Multiple Imputation of Family Income and Personal Earnings in the National Health Interview Survey: Methods and Examples

Nathaniel Schenker[a], Trivellore E. Raghunathan[b], Pei-Lu Chiu[a], Diane M. Makuc[a], Guangyu Zhang[b], and Alan J. Cohen[a]

[a] National Center for Health Statistics, [b] University of Michigan

AUGUST 31, 2010

Contents

Abstract iii

1 Introduction 1

1.1 Questions on Family Income and Personal Earnings in the NHIS 1

1.2 Missing Data on Family Income and Personal Earnings 5

1.3 Multiple Imputation of Income and Earnings Items 5

1.4 Objective and Contents of this Report 6

2 Multiple Imputation 7

2.1 Overview of Multiple Imputation 7

2.2 Analyzing Multiply Imputed Data 10
2.2.1 General Procedures 10
2.2.2 Combining Data Across Years of the NHIS 11

2.3 Analyzing Only a Single Completed Data Set 12

3 Procedure for Creating Imputations for the NHIS 13

3.1 Overview of the Imputation Procedure 14
3.1.1 Steps in the Imputation Procedure 14
3.1.2 Sequential Regression Multivariate Imputation 15
3.1.3 Reflecting the Sample Design in Creating the Imputations 18

3.2 Further Details of the Imputation Procedure 18
3.2.1 Step 1: Imputing Person-Level Covariates for Adults 18
3.2.2 Step 2: Creating Family-Level Covariates 19
3.2.3 Step 3: Imputing Family Income and Family Earnings (and Family-Level Covariates) 19
3.2.4 Step 4: Imputing Personal Earnings 20

3.3 Inconsistencies Between Family Income and Family Earnings 21

4 Software for Analyzing Multiply Imputed Data 23

4.1 Analysis Using SAS-Callable SUDAAN 26
4.1.1 SUDAAN Version 9.0 with a Built-In Option for Multiple Imputation 26
4.1.2 SAS Commands for Use with SUDAAN Version 7 or Higher without a Built-In Option for Multiple Imputation 28

4.2 Analysis Using SAS-Callable IVEware 28

Appendix A Technical Details for Analyzing Multiply Imputed Data 30

Appendix B Variables Included in the Imputation Process 34

Appendix C SAS Code for the Examples in Section 4 46

C.1 Code for Creating Completed Data Sets 46

C.2 Code for Use with SAS-Callable SUDAAN Version 9.0 with a Built-In Option for Multiple Imputation 50

C.3 Code for Use with SAS-Callable SUDAAN Version 7 or Higher without a Built-In Option for Multiple Imputation 51

C.4 Code for Use with SAS-Callable IVEware 54

Appendix D Sample Output from SAS-Callable SUDAAN Version 9.0 with a Built-In Option for Multiple Imputation 55

Appendix E Sample Output from SAS Commands for Use with SUDAAN Version 7 or Higher without a Built-In Option for Multiple Imputation 67

Appendix F Sample Output from SAS-Callable IVEware 74

References 80

Abstract

The National Health Interview Survey (NHIS) provides a rich source of data for studying relationships between income and health and for monitoring health and health care for persons at different income levels. However, the nonresponse rates are high for two key items, total family income in the previous calendar year and personal earnings from employment in the previous calendar year. To handle the problem of missing data on family income and personal earnings in the NHIS, multiple imputation of these items was performed for the survey years 1997 – 2009. (There are plans to create multiple imputations for the years 2010 and beyond as well, as the data become available.) For each survey year, data sets containing the imputed values, along with related documentation, can be obtained from the NHIS Web site (http://www.cdc.gov/nchs/nhis.htm). The objective of this report is to describe the approach used to create the multiple imputations and methods for analyzing the multiply imputed data.

1. Introduction

The National Health Interview Survey (NHIS) is a multi-purpose health survey and is the principal source of information on the health of the civilian, noninstitutionalized household population of the United States (National Center for Health Statistics 2010). The NHIS provides a rich source of data for studying relationships between income and health and for monitoring health and health care for persons at different income levels. There is particular interest in the health of vulnerable populations such as those with low income, as well as their access to health care and their use of health care. However, the nonresponse rates are high for two key items, total family income in the previous calendar year and personal earnings from employment in the previous calendar year.

1.1 Questions on Family Income and Personal Earnings in the NHIS

In 1997, the NHIS questionnaire underwent a major revision. The redesigned questionnaire consists of a Basic Module or Core as well as variable Supplements. The Basic Module, which remains largely unchanged from year to year, consists of three components: the Family Core; the Sample Adult Core; and the Sample Child Core. Data are collected through personal household interviews.

For the Family Core component, all members of the household 17 years of age and over who are at home at the time of the interview are invited to participate and to respond for themselves. For those under 17 years of age and those not at home during the interview, information is provided by a knowledgeable adult (18 years of age or over) family member residing in the household. The Family Core component collects information on everyone in the family and includes sections on family relationships, health status and limitations of activities, injuries, health care access and utilization, health insurance, socio-demographic background, and

income and assets. The questions on personal earnings and total family income are in different sections (socio-demographic background and income and assets, respectively).

The socio-demographic background section includes a question on total earnings in the last calendar year for each adult (18 years of age or over) who had at least one job or business: "What is your best estimate of {your/subject name's} earnings {including hourly wages, salaries, tips and commissions} before taxes and deductions from all jobs and businesses in {last calendar year}?"[1] The response to this question is not taken into account or used in the next section (income and assets).

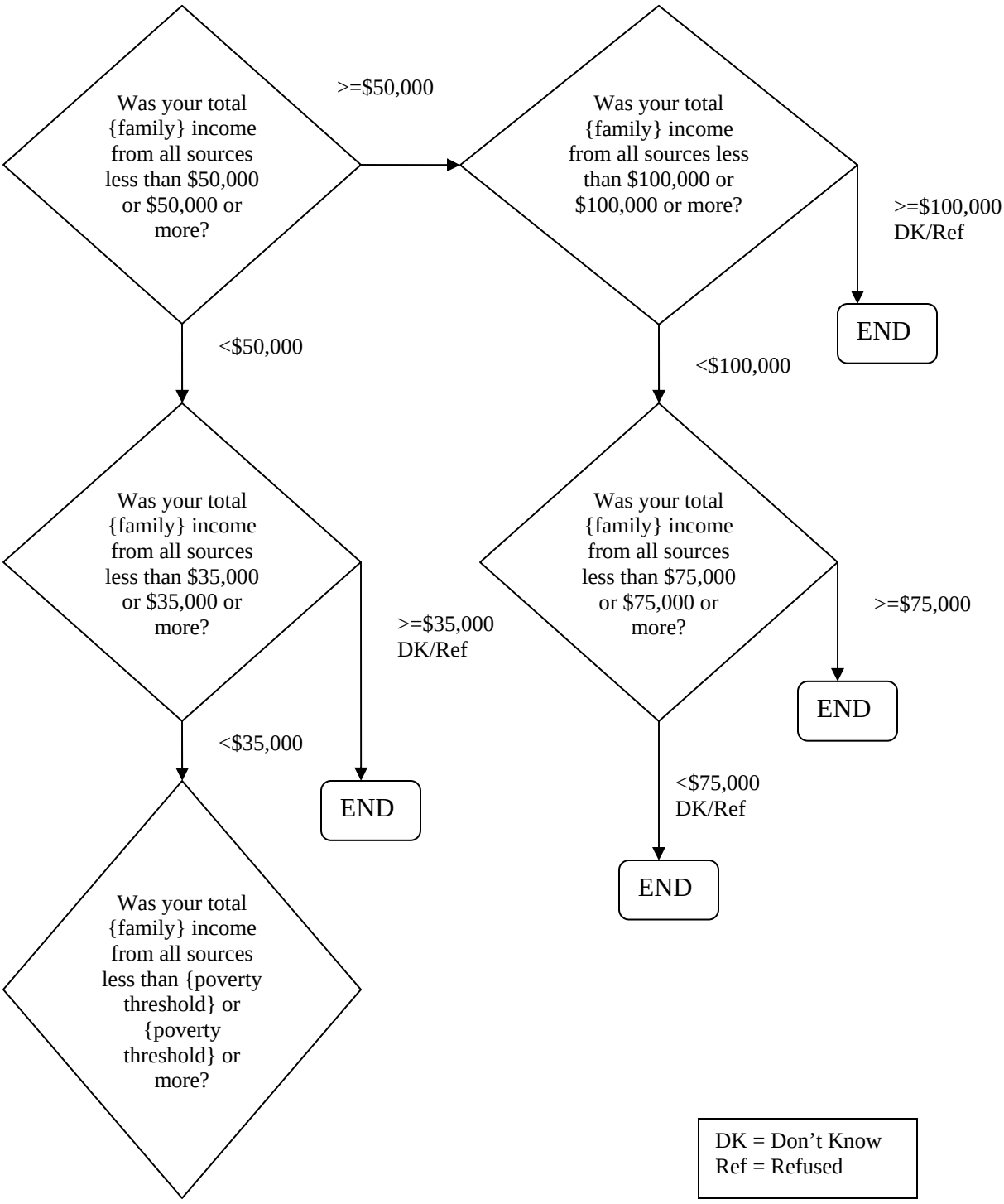
In the section on income and assets, the respondent is first asked whether any family members of any age (and if so, who) received income from each of several different sources.[2] In the 1997 to 2006 NHIS, the respondent is then asked about total combined family income for all family members including children as follows: "Now I am going to ask about the total combined income of your family in {last calendar year}, including income from all sources we have just talked about such as wages, salaries, Social Security or retirement benefits, help from relatives and so forth. Can you tell me that amount before taxes?" If the respondent does not know the amount, the following question is asked: "You may not be able to give us an exact figure for your total combined family income, but can you tell me if your income was $20,000 or more or less than $20,000?" If one of these two income groups is specified by the respondent, a card is shown to the respondent with the goal of placing the income into one of 44 detailed income categories, and the respondent is asked which category best represents the total combined family

[1] Earnings include wages, salaries, tips, commissions, Armed Forces pay and cash bonuses, and subsistence allowances, as well as net income from unincorporated businesses, professional practices, farms, or rental property (where "net" means after deducting business expenses, but before deducting personal taxes).

[2] Sources of income about which the respondent is questioned are: wages and salaries; self-employment including business and farm income; Social Security or Railroad Retirement; disability pension; retirement or survivor pension; Supplemental Security Income; cash assistance from a welfare program; other kind of welfare assistance; interest; dividends; net rental income; child support; and other sources.

income. Starting with the 2007 NHIS, the respondent is asked about total combined family income for all family members including children as follows: “What is your best estimate of {your total income/the total income of all family members} from all sources, before taxes, in {last calendar year}?” If the respondent refuses or does not know the amount, the following question is asked: “Was your total {family} income from all sources less than $50,000 or $50,000 or more?” If one of these two income groups is specified by the respondent, follow up questions of income ranges are asked based on the respondent’s answer. Figure 1 presents a diagram of the new sets of income questions that were implemented in the 2007 NHIS for collecting income ranges. Note that the total combined income of all family members is estimated by the respondent. An estimate of family income is not obtained by summing responses to more detailed questions, as is done in some surveys that include more extensive questions on income, such as the Current Population Survey, a monthly survey of households conducted by the Bureau of the Census for the Bureau of Labor Statistics.

Figure 1. Family income questions for nonresponders to exact family income for collecting family income ranges, starting with the 2007 NHIS.



1.2 Missing Data on Family Income and Personal Earnings

For the years 1997 – 2006, the weighted percentages of persons with unknown family income were within the following ranges: 24-34% for the "exact" value; 20-31% for the detailed categorical (44 categories) value; and 6-11% for the two-category ($20,000 or more, or less than $20,000) value. For the years 2007 – 2009, the weighted percentages of persons with unknown family income were within the following ranges: 33% and 25% for the "exact" value; 15% and 9% for any of the family income ranges questions; and 9% and 5% for the two-category ($50,000 or more, or less than $50,000) income range value respectively. For the years 1997 – 2009, the weighted percentages of employed adults with unknown personal earnings were within 24-33%. (The weighted missing-data rates given in this paragraph are all close to their unweighted counterparts.) There is evidence that the nonresponse on family income and personal earnings was related to several person-level and family-level characteristics, including items pertaining to health. Thus, the respondents cannot be treated as a random subset of the original sample. It follows that the most common method for handling missing data in software packages, "complete-case analysis" (Little and Rubin 2002, Section 3.2), also known as "listwise deletion," which deletes cases that are missing any of the variables involved in the analysis, will generally be biased. Moreover, since deletion of incomplete cases discards some of the observed data, complete-case analysis is generally inefficient as well; that is, it produces inferences that are less precise than those produced by methods that use all of the observed data.

1.3 Multiple Imputation of Income and Earnings Items

To handle the problem of missing data on family income and personal earnings in the NHIS, multiple imputation of these items was performed for the survey years 1997 – 2009, with five sets of imputed values created to allow the assessment of variability due to imputation. (There

are plans to create multiple imputations for the years 2010 and beyond as well, as the data become available.) Since personal earnings were only collected for employed adults, employment status was imputed as well for the small percentage (less than 4%) of adults for whom employment status was unknown. Finally, the ratio of family income to the applicable Federal poverty thresholds was derived for families with missing incomes, based on the imputed income values. The imputation procedure incorporated a large number of predictors, including demographic and health-related variables.

For each year in 1997 – 2009, the data base for the NHIS multiply imputed data includes five files, one for each set of imputed values. For each person, each file contains: the values of family income, personal earnings, employment status, and the poverty ratio; flags indicating whether the value of each variable was imputed; and information for linking the data to other data from the NHIS. In the public-use version of the multiply imputed data, family income, personal earnings, and poverty ratio values are given. However, for each variable they are top-coded at the 95th percentile, and that the top five percent of values are set to this top-coded value.

For each survey year, data sets containing the imputed values, along with related documentation, can be obtained from the NHIS Web site (http://www.cdc.gov/nchs/nhis.htm).

1.4 Objective and Contents of this Report

The objective of this report is to describe the approach used to multiply impute income and earnings items in the NHIS and methods for analyzing the multiply imputed data. Sample program code and output are also provided.

Section 2 provides an overview of multiple imputation and a discussion of how multiply imputed data are analyzed. Section 3 contains a description of the imputation procedure that was used in this project. Finally, in Section 4, two examples are discussed to illustrate how to

analyze the multiply imputed NHIS data using the software packages SAS-callable SUDAAN and SAS-callable IVEware.

2. Multiple Imputation

2.1 Overview of Multiple Imputation

Imputation is a popular approach to handling nonresponse on items in a survey for several reasons. First, imputation adjusts for observed differences between nonrespondents and respondents; such an adjustment is generally not made by complete-case analysis. Second, imputation results in a completed data set, so that the data can be analyzed using standard software packages without discarding any observed values. Third, when a data set is being produced for analysis by the public, imputation by the data producer allows the incorporation of specialized knowledge about the reasons for missing data in the imputation procedure, including confidential information that cannot be released to the public. Moreover, the nonresponse problem is addressed in the same way for all users, so that analyses will be consistent across users.

