for the model are a set of

fixed covariates assumed to be related to the variable we are imputing. The regression model allows modeling the missingness using the observed data.

Using the regression coefficients, SOLAS calculates the propensity that a subject would be a missing value in the variable in question. In other words, the propensity score is the conditional probability of missingness, given the vector of observed covariates.

Each missing data entry of the imputation variable *y* is imputed by values randomly drawn from a subset of observed values of *y* (i.e., its donor pool), with an assigned probability close to the missing data entry that is to be imputed. The donor pool defines a set of cases with observed values for that imputation variable.

For the propensity score method, gender, education, marital status, homeownership status and age were covariates. We requested, and the system generated, five multiple imputed data sets. Each data set was converted back to *.sas7bdata* for our regular statistical analysis.

3.6 Markov Chain Monte Carlo (MCMC) Method

The **Markov chain Monte Carlo (MCMC)** method is one of the three multiple imputation methods in SAS and is available under the PROC MI procedure. The method of choice depends on the type of missing data pattern. For monotone missing data patterns, either parametric regression method under the multivariate normal assumption can be used, or a nonparametric propensity score method. For an arbitrary missing pattern, an MCMC method under multivariate

normality assumption can be used (Schafer 1997). This method creates multiple imputations by using simulations from a Bayesian prediction distribution for normal data.

SAS PROC MI uses a three-step process. First, the imputation is carried out by PROC MI. Then, complete data methods are employed using any of the SAS procedures for complete data analysis (e.g., PROC GLM, GENMOD, PHREG, LOGISTIC). The BY statement repeats the analysis for each completed dataset. Finally, the results are combined using PROC MIANALYZE.

PROC MI incorporates a number of different imputation methods. For monotone missing patterns, the MCMC statement can be used either for all missing data or to impute enough data so that remaining missing data are monotone. The PROC MI with MCMC statement and PROC MIANALYZE were used to combine the results.

3.7 Imputation by Chained Equations (ICE) Method

Imputation by chained equations (ICE) is a code set developed by Patrick Royston for multiple imputation of missing data in STATA statistical software. ICE codes must be downloaded into the STATA software before imputation can begin. This step can be completed by using the "find" and "update" features in the STATA program. The download is free of cost. The ICE software provides support for categorical missing values (Royston 2005). Binary variables are predicted from other variables using logistic regression, while categorical variables with more levels use either a multinomial or ordered logistic regression.

There are two major approaches in multiple imputations. The first approach is based on the joint distribution of all variables considered for imputation or selected for imputing other variables;

this is used by SAS PROC MI and works very well with multivariate normal data. The second approach is based on each conditional density of a variable, given all other variables; this is the approach Royston's ICE program takes. The advantage of ICE is that it does not require a multivariate joint distribution assumption; it allows different types of variables to be imputed together. ICE allows using different kinds of weights when regression models allow them.

ICE is easy to use and understand. It has two major steps. The first step is the imputation of a single variable given a set of predictor variables, completed by Univariate Imputation Sampling (UVIS)—code statements in ICE. The second step is "regression switching," an algorithm that cycles through all variables to be imputed, using UVIS, which performs imputation of a single variable on a set of predictor variables by an appropriate regression model based on the predictors. The regression model can be ordinary least-squares (OLS) if the imputed variable is a continuous variable, or a logit model if it is a binary variable. Other models can also be used.

With the regression model, UVIS can create the imputed values for the missing observations by drawing from the posterior predictive distribution or by predictive matching. There are two types of distributions involved: regression coefficients and the residual standard deviation. With the boot option, the multivariate normality assumption can be relaxed, which has the advantage of robustness, since the distribution of the beta coefficient is no longer assumed to be multivariate normal.

STATA Release 10 was used to apply the ICE imputation method to the data set. ICE codes used for this study are included below in the Appendix.

3.8 CENSUS 2000 Median Income Data

In additional to the various imputation methods applied to the HOS data as described above, we also used the external imputation method by replicating the methodology developed by Bruess and Bikah using the CENSUS 2000 median income data. The CENSUS data files and SAS codes developed by Bruess and Bikah were used to impute the missing income categories with CENSUS income categories. Mean and the standard error of the mean of the imputed CENSUS income categories were derived and are reported in Table 3. CENSUS income imputed data set was also used to test the multiple regression model used for the other methods described above. Results of the multiple regression analyses are reported in Tables 12 and 13. Since the CENSUS 2000 income data was used for HOS Survey Cohort of 2006, the median income was adjusted for inflation. Two inflationary rates were used—2%and 3%.

4.0 COMPARISON OF IMPUTATION METHODS

This study reports results from application of various missing data imputation methods. Methods we have covered are imputation of external values (zipcode level CENSUS 2000 income adjusted for inflation) as well as methods using simple to complex statistical missing data imputation procedures. The primary focus of these methods was to impute the missing values representing the HOS survey income categories 1-9.

There were noticeable differences in the ability of the methods to impute all the income category missing values. Three of the nine methods imputed all the missing values. The three methods are: overall mean imputation method, STATA ICE and MCMC multiple imputation methods. The other methods—conditional means, hotdeck (SOLAS and SAS Macro), SOLAS Propensity score method and CENSUS 2000 income value imputations—still had some small percentages of missing cases. The residual missing percentages ranged from a high of 3.92% to a low of 0.04%. SAS Macro hotdeck had the lowest percentage of residual missing while conditional mean method had the highest. See the N sizes reported in Table 3.

Most of the means and their respective standard errors of the imputed income category variable yielded values that were very similar, except for the mean from SOLAS Propensity Score imputation method. The mean values of the income categories ranged from a low of 3.93 (SOLAS Propensity Score Method) to a high of 4.28 (CENSUS 3% adjusted). While these values were slightly different from each other, and in some instances statistically significant, the differences however, did not alter the income group they were representing. Because these

means were of income categories, after rounding, they all pointed to the fourth income category of $20,000 to $29,999. If the imputation involved actual dollar values, then any difference would have indicated real differences in the income of the respondents.