Although single imputation, that is, imputing one value for each missing datum, enjoys the positive attributes just mentioned, analysis of a singly imputed data set using standard software fails to reflect the uncertainty stemming from the fact that the imputed values are plausible replacements for the missing values but are not the true values themselves. As a result, such analyses of singly imputed data tend to produce estimated standard errors that are too small, confidence intervals that are too narrow, and significance tests that reject the null hypothesis too often when it is true. For example, large-sample results reported in Rubin and Schenker (1986)

suggest that when the rate of missing information is 20% to 30%, nominal 95% confidence intervals computed from singly imputed data have actual coverage rates between 85% and 90%. Moreover, the performance of single imputation can be even worse when inferences are desired for a multi-dimensional quantity. For example, large-sample results reported in Li, Raghunathan, and Rubin (1991) demonstrate that for testing hypotheses about multi-dimensional quantities, the actual rejection rate under the null hypothesis increases as the number of components being tested increases, and the actual rate can be much larger than the nominal rate.

Multiple imputation (Rubin 1978, 1987) is a technique that seeks to retain the advantages of single imputation while also allowing the uncertainty due to imputation to be reflected in the analysis. The idea is to simulate $M > 1$ plausible sets of replacements for the missing values, thereby generating M completed data sets. The M completed data sets are analyzed separately using a standard method for analyzing complete data, and then the results of the M analyses are combined in a way that reflects the uncertainty due to imputation. Details of how to analyze multiply imputed data are provided in Section 2.2 and Appendix A. For public-use data, M is not usually larger than five, which is the value that has been used in multiply imputing missing data for the NHIS. Rubin (1996) argues that a small value of M is appropriate for multiple imputation, because the simulation involved in multiple imputation is only being used to handle the missing information, whereas the observed information is handled by the complete-data method used to analyze the completed data sets.

With multiple imputation, the M sets of imputations for the missing values are ideally independent draws from the predictive distribution of the missing values conditional on the observed values. Consider, for example, the simple case in which there are two variables, X and Y, with Y subject to nonresponse and X fully observed. Suppose further that the

imputation model specifies that: Y has a normal linear regression on X, that is, $Y = \beta_0 + \beta_1 X + \varepsilon$, where ε has a normal distribution with mean 0 and variance σ^2; and given X, the missing values of Y are only randomly different from the observed values of Y. After the regression of Y on X is fitted to the complete cases, a single set of imputations for the missing Y-values can be generated in two steps. First, values β_0^*, β_1^*, and σ^{*2} are drawn randomly from the joint posterior distribution of the regression parameters. For example, the appropriately scaled chi-square distribution could be used for drawing σ^{*2}, and the appropriate bivariate normal distribution could be used for drawing β_0^* and β_1^* given σ^{*2}. Second, for each nonrespondent, say nonrespondent i, the missing value of Y is drawn randomly as $\beta_0^* + \beta_1^* X_i + \varepsilon_i^*$, where X_i is the X-value for nonrespondent i, and ε_i^* is a value drawn from a normal distribution with mean 0 and variance σ^{*2}. The first step reflects the uncertainty due to the fact that the regression model was fitted to just a sample of data, and the second step reflects the variability of the Y-values about the regression line. Multiple imputations of the missing Y-values are generated by repeating the two steps independently M times. Although most imputation problems, including the imputation of missing data in the NHIS, are much more complicated than the simple example just presented, the basic principle illustrated by the simple example, that is, reflecting all of the sources of variability across the M sets of imputations, still applies.

2.2 Analyzing Multiply Imputed Data

2.2.1 General Procedures

Suppose that the primary interest is in estimating a scalar population quantity, such as a mean, a proportion, or a regression coefficient. The analysis of the M completed data sets resulting from multiple imputation proceeds as follows:

- Analyze each of the M completed data sets separately using a suitable software package designed for complete data (for example, SUDAAN or STATA).
- Extract the point estimate and the estimated standard error from each analysis.
- Combine the point estimates and the estimated standard errors to arrive at a single point estimate, its estimated standard error, and the associated confidence interval or significance test.

Technical details of how to analyze multiply imputed data are given in Appendix A. Briefly, however, the combined point estimate is the average of the point estimates obtained from the M completed data sets. The estimated variance of the combined point estimate is computed by adding two components. The first component is the average of the estimated variances obtained from the M completed data sets. The second component is the variation among the point estimates obtained from the M completed data sets. The latter component represents the uncertainty due to imputing for the missing values. Confidence intervals and significance tests are constructed using a t reference distribution.

One can carry out a multiple-imputation analysis by using any appropriate software package for analyzing the completed data sets and then using a spreadsheet program, a macro, a specially written program, or even just a calculator to combine the results of the analyses. However, it can be quite time consuming to perform the multiple-imputation analysis, especially

if many quantities of interest are involved. Fortunately, several software packages are becoming available that implement the combining rules for a variety of analytical techniques. Section 4 provides information about some of these software packages.

2.2.2 Combining Data Across Years of the NHIS

A common practice with the NHIS, especially when rare events or small subsets of the population are being studied, is to combine more than one year of data in order to increase the sample size. For analyses of the combined data, the data files are typically concatenated and the analysis weights adjusted accordingly. Botman and Jack (1995) and National Center for Health Statistics (2010, Appendix VI) provide further information on how to conduct such analyses as well as on issues that arise.

With the NHIS multiply imputed data, there are $M = 5$ completed data sets for each year. If it is desired to combine more than one year of data, then the corresponding completed data sets from the years in question can be concatenated to obtain $M = 5$ concatenated completed data sets. Suppose, for example, that the data from 1999 and 2000 were to be combined. Then the first completed data set from 1999 and the first completed data set from 2000 would be concatenated to create the first concatenated completed data set for 1999 – 2000. The analogous concatenations would be carried out for the second through fifth completed data sets, with the end result being $M = 5$ concatenated completed data sets for 1999 – 2000.

After $M = 5$ concatenated completed data sets have been created by combining data across years, each of the concatenated completed data sets is analyzed using the standard techniques for concatenated data from multiple years of the NHIS, as described by Botman and Jack (1995) and National Center for Health Statistics (2010, Appendix VI). The results of the five analyses are then combined using the rules given in Section 2.2.1 and Appendix A.

2.3 Analyzing Only a Single Completed Data Set

Users of the multiply imputed NHIS data who are unfamiliar with multiple imputation or who find the analysis of multiply imputed data cumbersome might be tempted to analyze only a single completed data set, such as the first of the five. Such an analysis, which is equivalent to using single imputation, would produce point estimates that are unbiased (under the assumption that the imputation model is correct). However, as discussed in Section 2.1, it would produce underestimates of variability and resultant inferences that may be inaccurate, since it would not account for the additional variability due to imputation.

When applying a model-selection procedure such as stepwise regression, it is not clear how to formally combine the results from M completed data sets. Therefore, an analyst might decide to apply the model-selection procedure to, for example, just the first completed data set. Since variability would be underestimated, such an approach would tend to judge more variables as "statistically significant" than would be the case if variability were estimated correctly. Thus, fewer variables would tend to be eliminated from the model under single imputation.

3. Procedure for Creating Imputations for the NHIS

The imputation of family income and personal earnings in the NHIS was complicated by several issues. First, these variables are hierarchical in nature, since one is reported at the family level whereas the other is reported at the person level. Second, there are structural dependencies among the variables in the survey. For example, individuals can only have earnings (given by one variable) if they are employed (as indicated by other variables). Third, in some cases, the income and earnings items needed to be imputed within bounds. For example, as discussed in Section 1.2, some families did not report exact income values but did report coarser income categories; such categories were used to form bounds for imputing exact income. Finally, there were several variables that were used as predictors in the imputation procedure. Such variables were of many types (e.g., categorical, continuous, count, ...), and they often had small percentages of missing values that needed to be imputed as well.

The following two sections describe the imputation procedure. Section 3.1 provides an overview of the steps in the procedure, the general algorithm used, and how features of the sample design were incorporated into the procedure. In Section 3.2, some additional details of the steps in the imputation procedure are described.

Note that in the process of imputing family income and personal earnings, missing values of several additional variables were imputed, and several new variables were created as well. These additional variables and imputed values were not retained in the final public-use data base for the NHIS multiply imputed data, except for the adult employment status and family poverty ratio that were mentioned in Section 1.3.

3.1 Overview of the Imputation Procedure

3.1.1 Steps in the Imputation Procedure

To handle the hierarchical nature of family income and personal earnings, it was decided to first impute the missing values of family income, together with the "family earnings," that is, the family total of personal earnings, for each family that had any employed adults with unknown personal earnings. Once these family-level items were imputed, missing values of personal earnings were imputed via imputation of the proportion of family earnings to be allocated to those family members with missing personal earnings.

Family income and family earnings were imputed first because there were other variables available that were expected to be especially useful in predicting these items. For example, as described in Sections 1.1 and 1.2, although exact family income was not reported for 25% to 34% of the families, either a fine or coarse categorical income value was available for the majority of these families. In addition, some families with missing values of family income had information available on family earnings and vice versa, and these two family-level variables were expected to be highly correlated with each other. Finally, the (log) mean and (log) standard deviation of reported family incomes were calculated by secondary sampling unit (SSU), and these contextual variables were used as predictors. (The SSUs in the NHIS were small clusters of housing units.)

In the imputation of family income and family earnings, several family-level covariates were used, including many summaries of the person-level covariates within each family. Most of the person-level covariates had very low rates of missingness. To facilitate their use, their missing values were imputed for adults (since employment and earnings items, as well as many of the person-level covariates, apply only to adults in the NHIS) prior to the imputation of family

income and family earnings. Any remaining missing values in the family-level covariates, due to missingness in person-level covariates for children, were imputed together with family income and family earnings.

To summarize, the sequence of steps in the imputation procedure was as follows:

1. Impute missing values of person-level covariates for adults.
2. Create family-level covariates.
3. Impute missing values of family income and family earnings, and any missing values of family-level covariates due to missing person-level covariates for children.
4. Impute the proportion of family earnings to be allocated to each employed adult with missing personal earnings, and calculate the resulting personal earnings.

The income and earnings items were not used in the initial imputation of covariates in step 1. To incorporate any relationships between the income and earnings items and the covariates into the imputations, after steps 3 and 4 were carried out, the procedure cycled through steps 1 – 4 five more times, with the income and earnings items (including the imputed values) now included as predictors in step 1. In each of these five additional cycles, the SSU-level (log) mean and (log) standard deviation of family incomes were also recalculated, with the imputed values included in the calculations.

To create multiple imputations, the entire imputation process described above was repeated independently five times.

3.1.2 Sequential Regression Multivariate Imputation

The imputations in each of steps 1 – 4 described in Section 3.1.1 were created using sequential regression multivariate imputation (SRMI) (Raghunathan *et al*. 2001), as implemented by the module **IMPUTE** in the software package IVEware (www.isr.umich.edu/src/smp/ive).

A brief description of SRMI is as follows; see Raghunathan *et al*. (2001) for details. Let X denote the fully-observed variables, and let $Y_1, Y_2, ..., Y_k$ denote k variables with missing values, ordered by the amount of missingness, from least to most. The imputation process for $Y_1, Y_2, ..., Y_k$ proceeds in c rounds. In the first round: Y_1 is regressed on X, and the missing values of Y_1 are imputed (using a process analogous to that described in the simple example of Section 2.1); then Y_2 is regressed on X and Y_1 (including the imputed values of Y_1), and the missing values of Y_2 are imputed; then Y_3 is regressed on X, Y_1, and Y_2, and the missing values of Y_3 are imputed; and so on, until Y_k is regressed on $X, Y_1, Y_2, ..., Y_{k-1}$, and the missing values of Y_k are imputed.

In rounds 2 through c, the imputation process carried out in round 1 is repeated, except that now, in each regression, all variables except for the variable to be imputed are included as predictors. Thus: Y_1 is regressed on $X, Y_2, Y_3, ..., Y_k$, and the missing values of Y_1 are re-imputed; then Y_2 is regressed on $X, Y_1, Y_3, ..., Y_k$, and the missing values of Y_2 are re-imputed; and so on. After c rounds, the final imputations of the missing values in $Y_1, Y_2, ..., Y_k$ are used.

For the regressions in the SRMI procedure, IVEware allows the following models:

- A normal linear regression model if the Y-variable is continuous;
- A logistic regression model if the Y-variable is binary;
- A polytomous or generalized logit regression model if the Y-variable is categorical with more than two categories;
- A Poisson loglinear model if the Y-variable is a count;

- A two-stage model if the Y-variable is mixed (i.e., semi-continuous), where logistic regression is used to model the zero/non-zero status for Y, and normal linear regression is used to model the value of Y conditional upon its being non-zero.

In addition, IVEware allows restrictions and bounds to be placed on the variables being imputed. As an example of a restriction, the imputation of family earnings was restricted to families with one or more employed adults (see Section 3.2.3). As an example of bounds, if a category rather than an exact value was reported for a family's income, the category's bounds were used in the imputation (see Section 3.2.3).

Because SRMI requires only the specification of individual regression models for each of the Y-variables, it does not necessarily imply a joint model for all of the Y-variables conditional on X. The decision to use SRMI and IVEware to create the imputations for the NHIS was influenced in large part by the complicating factors summarized at the beginning of Section 3 and discussed further in Section 3.2, specifically, the structural dependencies, the bounds, and the large number of predictors of varying types that had missing values. These complicating factors would be very difficult to handle using a method based on a full joint model. Moreover, without the complicating factors, the SRMI-based imputation procedure used in this project would actually be equivalent to the following two steps, corresponding to steps 3 and 4 in Section 3.1.1:

i. Impute the missing values of family income and family earnings based on a bivariate normal model (given predictors and transformations).

ii. Impute the proportion of family earnings to be allocated to each employed adult with missing personal earnings, based on a normal linear regression model for the logit of the proportion, and calculate the resulting personal earnings.

3.1.3 Reflecting the Sample Design in Creating the Imputations

When using multiple imputation in the context of a sample survey with a complex design, it is important to include features of the design in the imputation model, so that approximately valid inferences will be obtained when the multiply imputed data are analyzed (Rubin 1996).

The sample design of the NHIS was reflected in the imputations for this project via the inclusion of the following covariates: indicators for the distinct combinations of stratum and primary sampling unit (PSU); the survey weights; and SSU-level summaries of family income, as mentioned in Section 3.1.1.

3.2 Further Details of the Imputation Procedure

Additional details of the steps outlined in Section 3.1.1 are now described.

3.2.1 Step 1: Imputing Person-Level Covariates for Adults

The variables included in the imputation of person-level covariates for adults are listed in Table 1 of Appendix B. The imputation of person-level covariates was carried out in two parts, because imputed values from the first part were needed to set restrictions for the imputations in the second part. In the first part, the variables for whether a person has a limitation of activity (`LIM_ACT`), for whether specific conditions caused the limitation (`LA_GP01`, `LA_GP02`, ..., `LA_GP09`), and for number of hours worked per week (`WRKHRS`), were omitted, and any missing values on the other variables were imputed. Then, the variable `LIM_ACT` was created from the individual items on limitations of activity (`PLAADL`, `PLAIADL`, etc.). In the second part, any missing values on `WRKHRS` and `LA_GP01`, `LA_GP02`, ..., `LA_GP09` were imputed, conditional on the values from the first part. An upper bound of 95 was set for `WRKHRS`.

Along with the person-level covariates, the log mean (`SSUFINL`) and the log standard deviation (`SSUSTDL`) of reported family incomes within the SSU were treated as person-level

variables and imputed if necessary. Missing values in SSUFINL occurred if no families in the SSU had reported incomes, if the mean reported family income was 0, or if the mean reported family income was top-coded (at $999,995), in which case the log top-code value was used as a lower bound in the imputation. Missing values in SSUSTDL occurred if fewer than two families in the SSU had reported incomes. If this was the case, the largest observed log standard deviation among the SSUs was used as an upper bound in the imputation. After the missing values of SSUFINL and SSUSTDL were imputed, averages of the values within each SSU were computed for use in subsequent steps.

3.2.2 Step 2: Creating Family-Level Covariates

The person-level variables from step 1 were summarized, by family, to create family-level covariates for use in imputing family income and family earnings. These family-level covariates are included in the listing in Table 2 of Appendix B. Examples include the total number of earners in a family (FM_EARN) and an indicator for whether a family has at least one person with Medicaid coverage (FM_MCAID).

After imputation of the person-level covariates for adults in step 1, some of the family-level covariates that were created still had small residual levels of missingness, due to missing values of some person-level covariates for children. These missing values in the family-level covariates were imputed together with family income and family earnings in step 3.

3.2.3 Step 3: Imputing Family Income and Family Earnings (and Family-Level Covariates)

The variables included in the imputation at the family level are listed in Table 2 of Appendix B.

To determine a good transformation for family income and family earnings to conform to the normality assumption in the imputation model, Box-Cox transformations (Box and Cox 1964) were estimated from the complete cases for the regressions predicting family income and

family earnings. The closest simple transformation suggested by the Box-Cox analysis was the cube-root transformation, which is also close to and consistent with the optimal transformation (the power 0.375) found by Paulin and Sweet (1996) in modeling income data from the Consumer Expenditure Survey of the Bureau of Labor Statistics. After the imputation procedure was completed, the variables were transformed back to their original scale.

The imputation of family earnings was restricted to families with one or more adult earners. For many families, there was partial information available on family earnings, because personal earnings were observed for some family members and missing for others. For each family with such partial information, the sum of the observed personal earnings was used as the lower bound in imputing the family earnings. With regard to family income, as mentioned previously, there were several families for which an exact income was not reported, but an income category was reported. In each such case, the bounds specified by the reported category were used in imputing the family income. In addition to the bounds just described, when a reported family income or family earnings value was top-coded, an exact value at least as large as the top-code value ($999,995 for income and $999,995 for earnings) was imputed. The imputation for top-coded values was just an intermediate step that was carried out so that the distribution from which other values were imputed would not be distorted by the top-coding. After the entire imputation process was completed, the top-coding of family income values larger than $999,996 was reinstated.

3.2.4 Step 4: Imputing Personal Earnings

For any family that had only one employed adult with missing personal earnings, once the family earnings were imputed in step 3, the person's missing earnings could be determined by

subtracting the observed personal earnings for members of the family from the imputed family earnings.

For families that had more than one employed adult with missing personal earnings, in the imputation of the proportion of family earnings to be allocated to each employed adult with missing personal earnings, the logit (log-odds) transformation was applied to the proportions, and a normal linear regression model was used for the logit. The variables included in this imputation step are listed in Table 3 of Appendix B.

After the logits were imputed, they were transformed back to proportions. Then, within each family, the proportions for the employed adults with missing personal earnings were rescaled so that they would sum to the total proportion of family earnings not accounted for by persons whose earnings had been observed. Imputed personal earnings were calculated from an imputed proportion via multiplication of the proportion by the family earnings.

During the imputation process, the imputed proportion corresponding to each top-coded reported value of personal earnings was bounded below so that the resulting imputed personal earnings value would be at least as large as the top-code value ($999,995). As with family incomes (see Section 3.2.3), after the entire imputation process was carried out, the top-coding of personal earnings values larger than $999,995 was reinstated.

3.3 Inconsistencies Between Family Income and Family Earnings

Because the items suggested to be included in family income in the NHIS questionnaire are all nonnegative and include the personal earnings of family members (see Section 1.1), it follows that family income should ideally be at least as large as family earnings. However, as noted in Section 1.1, family income in the NHIS is estimated by the respondent rather than being constructed by summing responses to more detailed questions, such as the question about

personal earnings of members of the family. Thus, some inconsistencies between family income and family earnings, in terms of the former being lower than the latter, might be expected.

In the 1997 – 2009 NHIS, 7% to 10% of responding families per year had reported family incomes that were lower than the reported family earnings. (The percentages presented in this section are weighted. As was the case in Section 1.2, the unweighted percentages are close to their weighted counterparts.) Moreover, the imputation procedure results in a larger percentage of families with family incomes lower than family earnings; 13% to 19% of the families in a completed data set (including both observed and imputed values) have such inconsistencies.

A reason for the higher rate of inconsistencies in the imputed data is as follows. In addition to the 7% to 10% inconsistency rate in the reported data from which the imputation model is estimated, 36% to 44% of responding families had reported family incomes exactly equal to their reported family earnings. Since the imputation model does not force equality of family income and family earnings for any families, the imputation procedure tends to produce differences between family income and family earnings that are close to zero for a large percentage of families, but several such differences will be positive and several others will be negative.

As part of this project, research has been conducted on restricting the imputed value of family income to be at least as large as the imputed value of family earnings, as well as on imputing new values of family income for those families whose reported family incomes and family earnings are inconsistent. The methods that have been developed to date tend to distort the marginal distribution of family income and the marginal distribution of family earnings. Given that the primary interest of data analysts is in each variable on its own, especially family income and its ratio to the poverty threshold, it was decided that family income and family

earnings would be imputed without imposing consistency. Research into resolving the issue of inconsistency will continue.

4. Software for Analyzing Multiply Imputed Data

As mentioned in Section 2.2.1, after analyzing each of the M completed data sets resulting from multiple imputation, one can combine the results of the M analyses by using a spreadsheet program, a macro, a specially written program, or even just a calculator. However, the increasing availability of software packages that implement the combining rules is helping to facilitate multiple-imputation analyses.

In this section, two examples are considered to illustrate analyses of the multiply imputed NHIS data using both SAS-callable SUDAAN and SAS-callable IVEware. STATA macros for performing multiple-imputation analyses have also been developed by Carlin *et al*. (2003), although examples of analyses using the macros are not given here. The STATA macros can be used to fit regression models with complex survey data. Obtaining multiple-imputation estimates and estimated standard errors based on cross-tabulations or descriptive measures is not possible without framing them as regression problems.

Both of the examples use data from the 2000 NHIS. The analyses of interest for the two examples, in terms of variables defined in the table on the next page, are as follows:

Example 1: Cross-tabulation of `POVERTYI` and `NOTCOV`.

Example 2: Logistic regression of the outcome `HSTAT` on the predictors `POVERTYI`, `AGEGR6R`, `HPRACE`, `USBORN`, `MSAR`, `REGIONR`, and `SEX`.

The SAS code given in Appendix C, Section C.1 was used to create five completed data sets (ANAL1-5) containing only the variables used in the two example analyses. The process involved in creating these data sets is as follows:

a. Extract the income-related variables from the files containing the five sets of imputations (INCMIMP1-5).
b. Extract the other necessary variables, including the design variables STRATUM, PSU, and WTFA, from the NHIS person-level file (e.g., PERSONSX).
c. Merge each of the five sets of income-related variables from step a with the other variables from step b, and perform the necessary recodes to create each of the five completed data sets for analysis.

Definitions of Variables Used in the Examples

Variable Name[a]	Definition	Code
STRATUM	Stratum	
PSU	Primary Sampling Unit	
WTFA	Survey Weight	
AGEGR6R	Age Category (Recode of variable AGE_P from file PERSONSX)	1 = <18 2 = 18-24 3 = 25-34 4 = 35-44 5 = 45-54 6 = 55-64 7 = 65+
SEX	Gender	1 = male 2 = female
HPRACE	Race/Ethnicity Category (Recode of variables ORIGIN_I and RACEREC_I from file PERSONSX)	1 = Hispanic 2 = black, non-Hispanic 3 = other, non-Hispanic 4 = white, non-Hispanic
REGIONR	Region of the Country (Recode of variable REGION from file PERSONSX)	1 = northeast 2 = south 3 = west 4 = midwest
MSAR	Residence in MSA (Recode of variable MSASIZEP from file PERSONSX)	1 = MSA 2 = not MSA
USBORN	Country of Birth (Recode of variable USBIRTH_P from file PESONSX)	1 = not born in US 2 = born in US
POVERTYI	Poverty Ratio Category (Recode of variable RAT_CATI from each of files INCMIMP1 – INCMIMP5)	1 = <100% 2 = 100-199% 3 = 200-399% 4 = 400%+
NOTCOV	Health Insurance Coverage	1 = uninsured 2 = insured
HSTAT	Health Status (Recode of variable PHSTAT from file PERSONSX)	1 = excellent to good 2 = fair/poor

[a] Except for POVERTYI, the variable name is either the same as in the 2000 public-use file PERSONSX, or is a recode of the variable(s) in PERSONSX specified in the Definition column. POVERTYI is a recode of a variable in the public-use files INCMIMP1 - INCMIMP5 that contain the NHIS multiply imputed data.

4.1 Analysis Using SAS-Callable SUDAAN

SAS-callable SUDAAN is a versatile software package for analyzing data from complex surveys. This section provides relevant code for carrying out the analyses for Examples 1 and 2 using SUDAAN Version 9.0, which includes a built-in option for analyzing multiply imputed data. For those who do not have access to this recent version of SUDAAN, an example is also provided of SAS commands to be used with SAS Version 6.12 or higher and SUDAAN Version 7 or higher, without a built-in option for analyzing multiply imputed data. Note that the examples of code provided in this section have to be modified for particular needs.

SUDAAN version 9.0 can process the NHIS multiply imputed data either from five separate files, or from a single file containing the five sets of imputed values, with each set in a separate variable. The examples described in this section are based on using five separate files. However, it may be more efficient to create a single file with the five sets of imputed values, and merge this file with the other analysis variables of interest before calling the SUDAAN procedure. SUDAAN processes the single file with a MI_VAR statement to identify the five variables containing the imputed values. An advantage of this approach is that it requires less storage space because it avoids replication of the variables that are not imputed. For more information on this approach and the MI_VAR statement see the section "Using multiple variables on the same data set" in the SUDAAN 9.0 documentation.

4.1.1 SUDAAN Version 9.0 with a Built-In Option for Multiple Imputation

The multiple files for the completed data sets can be identified in two different ways in SUDAAN Version 9.0. The first is to name the completed data sets with consecutive numbers at the end of the name as was done with `ANAL1-5` above. Setting the system variable `MI_COUNT` via the option

```
MI_COUNT=count
```

indicates the number of completed data sets, `count`, to be analyzed. Upon encountering this option, SUDAAN will automatically perform the multiple-imputation analysis. Note that `count` must be at least 2; otherwise, SUDAAN will produce an error message and halt. In addition, the files containing the completed data sets must all be located in the same directory and must be numbered consecutively. Each data set must be sorted by the `"NEST"` variables.

The second approach to identifying multiply imputed data is useful when the files containing the completed data sets either are not numbered consecutively or reside in different directories. The command

```
MI_FILES=file names;
```

identifies the completed data sets. For example, suppose that the SAS files were named `one`, `two`, `three`, `four`, and `five` and were located in the same directory, `C:\NHIS`. Then the following commands would be used:

```
proc anyprocedure data="c:\nhis\one" filetype=sas design=wr;
mi_files="c:\nhis\two" "c:\nhis\three" "c:\nhis\four"
"c:\nhis\five";
```

The first approach to identifying multiply imputed data will be followed here. For Example 1, `PROC CROSSTAB` is used (although `PROC DESCRIPT` could also be used after the recoding of `NOTCOV` as a binary variable). The syntax is the same as usual except that the multiple-imputation analysis is requested via a specification of the system variable `MI_COUNT` as one of the options. Without this option, SUDAAN will perform an analysis of only the first completed data set. For Example 2, the logistic regression model is fitted using `PROC RLOGIST`. The SUDAAN code for both examples is given in Appendix C, Section C.2, and the output is in Appendix D.

4.1.2 SAS Commands for Use with SUDAAN Version 7 or Higher without a Built-In Option for Multiple Imputation

The logistic regression analysis (i.e., Example 2) is now illustrated using commands in SAS and SAS-callable SUDAAN, for those who do not have access to SUDAAN Version 9.0. The three steps outlined in Section 2.2.1 are carried out. That is, each completed data set is analyzed; the point estimates and the estimated standard errors are stored; and the point estimates and estimated standard errors are combined using the rules given in Section 2.2.1 and Appendix A. The first two steps are performed by one macro, and then the combining of estimates is performed by further commands. The full set of commands is shown in Appendix C, Section C.3, and the output is in Appendix E.

4.2 Analysis Using SAS-Callable IVEware

SAS users can download IVEware, a free SAS-callable software package, from the Web site www.isr.umich.edu/src/smp/ive. IVEware has three modules for performing various multiple-imputation analyses incorporating complex sample designs. **DESCRIBE** performs descriptive analyses such as the estimation of means, proportions, and contrasts. It uses Taylor series methods to estimate variances in the analysis of each completed data set. **REGRESS** performs linear, logistic, polytomous, Poisson, Tobit and proportional hazards regression analyses. Variance estimates in the analysis of each completed data set are obtained using the jackknife repeated replication technique. **SAMOD** performs various other analyses such as CALIS (structural equation models), CATMOD (categorical data analysis), MIXED (random effects models), NLIN (nonlinear regression models), and GENMOD (generalized linear regression and GEE models), to name a few. Again, variance estimates for each completed data set are based on the jackknife repeated replication technique. Multiple-imputation analyses in IVEware are

performed using the combining rules described in Rubin and Schenker (1986) and summarized in Section 2.2.1 and Appendix A.

IVEware also contains a fourth module, **IMPUTE**, which actually performs multiple imputation for missing data. As discussed in Section 3.1.2, this module performs sequential regression multivariate imputation, and it was used to create the multiple imputations for the NHIS.

Details about the features of IVEware are provided in the documentation, "IVEware: Imputation and Variance Estimation Software," which can be downloaded from the Web site given above.

Code for using IVEware to perform analyses for Examples 1 and 2 is illustrated in Appendix C, Section C.4, with the corresponding output given in Appendix F.

Appendix A. Technical Details for Analyzing Multiply Imputed Data

Suppose that M completed data sets have been generated via multiple imputation, and let Q denote the scalar population quantity of interest. Application of the chosen method of analysis to the l^{th} completed data set yields the point estimate $\hat{Q}_l$ and its estimated variance (square of the estimated standard error) U_l, where $l = 1,2,...,M$. It is important to analyze each data set separately to derive the M point estimates and estimated variances.

The combined multiple-imputation point estimate is

$$\overline{Q}_M = \frac{1}{M}\sum_{l=1}^{M}\hat{Q}_l \,. \tag{1}$$

The estimated variance of this point estimate consists of two components. The first component, the "within-imputation variance"

$$\overline{U}_M = \frac{1}{M}\sum_{l=1}^{M}U_l \,,$$

is, approximately, the variance that one would have obtained had there been no missing data. The second component, the "between-imputation variance"

$$B_M = \frac{1}{M-1}\sum_{l=1}^{M}(\hat{Q}_l - \overline{Q}_M)^2 \,,$$

is the component of variation due to differences across the M sets of imputations.