The differences in the mean values discussed above also point to the fact that some methods yielded results that were much similar than others. The SOLAS Propensity score method yielded a mean that was smaller than the mean of the CENSUS 2000 income method. The MCMC, STATA ICE, and conditional mean imputation yielded mean values that were larger than the SOLAS Propensity score but lower than SOLAS hotdeck, SAS MACRO hotdeck, CENSUS 2000 income impute and overall mean imputation. See Table 3.

Like the mean test, the multiple regression model testing also yielded results that displayed some differences. But, in general the coefficients were very similar. The R^2 values for testing the model fit (not shown in the table) were all similar (0.08). The magnitude of the estimated beta parameters for all four independent variables was very similar across all the imputation methods. All the estimated parameters have similar signs and were statistically significant with very small standard errors. See Table 14.

5.0 RECOMMENDATIONS

Because we focused only on imputing missing income categories for a rather large data set, the test results do not point to any single method that may be considered superior to the rest of the methods. It seems the various methods considered here yielded results that were very similar. Therefore selecting a method to impute income categories in HOS data will be influenced by the availability of time and resources as well as the type of analysis the researcher is proposing. A researcher may not have the software program we used and hence adoption of a method requiring that software is then out of the question. Similarly, some methods require the use of software that takes a long time to conduct the imputation. Even if the researcher has the software, but not the time, then the researcher is constrained from using that imputation method (for example the SOLAS hot deck or the Propensity Scoring method).

If a researcher is interested in conducting a simple univariate analysis of the income variable in the HOS data set, the use of listwise deletion method may be recommended. Because of the size of the data set, even when 20% of the income categories are missing, listwise deletion method is likely to yield unbiased results. Even estimation of multiple linear regression model coefficients using the listwise deletion method for independent variables such as income categories, the results are robust. Logistic regression models using listwise deletion can tolerate nonrandom missingness on the dependent variable or an independent variable but not both (Allison 2002). Listwise deletion method is the recommended method if the amount of data deleted is small. As discussed in section 2.4 above, when there are missing values in most of the variables in the analysis applying the listwise deletion method then becomes an inefficient method.

Therefore when the rate of missing values across all the variables in the analysis is large, one or more imputation methods may be considered. If the missing variable is discrete with two or more categories, as in the income category variable, then a recommended imputation method would be ICE. ICE handles such variables more appropriately than any of the methods we have considered in our study. If the analysis involves using the cohort data (baseline and the follow-up) then missing income values in the two time periods may be characterized as monotone. The MCMC is then recommended in that situation.

6.0 CONCLUSION

It is important to address missing data issues in observational data sets like HOS because there are significant variables that can be used as predictors in model testing. If there are missing values in many of these predictors, then a large percentage of the cases will be dropped from analysis when using the default listwise deletion method. Our study lost 23% of the cases when we used this method, which is an inefficient way to use data collected at a high cost and could potentially lead to biased parameter estimates.

Admittedly, using one or more missing data imputation methods requires additional effort and time, but it is worth it to create the opportunity to make more efficient use of collected data. This practice also allows the use of advanced statistical methods (e.g., factor analysis, structural equation modeling), when appropriate, and results from these analyses are less likely to be biased.

7.0 TABLES 3-14

Table 3 First 30 Observations From Datasets With Imputed Values for Missing Income Data Category Values Using Various Imputation Methods

Original Data	Overall Mean	Conditional Mean	SOLAS Hot Deck	SAS MACRO Hot Deck	SOLAS Propensity	STATA ICE	SAS MCMC	CENSUS$ 2% adjusted	CENSUS$ 3% adjusted
.	***4.13***	***3.33***	***3***	***2***	***3***	***3***	***2.36***	***4***	***4***
4	4	4	4	4	4	4	4	4	4
3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4
4	4	4	4	4	4	4	4	4	4
3	3	3	3	3	3	3	3	3	3
7	7	7	7	7	7	7	7	7	7
2	2	2	2	2	2	2	2	2	2
4	4	4	4	4	4	4	4	4	4
6	6	6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7	7	7
2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3
.	***4.13***	***3.68***	***6***	***3***	***3***	***4***	***7.18***	***4***	***4***
3	3	3	3	3	3	3	3	3	3
3	3	3	3	3	3	3	3	3	3
.	***4.13***	**.**	***7***	**3**	**.**	***3***	***5.92***	***4***	***4***
.	***4.13***	***4.29***	***7***	***2***	***3***	***5***	***3.98***	***4***	***5***
3	3	3	3	3	3	3	3	3	3
.	***4.13***	***4.16***	**6**	***5***	***3***	***6***	***2.66***	***7***	***7***
5	5	5	5	5	5	5	5	5	5
3	3	3	3	3	3	3	3	3	3
.	***4.13***	***3.42***	***3***	***7***	***3***	***4***	***3.39***	***4***	***4***
4	4	4	4	4	4	4	4	4	4
6	6	6	6	6	6	6	6	6	6
8	8	8	8	8	8	8	8	8	8
.	***4.13***	***4.53***	***4***	***3***	***5***	***3***	***4.7***	**.**	**.**
3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4
8	8	8	8	8	8	8	8	8	8
Mean=4.131 SE=0.0058 N=92,693	Mean=4.131 SE=0.0047 N=116,098	Mean=4.10 SE=0.0048 N=111,540	Mean=4.14 SE=0.0053 N=112,758	Mean=4.13 SE=0.0053 N=116,046	Mean=3.93 SE=0.0052 N=115,074	Mean=4.11 SE=0.0052 N=116,098	Mean=4.10 SE=0.0059 N=116,098	Mean=4.24 SE=0.0049 N=115,653	Mean=4.28 SE=0.00496 N=115,653

Table 4 Multiple Regression Model Testing Results Using Listwise Deletion Method in SAS With General Health Status as the Dependent Variable (B9CMPHTH)