The total estimated variance of the multiple-imputation point estimate $\overline{Q}_M$ is

$$T_M = \overline{U}_M + \frac{M+1}{M}B_M \,. \tag{2}$$

The factor $(M+1)/M$ is a correction for small M. Furthermore, it is shown in Rubin and Schenker (1986) and Rubin (1987, Section 3.3) that, approximately,

$$T_M^{-1/2}(Q - \overline{Q}_M) \sim t_\nu$$

where the degrees of freedom ν for the t distribution are given by

$$\nu = (M-1)\hat{\gamma}_M^{-2},$$

with

$$\hat{\gamma}_M = \frac{M+1}{M}\frac{B_M}{T_M}.$$

The quantity $\hat{\gamma}_M$ measures the proportionate share of T_M that is due to between-imputation variability; it is also approximately the fraction of information about Q that is missing due to nonresponse (Rubin 1987, p. 93).

For a multi-dimensional population quantity Q, Li, Raghunathan, and Rubin (1991) developed multiple-imputation procedures for significance testing when the hypothesis to be tested involves several components of Q simultaneously. In addition, Li, Meng, Raghunathan, and Rubin (1991) developed procedures for combining test statistics and p-values (rather than point estimates and estimated variances) computed from multiply imputed data.

The procedures described above assume that the degrees of freedom that would be used for analyzing the complete data if there were no missing values, i.e., the "complete-data degrees of freedom," are large (or infinite); that is, a large-sample normal approximation would be valid for constructing confidence intervals or performing significance tests if there were no missing data. This is clearly not true in many survey settings, where the number of sampled PSUs may be small, and a t reference distribution would be used if there were no missing data. For example, for a survey involving H strata with 2 PSUs selected from each stratum, the complete-data degrees of freedom for inferences about the population mean are H.

Barnard and Rubin (1999) relaxed the assumption of large complete-data degrees of freedom and suggested the use of

$$\nu' = \left(\frac{1}{\nu} + \frac{1}{k} \right)^{-1}$$

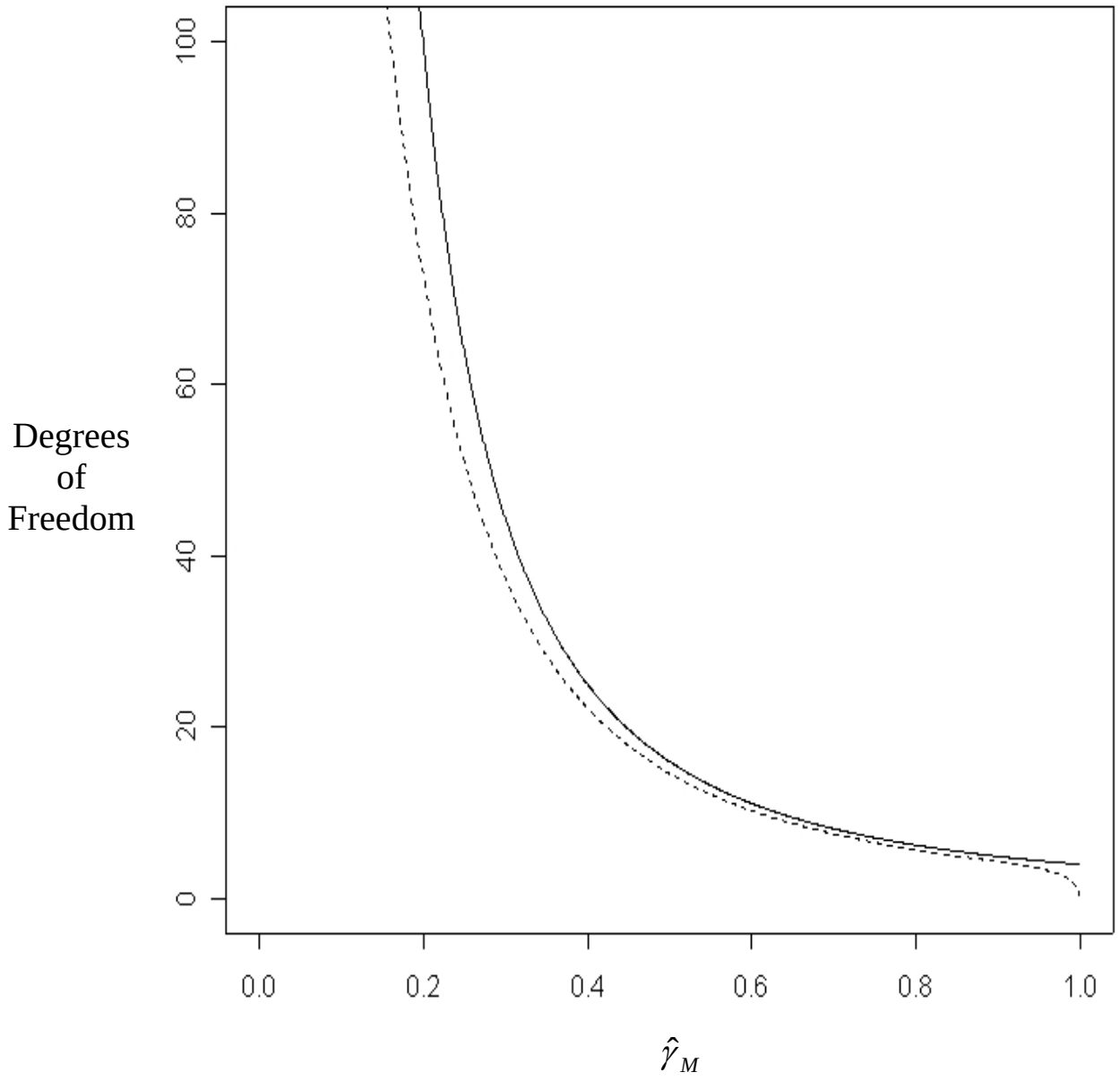
for the multiple-imputation analysis, where

$$k = \frac{df\,(df+1)}{df+3}(1-\hat{\gamma}_M)\,,$$

and df are the complete-data degrees of freedom.

For the NHIS multiply imputed data, $M = 5$, and the complete-data degrees of freedom, df, are 339 for many analyses. For ν' or ν greater than 100, the normal approximation is generally valid. To assess how different ν' and ν are when they are smaller, the figure on the next page provides a plot of ν and ν' as functions of $\hat{\gamma}_M$ when $\nu' \leq 100$. From the plot, it appears that, for many analyses of the NHIS data, use of either ν or ν' should give similar results, although use of ν' will be slightly more conservative (smaller degrees of freedom).

Rubin-Schenker (solid curve) and Barnard-Rubin (dashed curve) degrees of freedom as a function of the approximate fraction of missing information $\hat{\gamma}_M$, when the complete-data degrees of freedom are equal to 339



Appendix B. Variables Included in the Imputation Process

TABLE 1
Variables included in imputation of person-level covariates for adults (Step 1).

Variable Name	Label and Code values
SEX	Sex 1 = male 2 = female
AGEGROUP	Recoded age group 0 = under 18 years old 1 = 18 - 24 years old 2 = 25 – 44 years old 3 = 45 - 64 years old 4 = 65+ years old
ORIGIN	Ethnic origin 1 = Hispanic 2 = Non-Hispanic
RACEREC	Race recode (year 1997-2005) 1 = white 2 = black 3 = other Race recode (year 2006+) 1 = white 2 = black 3 = Asian 4 = other
MARRY	Marital status recode 1 = married with or without spouse in HH, or living w/ partner 2 = divorced, widowed, separated 3 = never married 4 = 14 or fewer years old
FM_SIZER	Family size recode 1 = 1 person family 2 = 2 person family 3 = 3 person family 4 = 4+ person family
URB_RRL	Urban/Rural 1 = Urban 2 = Rural
*MSA	MSA/non-MSA residence (From 2007, MSA is created from recoding CBSASTAT and CBSASIZE) 1 = in MSA; in central city 2 = in MSA, not in central city 3 = not in MSA
WTFA	Final person weight
STRATPSU	Stratum by PSU combination (year 1997 – 2005) Stratum and PSU combination recoded based on STRAT_I, STRAT_D, PSU_I from the NHIS inhouse data file.(year 2006+)
PLAADL	Needs help with ADL (age >= 3) 1 = yes 2 = no
PLAIADL	Needs help with chores, shopping, etc. (age >= 5) 1 = yes 2 = no
PLAWKNOW	Unable to work due to health problem (age >= 18) 1 = yes 2 = no
PLAWKLIM	Limited kind/amount of work due to health problem (age >= 18) 1 = yes 2 = no 3 = unable to work (PLAWKNW = 1)
PLAWALK	Has difficulty walking without equipment 1 = yes 2 = no
PLAREMEM	Limited by difficulty remembering 1 = yes 2 = no

PLIMANY	Limited in any other way 1 = yes 2 = no 3 = limitation previously mentioned
LIM_ACT	Limited in any way (at least mentioned one limitation) 1 = at least 1 limitation 2 = no limitation 3 = under 18 years old
LA_GP01	Vision or hearing problem causes limitation recode (18+ with at least 1 limitation) 1 = mentioned 2 = not mentioned 3 = no limitation or under 18 years old
LA_GP02	Arthritis/rheumatism, back/neck, or muscular-skeletal problem causes limitation recode (18+ with at least 1 limitation) 1 = mentioned 2 = not mentioned 3 = no limitation or under 18 years old
LA_GP03	Fracture/bone/joint injury, other injury, or missing limb/finger causes limitation recode (18+ with at least 1 limitation) 1 = mentioned 2 = not mentioned 3 = no limitation or under 18 years old
LA_GP04	Heart, stroke, hypertension, or circulatory problem causes limitation recode (18+ with at least 1 limitation) 1 = mentioned 2 = not mentioned 3 = no limitation or under 18 years old
LA_GP05	Diabetes or endocrine problem causes limitation recode (18+ with at least 1 limitation) 1 = mentioned 2 = not mentioned 3 = no limitation or under 18 years old
LA_GP06	Lung/breath problem causes limitation recode (18+ with at least 1 limitation) 1 = mentioned 2 = not mentioned 3 = no limitation or under 18 years old
LA_GP07	Senility or nervous system condition causes limitation recode (18+ with at least 1 limitation) 1 = mentioned 2 = not mentioned 3 = no limitation or under 18 years old
LA_GP08	Depression/anxiety/emotion, alcohol, drug, or other mental problem causes limitation recode (18+ with at least 1 limitation) 1 = mentioned 2 = not mentioned 3 = no limitation or under 18 years old
LA_GP09	Other problem causes limitation recode (18+ with at least 1 limitation) 1 = mentioned 2 = not mentioned 3 = no limitation or under 18 years old
PHSTAT	Recoded health status 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor
PDMED12M	Delayed medical care due to cost in the past 12 months 1 = yes 2 = no
PNMED12M	Did not get medical care due to cost in the past 12 months 1 = yes 2 = no
PHOSPYR	In a hospital overnight in the past 12 months 1 = yes 2 = no
P10DVYR	Received health care from doctor 10+ times in the past 12 months 1 = yes 2 = no

M_CARE	Medicare coverage recode 1 = yes (covered, with or without information) 2 = no (not covered)
M_CAID	Medicaid coverage recode 1 = yes 2 = no
MILITRY	Military coverage recode 1 = yes (military/VA/CHAMPUS/TRICARE/don’t know type) 2 = no
PRIVATW	Private insurance coverage recode; at least 1 plan is paid by employer 1 = at least 1 private plan was obtained through employer 2 = all the private plans were not obtained through employer
PRIVATS	Private insurance coverage recode; at least 1 plan is paid by self 1 = at least 1 private plan was obtained through self 2 = all the private plans were not obtained through self
USBRTH	Born in US 1 = yes 2 = no
EDUCR	Education recode 1 = high school or less 2 = HS grad or equiv 3 = some college 4 = college graduate 5 = more than college 6 = professional degree
EMPSTAT	Last year’s employment status recode 1 = 18+, worked for pay last year 2 = 18+, not worked for pay last year 3 = under 18 years old
WRKHRS	Hours worked per week 1 – 95 = worked 1 to 95+ hours
SSUFINL	Logarithm of mean family income in SSU 0 – 13.816 = logarithm of mean
SSUSTDL	Logarithm of (standard deviation + 1) of family income in SSU 0 – 13.199 = logarithm of standard deviation
PSAL	Person received income from wage/salary 1 = yes 2 = no 3 = under 18
PSEINC	Person received income from self-employment 1 = yes 2 = no 3 = under 18
PSSRR	Person received income from Social Security 1 = yes 2 = no
PPENS	Person received income from other pension 1 = yes 2 = no
PSSI	Person received income from Supplement Security Income (SSI) 1 = yes 2 = no
PSSRRDB	Person received income from Social Security Disability Insurance (SSDI) 1 = yes 2 = no 3 = ineligible due to age (65+), or received social security
PTANF	Person received income from welfare/AFDC 1 = yes 2 = no
PINTRST	Person received income from interest 1 = yes 2 = no
PDIVD	Person received income from dividend 1 = yes 2 = no
PCHLDSP	Person received child support 1 = yes 2 = no
PINCOT	Person received income from other sources 1 = yes 2 = no

HOUSER	House ownership recode 1 = owned or being bought 2 = rented or other arrangement
PSSID	Receive SSI due to a disability 1 = yes 2 = no 3 = did not receive SSI
PSSAPL	Person not receiving SSI and has ever applied for SSI 1 = yes 2 = no 3 = received SSI
PSSRRD	Receive SSDI due to a disability 1 = yes 2 = no 3 = AGE > 65 or did not receive SSDI
PSDAPL	Person not receiving SSDI and has ever applied for SSDI 1 = yes 2 = no 3 = received SSDI
PFSTP	Person was authorized to receive food stamps 1 = yes 2 = no

* Variable was recoded using CBSASTAT and CBSASIZE, so it is comparable to the definition of MSA which is used in the 1997-2005 files.

TABLE 2
Variables included in imputation at the family level (Steps 2 & 3).

Variable Name	Label and Code values
URB_RRL	Urban/Rural 1 = Urban 2 = Rural
*MSA	MSA/non-MSA residence (From 2007, MSA is created from recoding CBSASTAT and CBSASIZE) 1 = in MSA; in central city 2 = in MSA, not in central city 3 = not in MSA
WTFA_FAM	Final family weight
**STRATPSU	Stratum by PSU combination (year 1997 – 2005) Stratum and PSU combination recoded based on STRAT_I, STRAT_D, PSU_I from the NHIS inhouse data file.(year 2006+)
ADULT	Total number of adults in a family
CHILD	Total number of children in a family
M_TWRKHR	Total number of work hours of male family members
F_TWRKHR	Total number of work hours of female family members
M_ERNAGE	Average age of male earners in a family
F_ERNAGE	Average age of female earners in a family
FM_EARN	Total number of earners in a family
P_HISP	Proportion of Hispanics in a family
P_WHITE	Proportion of whites in a family
P_BLACK	Proportion of blacks in a family
FM_ADL1	Family having family members (age >= 3) with PLAADL = 1 (Needs help with ADL) 1 = at least one family member has 2 = none of the family members has
FM_IADL1	Family having family members (age >= 5) with PLAIADL = 1 (Needs help with chores, shopping, etc.) 1 = at least one family member has 2 = none of the family members has
FM_WKNW1	Family having family members (age >=18) with PLAWKNOW = 1 (Unable to work due to health problem) 1 = at least one family member has 2 = none of the family members has
FM_WKLM1	Family having family members (age >=18) with PLAWKLIM = 1 (Limited kind/amt of work due to health problem) 1 = at least one family member has 2 = none of the family members has
FM_WALK1	Family having family members with PLAWALK = 1 (Has difficulty walking without equipment) 1 = at least one family member has 2 = none of the family members has
FM_REM1	Family having family members with PLAREMEM = 1 (Limited by difficulty remembering) 1 = at least one family member has 2 = none of the family members has
FM_MANY1	Family having family members with PLIMANY = 1 (Limited in any other way) 1 = at least one family member has 2 = none of the family members has
FM_GP011	Family having family members with LA_GP01 = 1 (Vision or hearing problem causes limitation) 1 = at least one family member has 2 = none of the family members has
FM_GP021	Family having family members with LA_GP02 = 1 (Arthritis/rheumatism, back/neck, or muscular-skeletal problem causes limitation) 1 = at least one family member has 2 = none of the family members has

FM_GP031	Family having family members with LA_GP03 = 1 (Fracture/bone/joint injury, other injury, or missing limb/finger causes limitation) 1 = at least one family member has 2 = none of the family members has
FM_GP041	Family having family members with LA_GP04 = 1 (Heart, stroke, hypertension, or circulatory problem causes limitation) 1 = at least one family member has 2 = none of the family members has
FM_GP051	Family having family members with LA_GP05 = 1 (Diabetes or endocrine problem causes limitation) 1 = at least one family member has 2 = none of the family members has
FM_GP061	Family having family members with LA_GP06 = 1 (Lung/breath problem causes limitation) 1 = at least one family member has 2 = none of the family members has
FM_GP071	Family having family members with LA_GP07 = 1 (Senility or nervous system condition causes limitation) 1 = at least one family member has 2 = none of the family members has
FM_GP081	Family having family members with LA_GP08 = 1 (Depression/anxiety/emotion, alcohol, drug, or other mental problem causes limitation) 1 = at least one family member has 2 = none of the family members has
FM_GP091	Family having family members with LA_GP09 = 1 (Other problem causes limitation) 1 = at least one family member has 2 = none of the family members has
FM_DMED	Family having family members with PDMED12M = 1 (Delayed medical care due to cost in the past 12 month) 1 = at least one family member has 2 = none of the family members has
FM_NMED	Family having family members with PNMED12M = 1 (Did not get medical care due to cost in the past 12 month) 1 = at least one family member has 2 = none of the family members has
FM_HOSP	Family having family members with PHOSPYR = 1 (In a hospital overnight in the past 12 months) 1 = at least one family member has 2 = none of the family members has
FM_DVYR	Family having family members with P10DVYR = 1 (Received health care from doctor 10+ times in the past 12 months) 1 = at least one family member has 2 = none of the family members has
FM_MCARE	Family having family members with M_CARE = 1 (Recoded Medicare coverage) 1 = at least one family member has 2 = none of the family members has
FM_MCAID	Family having family members with M_CAID = 1 (Recoded Medicaid coverage) 1 = at least one family member has 2 = none of the family members has
FM_MILIT	Family having family members with MILITRY = 1 (Military coverage recode) 1 = at least one family member has 2 = none of the family members has
FM_PRIVW	Family having family members with PRIVATW = 1 (Private insurance coverage; at least 1 plan is paid by employer) 1 = at least one family member has 2 = none of the family members has
FM_PRIVS	Family having family members with PRIVATS = 1 (Private insurance coverage; at least 1 plan is paid by self) 1 = at least one family member has 2 = none of the family members has
FM_USBRN	Family having family members with USBRTH = 1 (Born in US) 1 = at least one family member has 2 = none of the family members has

FM_HLTH1	Family having family members with PHSTAT = 1 or 2 (Excellent or very good health) 1 = at least one family member has 2 = none of the family members has
FM_HLTH2	Family having family members with PHSTAT = 3 (Good health) 1 = at least one family member has 2 = none of the family members has
FM_HLTH3	Family having family members with PHSTAT = 4 or 5 (Fair or poor health) 1 = at least one family member has 2 = none of the family members has
FM_HIEDU	Highest education attainment of family members 1 = high school or less (1-12, 96) 2 = HS grad or equiv (13,14) 3 = Some college (15-17) 4 = College graduate (18) 5 = more than college or prof. degrees (19-21) 6 = all family members are under 18
FM_SSRR	Family having family members with PSSRR = 1 (Person received income from Social Security) 1 = at least one family member has 2 = none of the family members has
FM_PENS	Family having family members with PPENS = 1 (Person received income from other pension) 1 = at least one family member has 2 = none of the family members has
FM_SSI	Family having family members with PSSI = 1 (Person received income from Supplement Security Income (SSI)) 1 = at least one family member has 2 = none of the family members has
FM_SSDI	Family having family members with PSSRRDB = 1 (Person received income from Social Security Disability Insurance (SSDI)) 1 = at least one family member has 2 = none of the family members has
FM_AFDC	Family having family members with PAFDC = 1 (Person received income from welfare/AFDC) 1 = at least one family member has 2 = none of the family members has
FM_INST	Family having family members with PINTRST = 1 (Person received income from interest) 1 = at least one family member has 2 = none of the family members has
FM_DIVD	Family having family members with PDIVD = 1 (Person received income from dividend) 1 = at least one family member has 2 = none of the family members has
FM_CHSP	Family having family members with PCHLDSP = 1 (Person received child support) 1 = at least one family member has 2 = none of the family members has
FM_INCO	Family having family members with PINCOT = 1 (Person received income from other sources) 1 = at least one family member has 2 = none of the family members has
FM_HOUS	Family having family members with HOUSER = 1 (Recoded house ownership) 1 = at least one family member has 2 = none of the family members has
FM_RSSI	Family having family members with PSSID = 1 (Receive SSI due to a disability) 1 = at least one family member has 2 = none of the family members has
FM_ASSI	Family having family members with PSSAPL = 1 (Person not receiving SSI and has ever applied for SSI) 1 = at least one family member has 2 = none of the family members has
FM_RSSD	Family having family members with PSSRRD = 1 (Receive SSDI due to a disability) 1 = at least one family member has 2 = none of the family members has

FM_ASSD	Family having family members with PSDAPL = 1 (Person not receiving SSDI and has ever applied for SSDI) 1 = at least one family member has 2 = none of the family members has
FM_FSTP	Family having family members with PFSTP = 1 (Person was authorized to receive food stamps) 1 = at least one family member has 2 = none of the family members has
PHONSRV	Phone service 1 = yes (has phone service) 2 = no (no phone service or working phone)
INTV_LNG	Interview language 1 = English only 2 = Not English only
FM_INCB	Cube root of total male and female earnings
FAMINB	Cube root of total family income
SSUFINWL	Logarithm of mean family income in SSU (recalculated after imputation step 1)
SSUSDNWL	Logarithm of standard deviation of family income in SSU (recalculated after imputation step 1)
P_F_EARN	Proportion of female earners to the total family earners

* Variable was recoded using CBSASTAT and CBSASIZE, so it is comparable to the definition of MSA which is used in the 1997-2005 files.
** Variable definition changed starting with the 2006 NHIS.

TABLE 3
Variables included in imputation of person-level
earnings proportions (Step 4).

Variable Name	Label and Code values
SEX	Sex 1 = male 2 = female
AGEGROUP	Recoded age group 0 = under 18 years old 1 = 18 - 24 years old 2 = 25 - 44years old 3 = 45 - 64 years old 4 = 65+ years old
ORIGIN	Ethnic origin 1 = Hispanic 2 = Non-Hispanic
**RACEREC	Race recode (year 1997-2005) 1 = white 2 = black 3 = other Race recode (year 2006+) 1 = white 2 = black 3 = Asian 4 = other
MARRY	Marital status recode 1 = married with or without spouse in HH, or living w/ partner 2 = divorced, widowed, separated 3 = never married 4 = 14 or fewer years old
FM_SIZER	Family size recode 1 = 1 person family 2 = 2 person family 3 = 3 person family 4 = 4+ person family
URB_RRL	Urban/Rural 1 = Urban 2 = Rural
*MSA	MSA/non-MSA residence (From 2007, MSA is created from recoding CBSASTAT and CBSASIZE) 1 = in MSA; in central city 2 = in MSA, not in central city 3 = not in MSA
WTFA	Final person weight
*STRATPSU	Stratum by PSU combination (year 1997 – 2005) Stratum and PSU combination recoded based on STRAT_I, STRAT_D, PSU_I from the NHIS inhouse data file.(year 2006+)
PLAADL	Needs help with ADL (age >= 3) 1 = yes 2 = no
PLAIADL	Needs help with chores, shopping, etc. (age >= 5) 1 = yes 2 = no
PLAWKNOW	Unable to work due to health problem (age >= 18) 1 = yes 2 = no
PLAWKLIM	Limited kind/amt of work due to health problem (age >= 18) 1 = yes 2 = no 3 = unable to work (PLAWKNW = 1)
PLAWALK	Has difficulty walking without equipment 1 = yes 2 = no
PLAREMEM	Limited by difficulty remembering 1 = yes 2 = no
PLIMANY	Limited in any other way 1 = yes 2 = no 3 = limitation previously mentioned

LIM_ACT	Limited in any way (at least mentioned one limitation) 1 = at least 1 limitation 2 = no limitation 3 = under 18 years old . = don't know/refused/not ascertained
LA_GP01	Vision or hearing problem causes limitation (18+ with at least 1 limitation) 1 = mentioned 2 = not mentioned 3 = no limitation or under 18 years old
LA_GP02	Arthritis/rheumatism, back/neck, or muscular-skeletal problem causes limitation (18+ with at least 1 limitation) 1 = mentioned 2 = not mentioned 3 = no limitation or under 18 years old
LA_GP03	Fracture/bone/joint injury, other injury, or missing limb/finger causes limitation (18+ with at least 1 limitation) 1 = mentioned 2 = not mentioned 3 = no limitation or under 18 years old
LA_GP04	Heart, stroke, hypertension, or circulatory problem causes limitation (18+ with at least 1 limitation) 1 = mentioned 2 = not mentioned 3 = no limitation or under 18 years old
LA_GP05	Diabetes or endocrine problem causes limitation (18+ with at least 1 limitation) 1 = mentioned 2 = not mentioned 3 = no limitation or under 18 years old
LA_GP06	Lung/breath problem causes limitation (18+ with at least 1 limitation) 1 = mentioned 2 = not mentioned 3 = no limitation or under 18 years old
LA_GP07	Senility or nervous system condition causes limitation (18+ with at least 1 limitation) 1 = mentioned 2 = not mentioned 3 = no limitation or under 18 years old
LA_GP08	Depression/anxiety/emotion, alcohol, drug, or other mental problem causes limitation (18+ with at least 1 limitation) 1 = mentioned 2 = not mentioned 3 = no limitation or under 18 years old
LA_GP09	Other problem causes limitation (18+ with at least 1 limitation) 1 = mentioned 2 = not mentioned 3 = no limitation or under 18 years old
PHSTAT	Recoded health status 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor
PDMED12M	Delayed medical care due to cost in the past 12 months 1 = yes 2 = no
PNMED12M	Did not get medical care due to cost in the past 12 months 1 = yes 2 = no
PHOSPYR	In a hospital overnight in the past 12 months 1 = yes 2 = no
P10DVYR	Received health care from doctor 10+ times in the past 12 months 1 = yes 2 = no
M_CARE	Medicare coverage recode 1 = yes (covered, with or without information) 2 = no (not covered)
M_CAID	Medicaid coverage recode 1 = yes 2 = no

MILITRY	Military coverage recode 1 = yes (military/VA/CHAMPUS/TRICARE/don't know type) 2 = no
PRIVATW	Private insurance coverage recode; at least 1 plan is paid by employer 1 = at least 1 private plan was obtained through employer 2 = all the private plans were not obtained through employer
PRIVATS	Private insurance coverage recode; at least 1 plan is paid by self 1 = at least 1 private plan was obtained through self 2 = all the private plans were not obtained through self
USBRTH	Born in US 1 = yes 2 = no
EDUCR	Education recode 1 = high school or less 2 = HS grad or equiv 3 = some college 4 = college graduate 5 = more than college 6 = professional degree
EMPSTAT	Last year's employment status recode 1 = 18+, worked for pay last year 2 = 18+, not worked for pay last year 3 = under 18 years old
WRKHRS	Hours worked per week 1 - 95 = worked 1 to 95+ hours
SSUFINWL	Logarithm of mean family income in SSU 0 - 13.816 = logarithm of mean (re-calculated after imputation)
SSUSDNWL	Logarithm of (standard deviation + 1) of family income in SSU 0 - 13.406 = logarithm of standard deviation (re-calculated after imputation)
PSAL	Person received income from wage/salary 1 = yes 2 = no 3 = under 18
PSEINC	Person received income from self-employment 1 = yes 2 = no 3 = under 18
PSSRR	Person received income from Social Security 1 = yes 2 = no
PPENS	Person received income from other pension 1 = yes 2 = no
PSSI	Person received income from Supplement Security Income (SSI) 1 = yes 2 = no
PSSRRDB	Person received income from Social Security Disability Insurance (SSDI) 1 = yes 2 = no 3 = ineligible due to age (65+) or received social security
PTANF	Person received income from welfare/AFDC 1 = yes 2 = no
PINTRST	Person received income from interest 1 = yes 2 = no
PDIVD	Person received income from dividend 1 = yes 2 = no
PCHLDSP	Person received child support 1 = yes 2 = no
PINCOT	Person received income from other sources 1 = yes 2 = no
HOUSER	House ownership recode 1 = owned or being bought 2 = rented or other arrangement

PSSID	Receive SSI due to a disability 1 = yes 2 = no 3 = do not receive SSI
PSSAPL	Person not receiving SSI and has ever applied for SSI 1 = yes 2 = no 3 = receive SSI
PSSRRD	Receive SSDI due to a disability 1 = yes 2 = no 3 = AGE > 65 or do not receive SSDI
PSDAPL	Person not receiving SSDI and has ever applied for SSDI 1 = yes 2 = no 3 = receive SSDI
PFSTP	Person was authorized to receive food stamps 1 = yes 2 = no
M_ERNAGE	Average age of male earners in a family 0 – 95 = age of male earners
F_ERNAGE	Average age of female earners in a family 0 – 95 = age of female earners
PHONSRV	Phone service in the household 1 = yes (has phone service) 2 = no (no phone service or working phone)
INTV_LN	Interview language 1 = English only 2 = Not English only
FM_EARN	Total number of earners in the family 0 – 9 = number of earners
P_F_EARN	Proportion of female earners to the total family earners 0 – 1 = proportion of female earners
ADULT	Total number of adults in a family
CHILD	Total number of children in a family
FAMINB	Cube root of total family income 0 – 101.283 = cube root of family income
FM_INCB	Cube root of total family earnings 0 – 101.283 = cube root of family earnings
LOG_PFM	Logit transformation of proportion of individual earnings to the total family earnings -12.78 – 12.33 = logit transformation of proportion of individual earnings

* Variable was recoded using CBSASTAT and CBSASIZE, so it is comparable to the definition of MSA which is used in the 1997-2005 files.
** Variable definition changed starting with the 2006 NHIS.