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	T Value	PR >[t]
Intercept	Intercept	1	5.23190	0.03391	154.30	<.0001
B9AGE	B9Age (Exact Calculation)	1	-0.01929	0.00039702	-48.59	<.0001
B9SRVGEND	B9 Q55 Survey-Gender	1	-0.04830	0.00706	-6.84	<.0001
B9EDUC	B9 Q59 Education Level	1	-0.11645	0.00290	-40.17	<.0001
B9HHINC	B9 Q63 Household Income	1	-0.12700	0.00216	-58.83	<.0001

Table 5 Multiple Regression Model Testing Results Using Overall Mean Imputation Method in SAS With General Health Status as the Dependent Variable (B9CMPHTH)

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	T Value	PR >[t]
Intercept	Intercept	1	5.08919	0.03075	165.49	<.0001
B9AGE	B9Age (Exact Calculation)	1	-0.01738	0.00035759	-48.60	<.0001
B9SRVGEND	B9 Q55 Survey-Gender	1	-0.03968	0.00638	-6.22	<.0001
B9EDUC	B9 Q59 Education Level	1	-0.13156	0.00257	-51.14	<.0001
B9HHINC	B9 Q63 Household Income	1	-0.12106	-0.00212	-57.16	<.0001

Table 6 Multiple Regression Model Testing Results Using Conditional Means Imputation Method in SAS With General Health Status as the Dependent Variable (B9CMPHTH)

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	T Value	PR >[t]
Intercept	Intercept	1	5.13171	0.03073	166.99	<.0001
B9AGE	B9Age (Exact Calculation)	1	-0.01776	0.00035698	49.74	<.0001
B9SRVGEND	B9 Q55 Survey-Gender	1	-0.05943	0.00641	-9.27	<.0001
B9EDUC	B9 Q59 Education Level	1	-0.11281	0.00267	-42.30	<.0001
B9HHINC	B9 Q63 Household Income	1	-0.13242	0.00216	-61.38	<.0001

Table 7 Multiple Regression Model Testing Results from SAS Using Imputed Complete-Data With Hot Deck Method in SOLAS, With General Health Status as the Dependent Variable (B9CMPHTH)

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	T Value	PR >[t]
Intercept	Intercept	1	5.04459	0.03062	164.74	<.0001
B9AGE	B9Age (Exact Calculation)	1	-0.01767	0.00035798	-49.35	<.0001
B9SRVGEND	B9 Q55 Survey-Gender	1	-0.04338	0.00640	-6.78	<.0001
B9EDUC	B9 Q59 Education Level	1	-0.12552	0.00263	-47.79	<.0001
B9HHINC	B9 Q63 Household Income	1	-0.10890	0.00195	-55.80	<.0001

Table 8 Multiple Regression Model Testing Results from SAS Using Imputed Complete-Data With Hot Deck Method in SAS, With General Health Status as the Dependent Variable (B9CMPHTH)

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	T Value	PR >[t]
Intercept	Intercept	1	4.99533	0.03052	163.65	<.0001
B9AGE	B9Age (Exact Calculation)	1	-0.01722	0.00035826	-48.07	<.0001
B9SRVGEND	B9 Q55 Survey-Gender	1	-0.02977	0.00637	-4.67	<.0001
B9EDUC	B9 Q59 Education Level	1	-0.13725	0.00256	-53.53	<.0001
B9HHINC	B9 Q63 Household Income	1	-0.10040	0.00188	-53.33	<.0001

Table 9 Multiple Regression Model Testing Results from SAS Using Imputed Complete-Data With Propensity Method in SOLAS, With General Health Status as the Dependent Variable (B9CMPHTH)

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	T Value	PR >[t]
Intercept	Intercept	1	5.04881	0.03067	163.53	<.0001
B9AGE	B9Age (Exact Calculation)	1	-0.01801	0.00035915	-50.14	<.0001
B9SRVGEND	B9 Q55 Survey-Gender	1	-0.04081	0.00642	-6.35	<.0001
B9EDUC	B9 Q59 Education Level	1	-0.13837	0.00257	-53.79	<.0001
B9HHINC	B9 Q63 Household Income	1	-0.09921	0.00194	-51.48	<.0001

Table 10 Multiple Regression Model Testing Results Using MI MCMC Method in SAS With General Health Status as the Dependent Variable (B9CMPHTH)

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	T Value	PR >[t]
Intercept	Intercept	1	5.10286	0.03141	162.44	<.0001
B9AGE	B9Age (Exact Calculation)	1	-0.01757	0.000361	-48.68	<.0001
B9SRVGEND	B9 Q55 Survey-Gender	1	-0.05679	0.00645	-8.81	<.0001
B9EDUC	B9 Q59 Education Level	1	-0.11521	0.00266	-43.32	<.0001
B9HHINC	B9 Q63 Household Income	1	-0.12755	0.00216	-59.00	<.0001

Table 11 Multiple Regression Model Testing Results From STATA Using Imputed Chain Equation (ICE) Method in STATA, With General Health Status as the Dependent Variable (B9CMPHTH)

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	T Value	PR >[t]
Intercept	Intercept	1	5.11067	0.03097	165.01	<.0001
B9AGE	B9Age (Exact Calculation)	1	-0.01767	0.000355	-49.57	<.0001
B9SRVGEND	B9 Q55 Survey-Gender	1	-0.04338	0.00635	-9.08	<.0001
B9EDUC	B9 Q59 Education Level	1	-0.12552	0.00267	-42.95	<.0001
B9HHINC	B9 Q63 Household Income	1	-0.10890	0.00230	-56.00	<.0001

Table 12 Multiple Regression Model Testing Results From Imputation of CENSUS Income Categories (2% inflation adjusted median income) With General Health Status as the Dependent Variable (B9CMPHTH)

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	T Value	PR >[t]
Intercept	Intercept	1	5.04688	0.03056	165.16	<.0001
B9AGE	B9Age (Exact Calculation)	1	-0.01699	0.000358	-49.49	<.0001
B9SRVGEND	B9 Q55 Survey-Gender	1	-0.03407	0.00637	-5.35	<.0001
B9EDUC	B9 Q59 Education Level	1	-0.12821	0.00258	-49.63	<.0001
B9HHINC	B9 Q63 2% Inflation adjusted CENSUS Household Median Income	1	-0.11926	0.00203	-58.87	<.0001