APPENDIX C. SAS CODE FOR THE EXAMPLES IN SECTION 4

C.1 Code for Creating Completed Data Sets

```
title 'SAS code for creating completed data sets';

libname nhis 'd:\stat\sas\sas files\nhis2000';

proc format;
   value povertyi
      1 = '(1) <100%'
      2 = '(2) 100-199%'
      3 = '(3) 200-399%'
      4 = '(4) 400%+'
      ;
   value agegr6r
      1 = '(1) <18'
      2 = '(2) 18-24'
      3 = '(3) 25-34'
      4 = '(4) 35-44'
      5 = '(5) 45-54'
      6 = '(6) 55-64'
      7 = '(7) 65+'
      ;
   value hprace
      1 = '(1) Hispanic'
      2 = '(2) Black'
      3 = '(3) Other'
      4 = '(4) White'
      ;
   value notcov
      1 = '(1) Uninsured'
      2 = '(2) Insured'
      ;
   value hstat
      1 = '(1) Excellent to good'
      2 = '(2) Fair/poor'
      ;
   value usborn
      1 = '(1) Not born in US'
      2 = '(2) Born in US'
      ;
   value sex
      1 = '(1) Male'
      2 = '(2) Female'
      ;
   value msar
      1 = '(1) MSA'
      2 = '(2) Not MSA'
      ;
   value regionr
      1 = '(1) Northeast'
      2 = '(2) South'
      3 = '(3) West'
      4 = '(4) Midwest'
      ;
run;
```

```
   /* EXTRACT VARIABLES FROM PERSON LEVEL FILE */

data temp1 (keep = hhx fmx px age_p sex origin_i
                   racrec_i notcov msasizep phstat usbrth_p
                   stratum psu wtfa region);
set nhis.personsx;
run;

proc sort;
by hhx fmx px;
run;


   /* MACRO TO REPEAT THE RECODES FOR ALL THE IMPUTED DATA SETS */

%macro subset;
%do impno=1 %to 5;


   /* EXTRACT RELEVANT INCOME VARIABLES FROM EACH IMPUTED FILE */

 data temp2;
    set nhis.incmimp&impno (drop=rectype srvy_yr);
 run;
 proc sort;
    by hhx fmx px;
 run;


   /* MERGE THE TWO TEMPORARY FILES, CREATE RECODES AND STORE THEM AS
      PERMANENT SAS DATA SETS */

 data nhis.anal&impno;
   merge temp1 temp2;
     by hhx fmx px;
```

```
*** recode age group ***;
        if (      age_p <  18) then  agegr6r = 1;
   else if (18 <= age_p <= 24) then  agegr6r = 2;
   else if (25 <= age_p <= 34) then  agegr6r = 3;
   else if (35 <= age_p <= 44) then  agegr6r = 4;
   else if (45 <= age_p <= 54) then  agegr6r = 5;
   else if (55 <= age_p <= 64) then  agegr6r = 6;
   else                              agegr6r = 7;


*** recode race/ethnicity groups ***;

        if (origin_i = 1) then  hprace = 1;
   else if (racrec_i = 2) then  hprace = 2;
   else if (racrec_i = 3) then  hprace = 3;
   else                         hprace = 4;


*** recode health insurance ****;

        if notcov in (7, 8, 9) then  notcov = .;


*** recode health status ***;

        if phstat in (1, 2, 3) then hstat = 1;
   else if phstat in (4, 5)    then hstat = 2;
   else                             hstat = .;

   *** create 0-1 health status variable for predicting fair-poor health
       using SUDAAN logistic regression procedure ***;

       hstat_sud=hstat-1;

   *** create 1-0 health status variable for predicting fair-poor health
       using IVEware logistic regression procedure ***;


       hstat_ive=2-hstat;


*** recode born in the US ***;

        if usbrth_p=1 then usborn = 2;
   else if usbrth_p=2 then usborn = 1;
   else                    usborn = .;


*** recode MSA ***;

   if msasizep=7 then msar = 2;
   else               msar = 1;


*** recode region ***;

        if region=1 then regionr = 1;
   else if region=3 then regionr = 2;
   else if region=4 then regionr = 3;
   else regionr = 4;
```

```
   *** recode poverty status ***;

           if rat_cati in (1, 2, 3)      then povertyi = 1;
      else if rat_cati in (4, 5, 6, 7)   then povertyi = 2;
      else if rat_cati in (8, 9, 10, 11) then povertyi = 3;
      else if rat_cati in (12, 13, 14)   then povertyi = 4;
      else                                    povertyi = .;


   *** labels ***;

   label povertyi = 'Poverty status';
   label agegr6r = 'Age groups (7)';
   label hprace =  'Race/ethnicity';
   label hstat = 'Health status';
   label notcov = 'Health insurance coverage';
   label usborn = 'Born in US';
   label regionr = 'Region';
   label msar = 'MSA';

   *** formats ***;

   format povertyi povertyi.
          agegr6r  agegr6r.
          hprace   hprace.
          hstat    hstat.

          notcov   notcov.
          usborn   usborn.
          sex      sex.
          regionr  regionr.
          msar     msar.;

run;
%end;
%mend;

%subset;


%macro sortall;
%do impno=1 %to 5;
proc sort data=nhis.anal&impno;
by stratum psu;
run;
%end;
%mend;

%sortall;
```

C.2 Code for Use with SAS-Callable SUDAAN 9.0 with a Built-In Option for Multiple Imputation

```
   /* SAS-CALLABLE SUDAAN CODE TO ANALYZE ACROSS ALL 5 IMPUTATIONS
      USING SUDAAN v9.0 WITH MI_COUNT OPTION */


     /* Example 1: Crosstab */

title "Example 1: Crosstab";
proc crosstab data=nhis.anal1 filetype=sas design=wr mi_count=5;
rtitle "Example 1: Crosstab";
nest stratum psu / missunit;
weight wtfa;
subgroup povertyi notcov;
levels 4  2;
tables povertyi*notcov;
test chisq;
print nsum wsum sewgt rowper serow totper setot/style=nchs WSUMFMT=F13.2;
run;

     /* Example 2: Logistic regression */

title "Example 2: Logistic regression";
proc rlogist data=nhis.anal1 filetype=sas design=wr mi_count=5;
rtitle "Example 2: Logistic regression";
nest stratum psu / missunit;
weight wtfa;
subgroup  povertyi agegr6r hprace usborn sex regionr msar;
levels      4         7      4      2     2    4      2;
reflevel povertyi=4 agegr6r=7 hprace=4 usborn=2 sex=2 regionr=4 msar=2;
model hstat_sud = povertyi agegr6r hprace usborn sex regionr msar;
print /ddfbetafmt=f16.3;
run;
```

C.3 Code for Use with SAS-Callable SUDAAN Version 7 or Higher without a Built-In Option for Multiple Imputation

```
/* SAS-CALLABLE SUDAAN MACRO TO PERFORM LOGISTIC REGRESSION ANALYSIS
   USING VERSIONS OF SUDAAN (v7.0 OR HIGHER) WITHOUT BUILT-IN
   MULTIPLE IMPUTATION OPTION */


      /* Example 2: Logistic regression */

%macro milogit;


   /* STEP 1: ANALYZE EACH DATA SET SEPARATELY */

%do impno=1 %to 5;

data temp;
set nhis.anal&impno;
run;


   /* PERFORM LOGISTIC REGRESSION ON EACH DATA SET AND STORE THE ESTIMATES */

proc rlogist data=temp filetype=sas design=wr noprint;
rtitle "Example 2: Logistic regression";
nest stratum psu / missunit;
weight wtfa;
subgroup  povertyi agegr6r hprace usborn sex regionr msar;
levels      4         7      4      2     2    4      2;
reflevel povertyi=4 agegr6r=7 hprace=4 usborn=2 sex=2 regionr=4 msar=2;
model hstat_sud = povertyi agegr6r hprace usborn sex regionr msar;
output/betas=default filename=est filetype=sas replace;
run;


   /* STEP 2: STORE THE RELEVANT OUTPUT FROM EACH DATA SET */

data est&impno;
set est;
beta&impno=beta;
sebeta&impno=sebeta;
keep modelrhs beta&impno sebeta&impno;
run;


   /* SORT THE OUTPUT FOR LATER MERGING */

proc sort data=est&impno;
by modelrhs;
run;
%end;
%mend;

%milogit;
```

```
   /* STEP 3: COMBINE THE RESULTS.

      THIS EXAMPLE ILLUSTRATES BOTH RUBIN-SCHENKER AND
      BARNARD-RUBIN DEGREES OF FREEDOM.
      IN PRACTICE ONE WOULD COMPUTE ONLY ONE OF THEM */

   /* MERGE THE FIVE SETS OF ESTIMATES */

data all;
merge est1 est2 est3 est4 est5;
by modelrhs;

   /* COMPUTE THE COMBINED ESTIMATE, ITS VARIANCE,
      THE DEGREES OF FREEDOM AND CONFIDENCE INTERVAL */


   /* COMPUTE QBAR_M */

qbar_m=mean(beta1,beta2,beta3,beta4,beta5);


   /* COMPUTE UBAR_M */

ubar_m=mean(sebeta1**2,sebeta2**2,sebeta3**2,sebeta4**2,sebeta5**2);


   /* COMPUTE B_M */

b_m= var(beta1,beta2,beta3,beta4,beta5);


   /* COMPUTE TOTAL ESTIMATED VARIANCE, T_M, AND ESTIMATED STANDARD ERROR,
      SE_MI */

t_m=ubar_m+(1+1/5)*b_m;
se_mi=sqrt(t_m);


   /* RUBIN-SCHENKER DEGREES OF FREEDOM, NU */

if t_m ne 0 then gammah_m=(1+1/5)*b_m/t_m;  /* t_m=0 for reference groups */
if gammah_m ne 0 then nu=(5-1)/gammah_m**2;
else nu=100000;

   /* BARNARD-RUBIN DEGREES OF FREEDOM, NUPRIME */

df=339;
k=(df*(df+1)/(df+3))*(1-gammah_m);
nuprime=1/(1/nu+1/k);
```

```
   /* CALCULATE THE T-RATIO AND THE P-VALUES */

if se_mi ne 0 then tratio=qbar_m/se_mi;  /* se_mi=0 for reference groups */
pval_rs=2*(1-probt(abs(tratio),nu));
pval_br=2*(1-probt(abs(tratio),nuprime));


   /* CALCULATE CONFIDENCE INTERVAL FOR ODDS RATIO USING RUBIN-SCHENKER */

or_mi=exp(qbar_m);
ll95_rs=exp(qbar_m-tinv(0.975,nu)*se_mi);
ul95_rs=exp(qbar_m+tinv(0.975,nu)*se_mi);


   /* CALCULATE CONFIDENCE INTERVAL USING BARNARD-RUBIN */

ll95_br=exp(qbar_m-tinv(0.975,nuprime)*se_mi);
ul95_br=exp(qbar_m+tinv(0.975,nuprime)*se_mi);
run;


   /* PRINT OUT THE RESULTS */

proc print data=all;
title1 "Multiple Imputation Analysis for Logistic Regression Model";
title2 "Estimates, Standard errors, t ratios, and p-values";
var modelrhs qbar_m se_mi tratio pval_rs pval_br;
run;

proc print data=all;
where modelrhs>1;
title2 "Odds ratios and confidence intervals";
var modelrhs or_mi ll95_br ul95_br ll95_rs ul95_rs;
run;
```

C.4 Code for Use with SAS-Callable IVEware

```
   /* SAS-CALLABLE IVEWARE CODE FOR EXAMPLES IN SECTION 4 */


   /* Example 1: Crosstab */

title "IVEware Example 1: Crosstab";
%describe(setup=new,name=dessetup,dir=d:\stat\sas\sas files\nhis2000);
datain nhis.anal1 nhis.anal2 nhis.anal3 nhis.anal4 nhis.anal5;
stratum stratum;
cluster psu;
weight wtfa;
table povertyi*notcov;
run;


   /* Example 2: Logistic regression */

title "IVEware Example 2: Logistic regression";
%regress(setup=new,name=regsetup,dir=d:\stat\sas\sas files\nhis2000);
datain nhis.anal1 nhis.anal2 nhis.anal3 nhis.anal4 nhis.anal5;
stratum stratum;
cluster psu;
weight wtfa;
link logistic;
dependent hstat_ive;
predictor   povertyi agegr6r hprace usborn sex regionr msar;
categorical povertyi agegr6r hprace usborn sex regionr msar hstat_ive;
run;
```

APPENDIX D. SAMPLE OUTPUT FROM SAS-CALLABLE SUDAAN VERSION 9.0 WITH A BUILT-IN OPTION FOR MULTIPLE IMPUTATION

```
Example 1: Crosstab


                              S U D A A N
           Software for the Statistical Analysis of Correlated Data
          Copyright        Research Triangle Institute          July 2004
                                Release 9.0


Processing data for set 1 of imputed variables:

Number of observations read    : 100355    Weighted count :274018975
Number of observations skipped :    263
(WEIGHT variable nonpositive)
Denominator degrees of freedom :    339


Processing data for set 2 of imputed variables:

Number of observations read    : 100355    Weighted count :274018975
Number of observations skipped :    263
(WEIGHT variable nonpositive)
Denominator degrees of freedom :    339


Processing data for set 3 of imputed variables:

Number of observations read    : 100355    Weighted count :274018975
Number of observations skipped :    263
(WEIGHT variable nonpositive)
Denominator degrees of freedom :    339


Processing data for set 4 of imputed variables:

Number of observations read    : 100355    Weighted count :274018975
Number of observations skipped :    263
(WEIGHT variable nonpositive)
Denominator degrees of freedom :    339


Processing data for set 5 of imputed variables:

Number of observations read    : 100355    Weighted count :274018975
Number of observations skipped :    263
```

```
(WEIGHT variable nonpositive)
Denominator degrees of freedom :    339
```

Variance Estimation Method: Taylor Series (WR) Using Multiply Imputed Data
Results for Summary Over All Imputations
by: Poverty status, Health insurance coverage.