Table 13 Multiple Regression Model Testing Results From Imputation of CENSUS Income Categories (3% inflation adjusted median income) With General Health Status as the Dependent Variable (B9CMPHTH)

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	T Value	PR >[t]
Intercept	Intercept	1	5.03988	0.03053	165.06	<.0001
B9AGE	B9Age (Exact Calculation)	1	-0.01693	0.000358	-47.33	<.0001
B9SRVGEND	B9 Q55 Survey-Gender	1	-0.03199	0.00636	-5.03	<.0001
B9EDUC	B9 Q59 Education Level	1	-0.12909	0.00258	-50.07	<.0001
B9HHINC	B9 Q63 3% Inflation adjusted CENSUS Household Median Income	1	-0.11770	0.00200	-58.79	<.0001

Table 14 Summary Results for Regression Parameters for the Nine Missing Data Handling Methods in Tables 4–13

Methods	Intercept (SE)	B9AGE (SE)	B9SRVGEND (SE)	B9EDUC (SE)	B9HHINC (SE)
Listwise Deletion	5.23190 (0.03391)	-0.01929 (0.00039702)	-0.04830 (0.00706)	-0.11645 (0.00290)	-0.12700 (0.00216)
Overall Mean	5.08919 (0.03075)	-0.01738 (0.00035759)	-0.03968 (0.00638)	-0.13156 (0.00257)	-0.12106 (0.00212)
Conditional Mean	5.13171 (0.03073)	-0.01776 (0.00035698)	-0.05943 (0.00641)	-0.11281 (0.00267)	-0.13242 (0.00216)
Hot Deck (SOLAS)	5.04459 (0.03062)	-0.01767 (0.00035798)	-0.04338 (0.00640)	-0.12552 (0.00263)	-0.10890 (0.00195)
Hot Deck (SAS)	4.99533 (0.03052)	-0.01722 (0.00035826)	-0.02977 (0.00637)	-0.13725 (0.00256)	-0.10040 (0.00188)
Propensity Scores (SOLAS)	5.04881 (0.03067)	-0.01801 (0.00035915)	-0.04081 (0.00642)	-0.13837 (0.00257)	-0.09921 (0.00194)
MCMC (SAS)	5.10286 (0.03141)	-0.01757 (0.000361)	-0.05679 (0.00645)	-0.11521 (0.00266)	-0.12755 (0.00216)
ICE (STATA)	5.11067 (0.03097)	-0.01767 (0.000355)	-0.04338 (0.00635)	-0.12552 (0.00267)	-0.10890 (0.00230)
CENSUS (2% inflation adjusted)	5.04688 (0.03056)	-0.01699 (0.000358)	-0.03407 (0.00637)	-0.12821 (0.00258)	-0.11926 (0.00203)
CENSUS (3% inflation adjusted)	5.03988 (0.03053)	-0.01693 (0.000358)	-0.03199 (0.00636)	-0.12909 (0.00258)	-0.11770 (0.00200)

8.0 REFERENCES

Allison, P.D. 2002. *Missing Data.* Sage University Papers Series on Quantitative Applications in the Social Sciences. Thousand Oaks, CA: Sage.

Altmayer, L. *Hot-Deck Imputation: A Simple DATA Step Approach*. Paper presented at the 15th Annual North East SAS Users Group (NESUG) Conference held on September 29–October 2, 2002, at Adams Mark, Buffalo, New York. http://analytics.ncsu.edu/sesug/1999/075.pdf

Bruess, J., and M. Bikah. 2008. *Imputation Analysis for HOS Income Data: Phase One: External Imputation.* Technical Report. National Committee for Quality Assurance (NCQA), Washington, D.C. (October 2008)

Carpenter, J.R., M.G. Kenward, and S. Vansteellandt. 2006. A Comparison of Multiple Imputation and Doubly Robust Estimation for Analyses with Missing Data. *Journal of the Royal Statistical Society* Series A, 19: 571–84.

Collins, L. M., J.L. Schafer, and C.M. Kam. 2001. A Comparison of Inclusive and Restrictive Strategies in Modern Missing Data Procedures. *Psychological Methods* 6: 330–51.

Diggle, P., and M.G. Kenward. 1994. Informative Drop-Out in Longitudinal Data Analysis. *Applied Statistics* 43: 49–73.

Diggle, P.J., P. Heagerty., K.Y. Liang, and S.L. Zegar. 2002. *Analysis of Longitudinal Data* (second edition). Clarendon, Texas: Clarendon Press.

Horton, N.J., and K.P. Kleinman. 2007. Much Ado About Nothing: A Comparison of Missing Data Methods and Software to fit Incomplete Data Regression Models. *The American Statistician* (February) Vol. 61, No. 1: 79–90. (Includes appendix pp 1–19 with analysis results and software codes.)

Kazis, L.E., et al. Health Status of Veterans: Physical and Mental Component Summary Scores (SF-12V). *1997 National Survey of Ambulatory Care Patients, Executive Report.* Office of Performance and Quality, Health Assessment Project HSR&D Field Program, VHA National Customer Feedback Center, Washington, D.C., Bedford and West Roxbury, Massachusetts (April 1998).

Kazis, L.E., et al. Health Status and Outcomes of Veterans: Physical and Mental Component Summary Scores (Veterans SF-12). *1998 National Survey of Hospitalized Patients, Executive Report.* Office of Performance and Quality, Health Assessment Project, HSR&D Field Program, Washington, D.C., and Bedford, Massachusetts (April 1999).

Little, R.J.A. 1994. A Class of Pattern-Mixture Models for Normal Incomplete Data. *Biometrika* 81: 471–83.

Little, R.J.A., and D.B. Rubin. 1987. *Statistical Analysis With Missing Data.* New York: Wiley.