Poverty status / Health insurance coverage	Sample Size	Weighted Size	SE Weighted	Row Percent	SE Row Percent	Tot Percent	SE Tot Percent
Total							
Total	99272	271135638.00	2229205.41	100.00	0.00	100.00	0.00
(1) Uninsured	17500	40484575.00	720600.13	14.93	0.22	14.93	0.22
(2) Insured	81772	230651063.00	1917400.17	85.07	0.22	85.07	0.22
(1) <100%							
Total	15622	34492808.80	842954.98	100.00	0.00	12.72	0.29
(1) Uninsured	5253	10522088.00	351141.75	30.51	0.71	3.88	0.12
(2) Insured	10369	23970720.80	639643.61	69.49	0.71	8.84	0.22
(2) 100-199%							
Total	20728	51627593.80	1006238.61	100.00	0.00	19.04	0.33
(1) Uninsured	5887	13168341.20	402170.15	25.51	0.56	4.86	0.14
(2) Insured	14841	38459252.60	783658.23	74.49	0.56	14.18	0.26
(3) 200-399%							
Total	30607	85884217.20	1199479.08	100.00	0.00	31.68	0.35
(1) Uninsured	4461	11384821.60	397324.06	13.26	0.38	4.20	0.14
(2) Insured	26146	74499395.60	1015930.76	86.74	0.38	27.48	0.30
(4) 400%+							
Total	32315	99131018.20	1393716.99	100.00	0.00	36.56	0.44
(1) Uninsured	1899	5409324.20	217723.08	5.46	0.21	2.00	0.08
(2) Insured	30416	93721694.00	1347608.79	94.54	0.21	34.57	0.43

Example 2: Logistic regression

```
                              S U D A A N
           Software for the Statistical Analysis of Correlated Data
          Copyright        Research Triangle Institute          July 2004
                                 Release 9.0


Processing data for set 1 of imputed variables:

Processing data for set 2 of imputed variables:

Processing data for set 3 of imputed variables:

Processing data for set 4 of imputed variables:

Processing data for set 5 of imputed variables:
```

```
Processing data for set 1 of imputed variables:

Number of zero responses     : 90708
Number of non-zero responses :  9200

Independence parameters have converged in 7 iterations

Number of observations read       : 100355    Weighted count:274018975
Number of observations skipped    :    263
(WEIGHT variable nonpositive)
Observations used in the analysis :  99908    Weighted count:272718013
Denominator degrees of freedom    :    339


Maximum number of estimable parameters for the model is 19

File NHIS.ANAL1 contains  678 Clusters
 678 clusters were used to fit the model
Maximum cluster size is 389 records
Minimum cluster size is  19 records


Sample and Population Counts for Response Variable HSTAT_SUD
  0:  Sample Count    90708    Population Count 248661451
  1:  Sample Count     9200    Population Count  24056562

R-Square for dependent variable HSTAT_SUD (Cox & Snell, 1989): 0.105144
```

```
Processing data for set 2 of imputed variables:

Number of zero responses     : 90708
Number of non-zero responses :  9200

Independence parameters have converged in 7 iterations

Number of observations read       : 100355    Weighted count:274018975
Number of observations skipped    :    263
(WEIGHT variable nonpositive)
Observations used in the analysis :  99908    Weighted count:272718013
Denominator degrees of freedom    :    339


Maximum number of estimable parameters for the model is 19

File NHIS.ANAL2 contains  678 Clusters
 678 clusters were used to fit the model
Maximum cluster size is 389 records
Minimum cluster size is  19 records


Sample and Population Counts for Response Variable HSTAT_SUD
  0:  Sample Count    90708    Population Count 248661451
  1:  Sample Count     9200    Population Count  24056562

R-Square for dependent variable HSTAT_SUD (Cox & Snell, 1989): 0.105782
```

```
Processing data for set 3 of imputed variables:

Number of zero responses     : 90708
Number of non-zero responses :  9200

Independence parameters have converged in 7 iterations

Number of observations read       : 100355    Weighted count:274018975
Number of observations skipped    :    263
(WEIGHT variable nonpositive)
Observations used in the analysis :  99908    Weighted count:272718013
Denominator degrees of freedom    :    339


Maximum number of estimable parameters for the model is 19

File NHIS.ANAL3 contains  678 Clusters
 678 clusters were used to fit the model
Maximum cluster size is 389 records
Minimum cluster size is  19 records


Sample and Population Counts for Response Variable HSTAT_SUD
  0:  Sample Count    90708    Population Count 248661451
  1:  Sample Count     9200    Population Count  24056562

R-Square for dependent variable HSTAT_SUD (Cox & Snell, 1989): 0.105506
```

```
Processing data for set 4 of imputed variables:

Number of zero responses     : 90708
Number of non-zero responses :  9200

Independence parameters have converged in 7 iterations

Number of observations read       : 100355    Weighted count:274018975
Number of observations skipped    :    263
(WEIGHT variable nonpositive)
Observations used in the analysis :  99908    Weighted count:272718013
Denominator degrees of freedom    :    339


Maximum number of estimable parameters for the model is 19

File NHIS.ANAL4 contains  678 Clusters
 678 clusters were used to fit the model
Maximum cluster size is 389 records
Minimum cluster size is  19 records


Sample and Population Counts for Response Variable HSTAT_SUD
  0:  Sample Count    90708    Population Count 248661451
  1:  Sample Count     9200    Population Count  24056562

R-Square for dependent variable HSTAT_SUD (Cox & Snell, 1989): 0.105770
```

```
Processing data for set 5 of imputed variables:

Number of zero responses     : 90708
Number of non-zero responses :  9200

Independence parameters have converged in 7 iterations

Number of observations read       : 100355    Weighted count:274018975
Number of observations skipped    :    263
(WEIGHT variable nonpositive)
Observations used in the analysis :  99908    Weighted count:272718013
Denominator degrees of freedom    :    339


Maximum number of estimable parameters for the model is 19

File NHIS.ANAL5 contains  678 Clusters
 678 clusters were used to fit the model
Maximum cluster size is 389 records
Minimum cluster size is  19 records


Sample and Population Counts for Response Variable HSTAT_SUD
  0:  Sample Count    90708    Population Count 248661451
  1:  Sample Count     9200    Population Count  24056562

R-Square for dependent variable HSTAT_SUD (Cox & Snell, 1989): 0.105145

Overall degrees of freedom (Rubin): 30.96

-2 * Normalized Log-Likelihood with Intercepts Only : 59620.63
-2 * Normalized Log-Likelihood Full Model           : 48521.49
Approximate Chi-Square (-2 * Log-L Ratio)           : 11099.14
Degrees of Freedom                                  :       18

Note: The approximate Chi-Square is not adjusted for clustering.
      Refer to hypothesis test table for adjusted test.
```

Variance Estimation Method: Taylor Series (WR) Using Multiply Imputed Data
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Logit
Response variable HSTAT_SUD: HSTAT_SUD
Results for Summary Over All Imputations
by: Independent Variables and Effects.

Independent Variables and Effects	Beta Coeff.	SE Beta	Lower 95% Limit Beta	Upper 95% Limit Beta	T-Test B=0	P-value T-Test B=0	DDF Beta
Intercept	-1.96	0.07	-2.09	-1.84	-30.00	0.0000	164.500
Poverty status					.		
(1) <100%	1.83	0.05	1.72	1.94	33.82	0.0000	109.630
(2) 100-199%	1.38	0.05	1.28	1.48	26.92	0.0000	133.114
(3) 200-399%	0.79	0.06	0.67	0.90	14.20	0.0000	30.957
(4) 400%+	0.00	0.00	0.00	0.00		.	339.000
Age groups (7)					.		
(1) <18	-3.33	0.06	-3.45	-3.21	-54.59	0.0000	333.220
(2) 18-24	-2.64	0.07	-2.77	-2.50	-38.07	0.0000	333.215
(3) 25-34	-2.11	0.06	-2.22	-2.01	-38.29	0.0000	331.255
(4) 35-44	-1.50	0.04	-1.59	-1.42	-34.63	0.0000	321.390
(5) 45-54	-0.74	0.04	-0.83	-0.65	-16.76	0.0000	289.338
(6) 55-64	-0.32	0.04	-0.40	-0.23	-7.41	0.0000	282.767
(7) 65+	0.00	0.00	0.00	0.00		.	339.000
Race/ethnicity					.		
(1) Hispanic	0.37	0.05	0.27	0.48	6.91	0.0000	316.355
(2) Black	0.44	0.05	0.35	0.54	9.37	0.0000	304.406
(3) Other	0.25	0.09	0.08	0.42	2.96	0.0033	335.040
(4) White	0.00	0.00	0.00	0.00		.	339.000
Born in US					.		
(1) Not born in US	-0.30	0.06	-0.41	-0.19	-5.30	0.0000	333.009
(2) Born in US	0.00	0.00	0.00	0.00		.	339.000
SEX					.		
(1) male	0.02	0.02	-0.03	0.07	0.85	0.3956	335.232
(2) female	0.00	0.00	0.00	0.00		.	339.000
Region					.		
(1) Northeast	-0.01	0.06	-0.12	0.10	-0.20	0.8427	335.694
(2) South	0.20	0.05	0.11	0.29	4.41	0.0000	334.248
(3) West	0.07	0.05	-0.02	0.17	1.47	0.1426	319.804
(4) Midwest	0.00	0.00	0.00	0.00		.	339.000
MSA					.		
(1) MSA	-0.18	0.04	-0.26	-0.10	-4.19	0.0000	334.995
(2) Not MSA	0.00	0.00	0.00	0.00		.	339.000

Variance Estimation Method: Taylor Series (WR) Using Multiply Imputed Data
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Logit
Response variable HSTAT_SUD: HSTAT_SUD
Results for Summary Over All Imputations
by: Contrast.

Contrast	Degrees of Freedom	Wald F	P-value Wald F
OVERALL MODEL	19	977.44	0.0000
MODEL MINUS INTERCEPT	18	354.66	0.0000
INTERCEPT	.	446.44	. 0.0000
POVERTYI	3	736.79	0.0000
AGEGR6R	6	34.44	0.0000
HPRACE	3	28.09	0.0000
USBORN	1	0.72	0.4015
SEX	1	9.28	0.0002
REGIONR	3	17.54	0.0002
MSAR	1		

Variance Estimation Method: Taylor Series (WR) Using Multiply Imputed Data
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Logit
Response variable HSTAT_SUD: HSTAT_SUD
Results for Summary Over All Imputations
by: Independent Variables and Effects.

Independent Variables and Effects	Odds Ratio	Lower 95% OR Limit	Upper 95% OR Limit
Intercept	0.14	0.12	0.16
Poverty status			
(1) <100%	6.25	5.61	6.95
(2) 100-199%	3.98	3.60	4.41
(3) 200-399%	2.19	1.96	2.45
(4) 400%+	1.00	1.00	1.00
Age groups (7)			
(1) <18	0.04	0.03	0.04
(2) 18-24	0.07	0.06	0.08
(3) 25-34	0.12	0.11	0.13
(4) 35-44	0.22	0.20	0.24
(5) 45-54	0.48	0.44	0.52
(6) 55-64	0.73	0.67	0.79
(7) 65+	1.00	1.00	1.00
Race/ethnicity			
(1) Hispanic	1.45	1.31	1.61
(2) Black	1.56	1.42	1.71
(3) Other	1.29	1.09	1.52
(4) White	1.00	1.00	1.00
Born in US			
(1) Not born in US	0.74	0.67	0.83
(2) Born in US	1.00	1.00	1.00
SEX			
(1) male	1.02	0.97	1.07
(2) female	1.00	1.00	1.00
Region			
(1) Northeast	0.99	0.89	1.10
(2) South	1.22	1.12	1.34
(3) West	1.07	0.98	1.18
(4) Midwest	1.00	1.00	1.00
MSA			
(1) MSA	0.84	0.77	0.91
(2) Not MSA	1.00	1.00	1.00

APPENDIX E. SAMPLE OUTPUT FROM SAS COMMANDS FOR USE WITH SUDAAN VERSION 7 OR HIGHER WITHOUT A BUILT-IN OPTION FOR MULTIPLE IMPUTATION

```
Example 2: Logistic regression


                              S U D A A N
           Software for the Statistical Analysis of Correlated Data
          Copyright        Research Triangle Institute          July 2004
                                Release 9.0


Number of zero responses     : 90708
Number of non-zero responses :  9200

Independence parameters have converged in 7 iterations

Number of observations read       : 100355    Weighted count:274018975
Number of observations skipped    :    263
(WEIGHT variable nonpositive)
Observations used in the analysis :  99908    Weighted count:272718013
Denominator degrees of freedom    :    339


Maximum number of estimable parameters for the model is 19

File TEMP contains  678 Clusters
 678 clusters were used to fit the model
Maximum cluster size is 389 records
Minimum cluster size is  19 records


Sample and Population Counts for Response Variable HSTAT_SUD
  0:  Sample Count    90708    Population Count 248661451
  1:  Sample Count     9200    Population Count  24056562

R-Square for dependent variable HSTAT_SUD (Cox & Snell, 1989): 0.105144

-2 * Normalized Log-Likelihood with Intercepts Only : 59620.63
-2 * Normalized Log-Likelihood Full Model           : 48521.57
Approximate Chi-Square (-2 * Log-L Ratio)           : 11099.05
Degrees of Freedom                                  :       18

Note: The approximate Chi-Square is not adjusted for clustering.
      Refer to hypothesis test table for adjusted test.
```

```
                              S U D A A N
           Software for the Statistical Analysis of Correlated Data
          Copyright        Research Triangle Institute          July 2004
                                 Release 9.0


Number of zero responses     : 90708
Number of non-zero responses :  9200

Independence parameters have converged in 7 iterations

Number of observations read       : 100355    Weighted count:274018975
Number of observations skipped    :    263
(WEIGHT variable nonpositive)
Observations used in the analysis :  99908    Weighted count:272718013
Denominator degrees of freedom    :    339


Maximum number of estimable parameters for the model is 19

File TEMP contains  678 Clusters
 678 clusters were used to fit the model
Maximum cluster size is 389 records
Minimum cluster size is  19 records


Sample and Population Counts for Response Variable HSTAT_SUD
  0:  Sample Count    90708    Population Count 248661451
  1:  Sample Count     9200    Population Count  24056562

R-Square for dependent variable HSTAT_SUD (Cox & Snell, 1989): 0.105782

-2 * Normalized Log-Likelihood with Intercepts Only : 59620.63
-2 * Normalized Log-Likelihood Full Model           : 48450.36
Approximate Chi-Square (-2 * Log-L Ratio)           : 11170.27
Degrees of Freedom                                  :       18

Note: The approximate Chi-Square is not adjusted for clustering.
      Refer to hypothesis test table for adjusted test.
```

```
                              S U D A A N
           Software for the Statistical Analysis of Correlated Data
          Copyright        Research Triangle Institute          July 2004
                                 Release 9.0


Number of zero responses     : 90708
Number of non-zero responses :  9200

Independence parameters have converged in 7 iterations

Number of observations read       : 100355    Weighted count:274018975
Number of observations skipped    :    263
(WEIGHT variable nonpositive)
Observations used in the analysis :  99908    Weighted count:272718013
Denominator degrees of freedom    :    339


Maximum number of estimable parameters for the model is 19

File TEMP contains  678 Clusters
 678 clusters were used to fit the model
Maximum cluster size is 389 records
Minimum cluster size is  19 records


Sample and Population Counts for Response Variable HSTAT_SUD
  0:  Sample Count    90708    Population Count 248661451
  1:  Sample Count     9200    Population Count  24056562

R-Square for dependent variable HSTAT_SUD (Cox & Snell, 1989): 0.105506

-2 * Normalized Log-Likelihood with Intercepts Only : 59620.63
-2 * Normalized Log-Likelihood Full Model           : 48481.20
Approximate Chi-Square (-2 * Log-L Ratio)           : 11139.43
Degrees of Freedom                                  :       18

Note: The approximate Chi-Square is not adjusted for clustering.
      Refer to hypothesis test table for adjusted test.
```

```
                              S U D A A N
           Software for the Statistical Analysis of Correlated Data
          Copyright        Research Triangle Institute          July 2004
                                 Release 9.0


Number of zero responses     : 90708
Number of non-zero responses :  9200

Independence parameters have converged in 7 iterations

Number of observations read       : 100355    Weighted count:274018975
Number of observations skipped    :    263
(WEIGHT variable nonpositive)
Observations used in the analysis :  99908    Weighted count:272718013
Denominator degrees of freedom    :    339


Maximum number of estimable parameters for the model is 19

File TEMP contains  678 Clusters
 678 clusters were used to fit the model
Maximum cluster size is 389 records
Minimum cluster size is  19 records


Sample and Population Counts for Response Variable HSTAT_SUD
  0:  Sample Count    90708    Population Count 248661451
  1:  Sample Count     9200    Population Count  24056562

R-Square for dependent variable HSTAT_SUD (Cox & Snell, 1989): 0.105770

-2 * Normalized Log-Likelihood with Intercepts Only : 59620.63
-2 * Normalized Log-Likelihood Full Model           : 48451.73
Approximate Chi-Square (-2 * Log-L Ratio)           : 11168.90
Degrees of Freedom                                  :       18

Note: The approximate Chi-Square is not adjusted for clustering.
      Refer to hypothesis test table for adjusted test.
```

```
                              S U D A A N
           Software for the Statistical Analysis of Correlated Data
          Copyright        Research Triangle Institute          July 2004
                                 Release 9.0


Number of zero responses     : 90708
Number of non-zero responses :  9200

Independence parameters have converged in 7 iterations

Number of observations read       : 100355    Weighted count:274018975
Number of observations skipped    :    263
(WEIGHT variable nonpositive)
Observations used in the analysis :  99908    Weighted count:272718013
Denominator degrees of freedom    :    339


Maximum number of estimable parameters for the model is 19

File TEMP contains  678 Clusters
 678 clusters were used to fit the model
Maximum cluster size is 389 records
Minimum cluster size is  19 records


Sample and Population Counts for Response Variable HSTAT_SUD
  0:  Sample Count    90708    Population Count 248661451
  1:  Sample Count     9200    Population Count  24056562

R-Square for dependent variable HSTAT_SUD (Cox & Snell, 1989): 0.105145

-2 * Normalized Log-Likelihood with Intercepts Only : 59620.63
-2 * Normalized Log-Likelihood Full Model           : 48521.49
Approximate Chi-Square (-2 * Log-L Ratio)           : 11099.14
Degrees of Freedom                                  :       18

Note: The approximate Chi-Square is not adjusted for clustering.
      Refer to hypothesis test table for adjusted test.
```