Ibid. 2002. *Statistical Analysis With Missing Data* (second edition). New York: Wiley.

Royston, P. 2005. Multiple Imputation of Missing Values. *State Technical Journal* 5: 327–536.

Schafer, J.L. 1997. *Analysis of Incomplete Multivariate Data.* New York: Chapman and Hall.

9.0 APPENDIX

9.1 SAS Codes for Hot Deck

```
*************************************************************************;
**  Project:   HOS Income Hot Deck Imputation Using SAS Macro developed by *;
**             Lawrence Altmayer, U.S. Census Bureau                      **;
**  Purpose:   Impute income from HOS Cohort 9 baseline data set using    **;
**              hot-deck methodology                                      **;
**  Date:      01-08-09                                                   **;
*************************************************************************;
**  note:      Draft version                                              **;
*************************************************************************;

libname hos10 '\\profs5\Secure\HOS\DATA\Round 10 (B10_FU8) final 4-24-08';
libname hos   '\\profs5\Secure\HOS\DATA\Cohort 9 baseline 1-22-07\';
libname hos09 'W:\HOS\Income Imputation -
2008\Internal_income_imputation_120808\data';
options nonumber;

     proc contents data=hos09.b9data;
     run;
       data ordered;
            set hos09.b9data;
                  order = _n_;
     run;
/*   proc freq data=hos.b9data1;
               tables b9srvgend b9race b9age b9gender b9state b9stateabv
B9HHINC B9EDUC;
     run; */
data manage(rename=(b9contract=CTRLNUM b9race=RACE b9age=AGE
                    b9zip=ZIP b9city=CITY b9state=STATE));
     attrib SEX length=$1 INCOME length=$2;
     set hos09.b9data(keep=hicnum b9contract b9srvgend b9race b9age b9gender
b9hhinc
                          b9educ b9zip b9zipcde b9city b9state b9stateabv
                          b9cmphth b9agecat b9marital b9hmown);
     SEX = b9srvgend;
     INCOME = b9hhinc;
run;
proc sort data=work.manage
          out =b9data;
     by CTRLNUM STATE CITY ZIP ;
run;

     /* Stiller/Dalzell method - not used
      * Driver Program for hot-deck imputation *

     data work.b9imputed     * imputed dataset *;
          work.mcounts       * matrix counts   *;

            array educ{6,6,6,6,6,6,1,1,9} _temporary_
```

```
                  (666666000   * for cell (1,1)  *;
                  (666666000   * for cell (1,1)  *;
                  (666666000   * for cell (1,1)  *;
                  (666666000   * for cell (1,1)  *;
                  (666666000   * for cell (1,1)  *;
                    );

     run; */

/* Version by Lawrence Altmayer, U.S. Census Bureau  */

/* AUTHOR
   Lawrence Altmayer
   U.S. Bureau of the Census
   ESMPD, Room 1223-4
   Washington, DC 20233-0001
   (301)457-2581
   lawrence.w.altmayer@ccmail.census.gov  */

/* This program imputes the values for age, race, and
   sex variables on the line level sap1 dataset.
   If necessary to run this program, delete sap1imp data-
   set beforehand, since it is created using proc append.  */

options ls=78 compress=yes;
/* options symbolgen; */
/* libname d '';      */

data sap1e;
     set work.b9data;
run;

data sap1ctl (keep=ctrlnum numlines);  /* <1> Produce a dataset containing
the
                                          number of persons per CTRLNUM
*/
     set sap1e;
     by ctrlnum;
     if first.ctrlnum then numlines=0;
        numlines+1;
     if last.ctrlnum then output;
run;

data _null_;  /* <2> Create macro variable to store the
                    values of CTRLNUM and NUMLINES     */
     set sap1ctl end=eof;
     call symput('ctrl'||left(_n_),trim(ctrlnum));
     call symput('line'||left(_n_),left(trim(numlines)));
     if eof then call symput('n',left(_n_));
run;

%macro subseta;

%do i=1 %to &n;

/* subset sap1e by ctrlnum, and first resolve race based */
```

```
/* on other data - create flag array for imputing race   */

data sap1e&i;  /* <3> The 1st step of the imputation process: attempt to
                      resolve RACE based on other data from the same
observation.
                      Subset sap1e by CTRLNUM to get sap1e&i.            */
     set sap1e;
     if ctrlnum="&&ctrl&i";
run;

/*   *--The following section is not useful unless you know/have the values
to assign--*
     data sap1r&i (keep=imprace) sap1a&i (keep=ctrlnum racfl1-racfl&&line&i);
     set sap1e&i;
         by ctrlnum;
            if ctrlnum="&&ctrl&i";
     retain racfl1-racfl&&line&i;

     array arrfl{&&line&i} $ racfl1-racfl&&line&i;
           if (b4='5' or b4='8' or b4='') and b5='1' then do;
              imprace='1'; arrfl{_n_}='1'; end;

           if ctrlnum='1001' and (b4='8' or b4='') then do;
              imprace='1'; arrfl{_n_}='1'; end;

     if last.ctrlnum then output sap1a&i;
     output sap1r&i;
run;

data sap1e&i;
     merge sap1e&i sap1r&i;
run;

data sap1e&i;
     merge sap1e&i sap1a&i;
     by ctrlnum;
run; --*/

/* create arrays necessary for imputing age, */
/* sex and race using hot deck method        */

data sap1b&i (keep=ctrlnum income1-income&&line&i rac1-rac&&line&i);
     /* <4> hot-deck portion of imputation process: use corresponding data
from
            previous or subsequent observations in a group. Create CTRLNUM-
level
            arrays for RACE AGE GENDER to be added to each observation in the
group */
     set sap1e&i;
         by ctrlnum;
     retain income1-income&&line&i rac1-rac&&line&i;

     array arinc{&&line&i} $ income1-income&&line&i;
     array arrac{&&line&i} $ rac1-rac&&line&i;
     array arrfl{&&line&i} $ racfl1-racfl&&line&i;
```

```
          arinc{_n_}=INCOME; * =b2;
          arrac{_n_}=RACE;   * =b4;
     if last.ctrlnum then output;
run;

data sap1e&i; /* <5> Merge sap1b&i with sap1e&i from above to get a new
                    version of sap1e&i with 1 observation for each person */
     merge sap1e&i sap1b&i;
     by ctrlnum;
run;

/* hot deck imputation for age, sex and race */
data sap1f&i (drop=i rac1-rac&&line&i racfl1-racfl&&line&i
                    inc1-inc&&line&i income1-income&&line&i incfl1-
incfl&&line&i
                    trcfl1-trcfl&&line&i);
     attrib impinc length=$2;
     set sap1e&i;
     retain incfl1-incfl&&line&i trcfl1-trcfl&&line&i;
     /* <6> create flags to indicate whether an
              observation has been used for imputation */

     array arinc{&&line&i} $ income1-income&&line&i;
     array arrac{&&line&i} $ rac1-rac&&line&i;
     array arifl{&&line&i} $ incfl1-incfl&&line&i;
     array arrfl{&&line&i} $ racfl1-racfl&&line&i;
     array trcfl{&&line&i} $ trcfl1-trcfl&&line&i;

     /* create temporary array for race flags, since */
     /* previously created array cannot be retained  */

     if _n_=1 then do;
        %do j=1 %to &&line&i;
            trcfl&j=racfl&j;
        %end;
     end;

     /* hot deck imputation for INCOME */
      * if b2='' then do;
     if INCOME=' .' then do;
        if _n_ > 1 then do;
          do i=_n_-1 to 1 by -1;
             if arinc{i}^=' .' and arifl{i}^='1' then do;
                arifl{i}='1'; impinc=arinc{i};
                goto nextrace;
             end;
        end;
        if impinc=' .' then do;
           do i=_n_+1 to &&line&i;
              if arinc{i}^=' .' and arifl{i}^='1' then do;
                 arifl{i}='1'; impinc=arinc{i};
                 goto nextrace;
              end;
        end;
     end;
     if impinc=' .' then impinc='9';
```

```
     end;
     /* _n_=1 */
     else do;
          do i=_n_+1 to &&line&i;
             if arinc{i}^=' .' and arifl{i}^='1' then do;
                arifl{i}='1'; impinc=arinc{i};
                goto nextrace;
             end;
     end;
     if impinc=' .' then impinc='9';
     end;
     end;
     nextrace:
     /* hot deck imputation for race */
      * if (b4='' or b4='8') and imprace='' then do;
     if RACE='' then do;
        if _n_>1 then do;
           do i=_n_-1 to 1 by -1;
           if arrac{i}^='' and arrac{i}^='8' and trcfl{i}^='1' then do;
              trcfl{i}='1'; imprace=arrac{i};
              goto next;
           end;
     end;
     if imprace='' then do;
        do i=_n_+1 to &&line&i;
           if arrac{i}^='' and arrac{i}^='8' and trcfl{i}^='1' then do;
              trcfl{i}='1'; imprace=arrac{i};
              goto next;
           end;
     end;
     end;
     if imprace='' then imprace='9';
     end;
     /* _n_=1 */
     else do;
          do i=_n_+1 to &&line&i;
             if arrac{i}^='' and arrac{i}^='8' and trcfl{i}^='1' then do;
                trcfl{i}='1'; imprace=arrac{i};
                goto next;
             end;
     end;
     if imprace='' then imprace='9';
     end;
     end;
     next:
run;

   proc append base=hos09.hotd_imp data=sap1f&i
        force;  /* d.sap1imp data=sap1f&i; */
   run;

%end;
%mend subseta;

%subseta
```

```
/* proc contents data=hos.hotd_imp position;  * uncomment if proc append used
*;
   run; */

data Sap1g1(keep=hicnum b9hhinc);
     set hos09.hotd_imp;
          b9hhinc = INCOME;
       If INCOME = ' .' then b9hhinc = impinc; else
          b9hhinc = INCOME;
run;

proc sort data=ordered; by hicnum; run;
proc sort data=Sap1g1 ; by hicnum; run;

data kazi2;
     merge work.ordered (in=A drop=b9hhinc)
             work.Sap1g1  (in=B);
     by hicnum;
       if A;
run;

proc sort data=work.kazi2 ; by order; run;

/* statistical analysis for Kazi Ahmed */
proc univariate data=kazi2;
     var b9hhinc b9cmphth b9age b9srvgend b9educ b9agecat b9marital b9hmown
b9agecat;
run;

proc reg data=kazi2;
     model b9cmphth=b9age b9srvgend b9educ b9hhinc/vif;
run;
quit;
/*eof*/
```