Multiple Imputation Analysis for Logistic Regression Model
Estimates, Standard errors, t ratios, and p-values

Obs	MODELRHS	qbar_m	se_mi	tratio	pval_rs	pval_br
1	1	-1.96467	0.065479	-30.0045	0.00000	0.00000
2	2	1.83210	0.054165	33.8245	0.00000	0.00000
3	3	1.38247	0.051350	26.9227	0.00000	0.00000
4	4	0.78531	0.055299	14.2013	0.00000	0.00000
5	5	0.00000	0.000000	.	.	.
6	6	-3.32730	0.060946	-54.5938	0.00000	0.00000
7	7	-2.63543	0.069228	-38.0690	0.00000	0.00000
8	8	-2.11470	0.055227	-38.2907	0.00000	0.00000
9	9	-1.50435	0.043436	-34.6338	0.00000	0.00000
10	10	-0.74168	0.044261	-16.7569	0.00000	0.00000
11	11	-0.31558	0.042604	-7.4072	0.00000	0.00000
12	12	0.00000	0.000000	.	.	.
13	13	0.37293	0.053986	6.9078	0.00000	0.00000
14	14	0.44469	0.047446	9.3725	0.00000	0.00000
15	15	0.25133	0.085009	2.9565	0.00311	0.00333
16	16	0.00000	0.000000	.	.	.
17	17	-0.29668	0.055972	-5.3004	0.00000	0.00000
18	18	0.00000	0.000000	.	.	.
19	19	0.02037	0.023943	0.8506	0.39499	0.39560
20	20	0.00000	0.000000	.	.	.
21	21	-0.01103	0.055538	-0.1987	0.84253	0.84265
22	22	0.20279	0.045940	4.4141	0.00001	0.00001
23	23	0.07180	0.048855	1.4697	0.14167	0.14262
24	24	0.00000	0.000000	.	.	.
25	25	-0.17951	0.042864	-4.1880	0.00003	0.00004
26	26	0.00000	0.000000	.	.	.

Multiple Imputation Analysis for Logistic Regression Model
Odds ratios and confidence intervals

Obs	MODELRHS	or_mi	ll95_br	ul95_br	ll95_rs	ul95_rs
1	2	6.24697	5.61112	6.95487	5.61368	6.95170
2	3	3.98474	3.59990	4.41073	3.60141	4.40887
3	4	2.19309	1.95917	2.45493	1.96040	2.45339
4	5	1.00000	.		.	
5	6	0.03589	0.03184	0.04046	0.03185	0.04044
6	7	0.07169	0.06256	0.08215	0.06259	0.08211
7	8	0.12067	0.10825	0.13452	0.10829	0.13446
8	9	0.22216	0.20396	0.24198	0.20403	0.24190
9	10	0.47631	0.43657	0.51967	0.43672	0.51950
10	11	0.72937	0.67070	0.79317	0.67091	0.79292
11	12	1.00000	.		.	
12	13	1.45198	1.30566	1.61469	1.30617	1.61406
13	14	1.56000	1.42095	1.71267	1.42144	1.71207
14	15	1.28574	1.08775	1.51976	1.08841	1.51884
15	16	1.00000	.		.	
16	17	0.74329	0.66579	0.82980	0.66606	0.82947
17	18	1.00000	.		.	
18	19	1.02057	0.97362	1.06979	0.97379	1.06961
19	20	1.00000	.		.	
20	21	0.98903	0.88667	1.10320	0.88702	1.10276
21	22	1.22481	1.11898	1.34065	1.11934	1.34021
22	23	1.07444	0.97598	1.18284	0.97632	1.18243
23	24	1.00000	.		.	
24	25	0.83568	0.76810	0.90919	0.76834	0.90892
25	26	1.00000	.	.	.	.
26						

APPENDIX F. SAMPLE OUTPUT FROM SAS-CALLABLE IVEWARE

Example 1: Crosstab

```
IVEware Setup Checker, Wed Nov 05 10:20:18 2003                          1

Setup listing:

datain nhis.anal1 nhis.anal2 nhis.anal3 nhis.anal4 nhis.anal5;
stratum stratum;
cluster psu;
weight wtfa;
table povertyi*notcov;
run;



IVEware Design-Based Descriptive Statistics Procedure, Wed Nov 05 10:21:35 2003
                                                                         1
Stratum variable:       STRATUM  Stratum for Variance
Cluster variable:       PSU  PSU
Weight variable:        WTFA  WEIGHT - FINAL ANNUAL

Analysis description:

         6    Variables
       339    Strata
       678    Secus

    Strata    Model
       339    Multiple PSU
         0    Paired Selection
         0    Successive Differences

    501775    Cases Read
```

All imputations

Problem 1

Degrees of freedom
15.9247

Factor Covariance of denominator
None 0.00822

Table povertyi	NOTCOV	Number of Cases	Sum of Weights	Weighted Proportion	Standard Error
1	1	5253	1.052209e+007	0.03881	0.00123
1	2	10369	2.397072e+007	0.08841	0.00224
1	1	5887	1.316834e+007	0.04857	0.00142
2	2	14841	3.845925e+007	0.14185	0.00258
2	1	4461	1.138482e+007	0.04199	0.00140
3	2	26146	7.44994e+007	0.27477	0.00301
3	1	1899	5409324	0.01995	0.00078
4	2	30416	9.372169e+007	0.34566	0.00438
4					

		Lower Bound	Upper Bound	T Test	Prob > |T|
	1	0.03621	0.04141	31.62450	0.00000
1	2	0.08366	0.09316	39.46063	0.00000
1	1	0.04556	0.05157	34.29998	0.00000
2	2	0.13638	0.14731	54.99516	0.00000
2	1	0.03903	0.04495	30.05772	0.00000
3	2	0.26838	0.28116	91.22695	0.00000
3	1	0.01829	0.02161	25.42298	0.00000
4	2	0.33637	0.35495	78.91966	0.00000
4					

		Unweighted Proportion	Bias	Design Effect
	1	0.05292	36.36357	1.89990
1	2	0.10445	18.14052	3.81654
1	1	0.05930	22.10615	1.83799
2	2	0.14950	5.39552	3.12661
2	1	0.04494	7.01542	1.49306
3	2	0.26338	-4.14554	3.14061
3	1	0.01913	-4.13716	2.52326
4	2	0.30639	-11.36046	6.15126
4				

Example 2: Logistic regression

```
IVEware Setup Checker, Wed Nov 05 10:21:36 2003                                 1

Setup listing:

datain nhis.anal1 nhis.anal2 nhis.anal3 nhis.anal4 nhis.anal5;
stratum stratum;
cluster psu;
weight wtfa;
link logistic;
dependent hstat_ive;
predictor   povertyi agegr6r hprace usborn sex regionr msar;
categorical povertyi agegr6r hprace usborn sex regionr msar hstat_ive;
run;



IVEware Jackknife Regression Procedure, Wed Nov 05 10:23:03 2003                1

Regression type:        Logistic
Dependent variable:     hstat_ive
Predictors:             povertyi  Poverty status
                        agegr6r  Age groups (7)
                        hprace  Race/ethnicity
                        usborn  Born in US
                        SEX
                        regionr  Region
                        msar  MSA
Cat. var. ref. codes:   SEX  2
                        agegr6r  7
                        hprace  4
                        hstat_ive  1
                        usborn  2
                        msar  2
                        regionr  4
                        povertyi  4
Stratum variable:       STRATUM  Stratum for Variance
Cluster variable:       PSU  PSU
Weight variable:        WTFA  WEIGHT - FINAL ANNUAL
```

All imputations

Valid cases 99908
Sum weights 272718013

Degr freedom 16.46093837

-2 LogLike 132349824.5

Variable	Estimate	Std Error	Wald test	Prob > Chi
Intercept	-1.9646743	0.0654952	899.83227	0.00000
povertyi.1	1.8320965	0.0541827	1143.33931	0.00000
povertyi.2	1.3824722	0.0513700	724.25942	0.00000
povertyi.3	0.7853111	0.0553181	201.53444	0.00000
agegr6r.1	-3.3272968	0.0609781	2977.39121	0.00000
agegr6r.2	-2.6354280	0.0692151	1449.77603	0.00000
agegr6r.3	-2.1147002	0.0552570	1464.61299	0.00000
agegr6r.4	-1.5043548	0.0434417	1199.18474	0.00000
agegr6r.5	-0.7416825	0.0442917	280.40819	0.00000
agegr6r.6	-0.3155753	0.0426026	54.86973	0.00000
hprace.1	0.3729262	0.0540541	47.59795	0.00000
hprace.2	0.4446873	0.0474451	87.84693	0.00000
hprace.3	0.2513328	0.0850379	8.73520	0.00312
usborn	-0.2966752	0.0560029	28.06349	0.00000
SEX	0.0203657	0.0239452	0.72337	0.39504
regionr.1	-0.0110329	0.0555605	0.03943	0.84260
regionr.2	0.2027855	0.0459095	19.51051	0.00001
regionr.3	0.0718029	0.0488095	2.16409	0.14127
msar	-0.1795138	0.0428418	17.55737	0.00003

Variable	Odds Ratio	95% Confidence Interval Lower	Upper
Intercept			
povertyi.1	6.2469697	5.5705566	7.0055174
povertyi.2	3.9847405	3.5744806	4.4420879
povertyi.3	2.1930892	1.9509339	2.4653014
agegr6r.1	0.0358900	0.0315472	0.0408306
agegr6r.2	0.0716883	0.0619254	0.0829903
agegr6r.3	0.1206695	0.1073593	0.1356298
agegr6r.4	0.2221606	0.2026574	0.2435406
agegr6r.5	0.4763118	0.4337167	0.5230902
agegr6r.6	0.7293691	0.6665208	0.7981436
hprace.1	1.4519772	1.2951114	1.6278429
hprace.2	1.5600022	1.4110533	1.7246741
hprace.3	1.2857379	1.0740855	1.5390971
usborn	0.7432854	0.6602567	0.8367551
SEX	1.0205745	0.9701732	1.0735942
regionr.1	0.9890277	0.8793709	1.1123586
regionr.2	1.2248097	1.1114690	1.3497082
regionr.3	1.0744435	0.9690552	1.1912932
msar	0.8356764	0.7632816	0.9149377

Variable	Design Effect	SRS Estimate	% Diff SRS v Est
Intercept	1.51026	-1.9287906	-1.82645
povertyi.1	1.49136	1.7978198	-1.87090
povertyi.2	1.38702	1.3683918	-1.01850
povertyi.3	1.32103	0.7724658	-1.63570
agegr6r.1	1.41998	-3.3338239	0.19617
agegr6r.2	1.20875	-2.6557082	0.76952
agegr6r.3	1.32854	-2.1683692	2.53790
agegr6r.4	1.24336	-1.5161737	0.78564
agegr6r.5	1.46884	-0.7409085	-0.10436
agegr6r.6	1.27075	-0.2915744	-7.60545
hprace.1	1.90751	0.3671522	-1.54831
hprace.2	1.84684	0.4718635	6.11132
hprace.3	1.49689	0.2193086	-12.74176
usborn	2.09977	-0.2861391	-3.55142
SEX	0.99894	-0.0045110	-122.14999
regionr.1	1.95092	0.0318232	-388.43891
regionr.2	1.90428	0.1565690	-22.79084
regionr.3	1.59708	0.0664275	-7.48634
msar	2.07600	-0.1684320	-6.17324

References

Barnard, J., and Rubin, D.B. (1999), "Small-Sample Degrees of Freedom with Multiple Imputation," *Biometrika*, 86, 948-955.

Botman, S.L., and Jack, S.S. (1995), "Combining National Health Interview Survey Datasets: Issues and Approaches," *Statistics in Medicine*, 14, 669-677.

Box, G.E.P., and Cox, D.R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society*, Series B, 26, 211-243.

Carlin, J.B., Li, N., Greenwood, P., and Coffey, C. (2003), "Tools for Analysing Multiple Imputed Data Sets," *The Stata Journal*, 3, 226-244.

Li, K.H., Meng, X.L., Raghunathan, T.E., and Rubin, D.B. (1991), "Significance Levels from Repeated p values with Multiply-Imputed Data," *Statistica Sinica*, 1, 65-92.

Li, K.-H., Raghunathan, T.E., and Rubin, D.B. (1991), "Large Sample Significance Levels from Multiply-Imputed Data Using Moment-Based Statistics and an F Reference Distribution," *Journal of the American Statistical Association*, 86, 1065-1073.

Little, R.J.A., and Rubin, D.B. (2002), *Statistical Analysis with Missing Data*, 2nd edition, Hoboken: Wiley.

National Center for Health Statistics (2010), "2009 National Health Interview Survey (NHIS) Public Use Data Release: NHIS Survey Description," Division of Health Interview Statistics, National Center for Health Statistics, Centers for Disease Control and Prevention. Available from the NHIS Web site (www.cdc.gov/nchs/nhis.htm).

Paulin, G.D., and Sweet, E.M. (1996), "Modeling Income in the U.S. Consumer Expenditure Survey," *Journal of Official Statistics*, 12, 403-419.

Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., and Solenberger, P. (2001), "A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models, *Survey Methodology*, 27, 85-95.

Rubin, D.B. (1978), "Multiple Imputation in Sample Surveys – A Phenomenological Bayesian Approach to Nonresponse," in *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp. 20-34.

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

Rubin, D.B. (1996), "Multiple Imputation After 18+ Years," *Journal of the American Statistical Association*, 91, 473-489.

Rubin, D.B., and Schenker, N. (1986), "Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse," *Journal of the American Statistical Association*, 81, 366-374.