9.2 SAS CODES

```
*Kazi Ahmed*;
*Data=b9data*;
*Date: January 15, 2009*;
*************************************************************************;
*HOS Cohort 9 Baseline data set B9data.sas7bdat was used to apply the various
*missing value imputation methods. The following codes were used to apply the
*overall mean imputation; conditional mean imputation; and multiple
*imputation using PROC MI and PROC MIANALYZE. The codes also include
*computation of the newgroup variable for conditional mean;
*************************************************************************;
data kazi1;
set 'W:\HOS\Income Imputation - 2008\Internal_income_imputation_120808\data\b9data';
proc contents data=kazi1;
run;
proc freq data =kazi1; tables b9hhinc b9cmphth b9age b9srvgend b9educ;
run;
proc univariate data=kazi1;
 var b9hhinc b9cmphth b9age b9srvgend b9educ b9agecat b9marital b9hmown;
run;

*Test a regression model with the original data prior to imputation;
proc reg data=kazi1;
model b9cmphth=b9age b9srvgend b9educ b9hhinc/vif;
run;

/*impute the overall mean value for the missing values in the income
variable. The original income b9hhinc was not changed but a newly computed
duplicate variable called reinc was. This is useful for imputing missing
value with a single mean value*/

data kazi2;
set kazi1;reinc=b9hhinc;
if reinc= . then reinc=4.130560; /*Comment it after running overall mean
imputation. Run from top to create the newgrp variable and later impute the
conditional means to missing reinc*/

*Compute a new variable called newgrp by combining the response categories of
the three demographic variables-gender, age, and education. Age was collapsed
into 5 categories. The new variable newgrp has 60 categories;

if b9srvgend =1 and b9agecat=0 and b9educ=1 then newgrp=1;
if b9srvgend =1 and b9agecat=0 and b9educ=2 then newgrp=2;
if b9srvgend =1 and b9agecat=0 and b9educ=3 then newgrp=3;
if b9srvgend =1 and b9agecat=0 and b9educ=4 then newgrp=4;
if b9srvgend =1 and b9agecat=0 and b9educ=5 then newgrp=5;
if b9srvgend =1 and b9agecat=0 and b9educ=6 then newgrp=6;
if b9srvgend =1 and b9agecat=1 and b9educ=1 then newgrp=7;
if b9srvgend =1 and b9agecat=1 and b9educ=2 then newgrp=8;
if b9srvgend =1 and b9agecat=1 and b9educ=3 then newgrp=9;
if b9srvgend =1 and b9agecat=1 and b9educ=4 then newgrp=10;
if b9srvgend =1 and b9agecat=1 and b9educ=5 then newgrp=11;
if b9srvgend =1 and b9agecat=1 and b9educ=6 then newgrp=12;
if b9srvgend =1 and b9agecat=2 and b9educ=1 then newgrp=13;
```

```
if b9srvgend =1 and b9agecat=2 and b9educ=2 then newgrp=14;
if b9srvgend =1 and b9agecat=2 and b9educ=3 then newgrp=15;
if b9srvgend =1 and b9agecat=2 and b9educ=4 then newgrp=16;
if b9srvgend =1 and b9agecat=2 and b9educ=5 then newgrp=17;
if b9srvgend =1 and b9agecat=2 and b9educ=6 then newgrp=18;
if b9srvgend =1 and b9agecat=3 and b9educ=1 then newgrp=19;
if b9srvgend =1 and b9agecat=3 and b9educ=2 then newgrp=20;
if b9srvgend =1 and b9agecat=3 and b9educ=3 then newgrp=21;
if b9srvgend =1 and b9agecat=3 and b9educ=4 then newgrp=22;
if b9srvgend =1 and b9agecat=3 and b9educ=5 then newgrp=23;
if b9srvgend =1 and b9agecat=3 and b9educ=6 then newgrp=24;
if b9srvgend =1 and b9agecat=4 and b9educ=1 then newgrp=25;
if b9srvgend =1 and b9agecat=4 and b9educ=2 then newgrp=26;
if b9srvgend =1 and b9agecat=4 and b9educ=3 then newgrp=27;
if b9srvgend =1 and b9agecat=4 and b9educ=4 then newgrp=28;
if b9srvgend =1 and b9agecat=4 and b9educ=5 then newgrp=29;
if b9srvgend =1 and b9agecat=4 and b9educ=6 then newgrp=30;
if b9srvgend =2 and b9agecat=0 and b9educ=1 then newgrp=31;
if b9srvgend =2 and b9agecat=0 and b9educ=2 then newgrp=32;
if b9srvgend =2 and b9agecat=0 and b9educ=3 then newgrp=33;
if b9srvgend =2 and b9agecat=0 and b9educ=4 then newgrp=34;
if b9srvgend =2 and b9agecat=0 and b9educ=5 then newgrp=35;
if b9srvgend =2 and b9agecat=0 and b9educ=6 then newgrp=36;
if b9srvgend =2 and b9agecat=1 and b9educ=1 then newgrp=37;
if b9srvgend =2 and b9agecat=1 and b9educ=2 then newgrp=38;
if b9srvgend =2 and b9agecat=1 and b9educ=3 then newgrp=39;
if b9srvgend =2 and b9agecat=1 and b9educ=4 then newgrp=40;
if b9srvgend =2 and b9agecat=1 and b9educ=5 then newgrp=41;
if b9srvgend =2 and b9agecat=1 and b9educ=6 then newgrp=42;
if b9srvgend =2 and b9agecat=2 and b9educ=1 then newgrp=43;
if b9srvgend =2 and b9agecat=2 and b9educ=2 then newgrp=44;
if b9srvgend =2 and b9agecat=2 and b9educ=3 then newgrp=45;
if b9srvgend =2 and b9agecat=2 and b9educ=4 then newgrp=46;
if b9srvgend =2 and b9agecat=2 and b9educ=5 then newgrp=47;
if b9srvgend =2 and b9agecat=2 and b9educ=6 then newgrp=48;
if b9srvgend =2 and b9agecat=3 and b9educ=1 then newgrp=49;
if b9srvgend =2 and b9agecat=3 and b9educ=2 then newgrp=50;
if b9srvgend =2 and b9agecat=3 and b9educ=3 then newgrp=51;
if b9srvgend =2 and b9agecat=3 and b9educ=4 then newgrp=52;
if b9srvgend =2 and b9agecat=3 and b9educ=5 then newgrp=53;
if b9srvgend =2 and b9agecat=3 and b9educ=6 then newgrp=54;
if b9srvgend =2 and b9agecat=4 and b9educ=1 then newgrp=55;
if b9srvgend =2 and b9agecat=4 and b9educ=2 then newgrp=56;
if b9srvgend =2 and b9agecat=4 and b9educ=3 then newgrp=57;
if b9srvgend =2 and b9agecat=4 and b9educ=4 then newgrp=58;
if b9srvgend =2 and b9agecat=4 and b9educ=5 then newgrp=59;
if b9srvgend =2 and b9agecat=4 and b9educ=6 then newgrp=60;

proc freq data =kazi2; tables newgrp;/*Run frequency to check whether newgrp
was properly computed and its distribution*/
run;
proc means data=kazi2; /*Compute the 60 means for each of the 60 categories
in the newgrp variable.*/
 class newgrp;
 var b9hhinc;
 output out=cmean;
```

```
 run;

*Impute the conditional means by recoding the variable reinc with the mean
values for the respective newgrp categories. The variable reinc will be used
to test the regression model;

data kazi_9;
set kazi2;
if newgrp=1 and reinc=. then reinc=3.3888889;
if newgrp=2 and reinc=. then reinc=3.9090909;
if newgrp=3 and reinc=. then reinc=3.3437500;
if newgrp=4 and reinc=. then reinc=4.2857143;
if newgrp=5 and reinc=. then reinc=3.0000000;
if newgrp=6 and reinc=. then reinc=4.5000000;
if newgrp=7 and reinc=. then reinc=4.6875000;
if newgrp=8 and reinc=. then reinc=3.7777778;
if newgrp=9 and reinc=. then reinc=5.0000000;
if newgrp=10 and reinc=. then reinc=4.2608696;
if newgrp=11 and reinc=. then reinc=6.3333333;
if newgrp=12 and reinc=. then reinc=5.2500000;
if newgrp=13 and reinc=. then reinc=3.9047619;
if newgrp=14 and reinc=. then reinc=4.8461538;
if newgrp=15 and reinc=. then reinc=5.5714286;
if newgrp=16 and reinc=. then reinc=5.7692308;
if newgrp=17 and reinc=. then reinc=5.5714286;
if newgrp=18 and reinc=. then reinc=5.5000000;
if newgrp=19 and reinc=. then reinc=5.6666667;
if newgrp=20 and reinc=. then reinc=2.8571429;
if newgrp=21 and reinc=. then reinc=4.6315789;
if newgrp=22 and reinc=. then reinc=5.8888889;
if newgrp=23 and reinc=. then reinc=7.6666667;
if newgrp=24 and reinc=. then reinc=7.0000000;
if newgrp=25 and reinc=. then reinc=4.4666667;
if newgrp=26 and reinc=. then reinc=4.3636364;
if newgrp=27 and reinc=. then reinc=4.9333333;
if newgrp=28 and reinc=. then reinc=8.3333333;
if newgrp=29 and reinc=. then reinc=8.7500000;
if newgrp=30 and reinc=. then reinc=6.3333333;
if newgrp=31 and reinc=. then reinc=4.2500000;
if newgrp=32 and reinc=. then reinc=4.2000000;
if newgrp=33 and reinc=. then reinc=3.9761905;
if newgrp=34 and reinc=. then reinc=3.3571429;
if newgrp=35 and reinc=. then reinc=2.3333333;
if newgrp=36 and reinc=. then reinc=5.2500000;
if newgrp=37 and reinc=. then reinc=4.3125000;
if newgrp=38 and reinc=. then reinc=3.1250000;
if newgrp=39 and reinc=. then reinc=3.7857143;
if newgrp=40 and reinc=. then reinc=5.1000000;
if newgrp=41 and reinc=. then reinc=3.0000000;
if newgrp=42 and reinc=. then reinc=8.5000000;
if newgrp=43 and reinc=. then reinc=3.2222222;
if newgrp=44 and reinc=. then reinc=4.1621622;
if newgrp=45 and reinc=. then reinc=4.6078431;
if newgrp=46 and reinc=. then reinc=4.9677419;
if newgrp=47 and reinc=. then reinc=6.0000000;
if newgrp=48 and reinc=. then reinc=5.0000000;
```

```
if newgrp=49 and reinc=. then reinc=3.8260870;
if newgrp=50 and reinc=. then reinc=3.9166667;
if newgrp=51 and reinc=. then reinc=4.2400000;
if newgrp=52 and reinc=. then reinc=5.0000000;
if newgrp=53 and reinc=. then reinc=4.7142857;
if newgrp=54 and reinc=. then reinc=4.0000000;
if newgrp=55 and reinc=. then reinc=4.3200000;
if newgrp=56 and reinc=. then reinc=4.5769231;
if newgrp=57 and reinc=. then reinc=4.4615385;
if newgrp=58 and reinc=. then reinc=4.1111111;
if newgrp=59 and reinc=. then reinc=10.0000000;
if newgrp=60 and reinc=. then reinc=7.0000000;
proc freq data=kazi_9; tables reinc;
run;

*Test the regression model with the imputed income variable;
proc reg data=kazi_9;
model b9cmphth=b9age b9srvgend b9educ reinc/vif;
run;

*Multiple Imputation using Markov chain Monte Carlo (MCMC) method in SAS;
proc mi data= 'c:\missing data imputation\b9dtkz' out='c:\missing data
imputation\mioutmc';
 mcmc chain=multiple displayinit initial=em(itprint);
 var b9cmphth b9age b9hhinc b9educ b9srvgend;
 run;
 proc reg data='c:\missing data imputation\mioutmc' outest=outreg covout
noprint;
 model b9cmphth=b9age b9srvgend b9educ b9hhinc;
 by _imputation_;
 run;
 proc print data=outreg(obs=10);
 var _imputation_ _type_ _name_
 intercept b9age b9srvgend b9educ b9hhinc;
 title 'Parameter Estimates from Imputed Data Sets-KZ';
 run;
*Combining regression parameters from multiple imputed data sets;
 proc mianalyze data=outreg mult edf=40;
 var intercept b9age b9srvgend b9educ b9hhinc;
run;
```

9.3 STATA ICE Codes

```
. findit mvpatterns
. findit ice
. findit mim
. which ice
. ssc describe ice
. ssc install ice, replace
. set memory 650m
. compress
. recode b9srvgen (1=0) (2=1), generate (gender)
. summarize gender
. ice b9hhinc b9cmphth gender b9educ b9age b9marita, m(5) cmd(b9hhinc:ologit) seed(123456)
clear
. mim: reg b9cmphth b9age b9srvgen b9educ b9hhinc
. mim: mlogit b9cmphth b9age b9srvgen b9educ b9hhinc
```