United States
Department of
Agriculture

Forest Service

**Southern
Research Station**

Research Paper
SRS–49

# Multiple Imputation: An Application to Income Nonresponse in the National Survey on Recreation and the Environment

**Stanley J. Zarnoch, H. Ken Cordell, Carter J. Betz, and John C. Bergstrom**

## The Authors:

**Stanley J. Zarnoch,** Mathematical Statistician, U.S. Department of Agriculture, Forest Service, Southern Research Station, Asheville, NC 28804; **H. Ken Cordell,** Pioneering Research Scientist, U.S. Department of Agriculture, Forest Service, Southern Research Station, Athens, GA 30602; **Carter J. Betz,** Outdoor Recreation Planner, U.S. Department of Agriculture, Forest Service, Southern Research Station, Athens, GA 30602; and **John C. Bergstrom,** Professor, Department of Agricultural and Applied Economics, University of Georgia, Athens, GA 30602.

Cover background photo: Old Fort Picnic Area, Pisgah National Forest, Pisgah Forest, NC.
(photo courtesy of Bill Lea)

May 2010

Southern Research Station
200 W.T. Weaver Blvd.
Asheville, NC 28804

# Multiple Imputation: An Application to Income Nonresponse in the National Survey on Recreation and the Environment

**Stanley J. Zarnoch, H. Ken Cordell, Carter J. Betz, and John C. Bergstrom**

**Abstract**

Multiple imputation is used to create values for missing family income data in the National Survey on Recreation and the Environment. We present an overview of the survey and a description of the missingness pattern for family income and other key variables. We create a logistic model for the multiple imputation process and to impute data sets for family income. We compare results between estimates of the income distribution based on no imputation, single imputation, and multiple imputation. Although the imputation methodology has been applied to the income variable, it is transferable as a general approach to dealing with item nonresponse for other variables in this and other survey studies.

**Keywords**: MAR assumption, MCAR assumption, nonresponse bias, outdoor recreation, single imputation.

## Introduction

The National Survey on Recreation and the Environment is a national random digit dialing telephone survey of U.S. households on outdoor recreation and includes demographics of the civilian non-institutionalized population and its attitudes toward environmental and natural resources issues. When it was first conducted in 1960 as the National Recreation Survey, the survey consisted of interviewing people in their homes about outdoor recreation participation. The next five surveys changed in methodology and sponsorship. In 1994–95, the survey was revamped as a household telephone survey and renamed the National Survey on Recreation and the Environment (NSRE) to address public interest and concerns with the natural environment. The new survey added questions about wildlife, environmental values, public management issues, and the needs of physically challenged recreationists. The current modification, NSRE 1999–2007, is a national household survey on outdoor recreation participation and includes an extensive demographic profile of the interviewees along with information about the interviewees' attitudes on environmental issues, natural resource values, and management policy issues. All told, 19 versions of the survey have been conducted over the years, with various modifications; each modification entailed completing nearly 5,000 questionnaires, for a total of 92,558 questionnaires. Sponsoring agencies include the U.S. Forest Service, U.S. Department of Commerce National Oceanic and Atmospheric Administration, USDA Economic Research Service, U.S. Environmental Protection Agency, USDA Bureau of Land Management, USDI National Park Service, the University of Georgia, and the University of Tennessee. Further information on the history, survey questions, reports, and data can be found at http://www.srs.fs.usda.gov/trends/Nsre/nsre2.html.

Besides the survey's sponsoring organizations, many agencies and research groups use NSRE data to address various recreational, environmental, and natural resources issues, including the study of outdoor recreation use by demographic profile (Bowker and others 2008, Johnson and others 2001, Johnson and others 2004), and recreation activity participation models (Bowker and others 2006). Often a key variable in such a study is family income or household income (hereafter referred to as income), but a survey question about income often receives disproportionately high-item nonresponse (Schenker and others 2006). In NSRE surveys, about one-third of interviewees refused to divulge income. With such a high item nonresponse rate for income, the integrity and quality of estimates and analyses that use income in their formulation can become compromised because these entire interviews (not just income) are deleted from such analyses, leading to possible bias and larger standard errors resulting from reduced sample size. In an attempt to alleviate such problems, multiple imputation can create a more complete data set by substituting income responses for those missing observations.

The objectives of this research are, first, to describe the extent of the problem of nonresponse on income for the NSRE survey and the impact of this nonresponse on the bias and standard errors of subsequent survey estimates; second, to introduce multiple imputation methodology by using a logistic income model as a function of demographic variables, outdoor recreation activity participation variables, and a survey design variable; third, to create a more complete NSRE data set by performing multiple imputation on the income variable; and fourth, to evaluate and compare estimates and analyses based on multiple imputation to alternatives with no imputation and single imputation. The data used in this paper are from the most recent NSRE survey (1999–2007).

## Multiple Imputation

### Background

Most large scale surveys are subject to some nonresponse. The nonresponse, in the form of either unit or item nonresponse, may result in biased estimates and increased standard errors, leading to inefficient use of the data. Unit nonresponse occurs when an interviewee does not answer the survey at all, resulting in missing data for all the survey questions for that individual. In such situations, the conductor of the survey could take such corrective action as deleting those observations and weighting the data to better represent a known, reliable population base, like that constructed by the U.S. Census Bureau. Item nonresponse occurs when the interviewee does not answer a subset of survey questions. The conductor of the survey could augment these incomplete question subsets by using imputation that creates values for the missing data based on the questions that the individual answered during the survey (Rubin 1987). Item nonresponse is usually handled by multiple imputation and not by weighting. The focus of this paper is to address how multiple imputation can handle item nonresponse.

Traditionally, item nonresponse has been handled by simply analyzing the data with as many observations as possible for a given type of analysis. For instance, survey means might be computed on different sample sizes depending on the pattern of missingnesss in the data. If modeling is performed, sample size might be substantially reduced because an entire observation would be deleted whenever any variable in the model had item nonresponse for that observation. Although this approach is simple, it creates potential problems for the analyses. First, if the missingness is not at random, a bias may result because the sample will not represent the target population of interest if the nonrespondents differ from the respondents in certain ways, which are usually unknown. Second, item nonresponse results in reduced sample size which yields an increase in the variance of the estimates, in turn leading to loss of precision and inefficient use of the data.

An alternative approach is single imputation, which consists of using the sample mean from nonmissing observations to fill in the missing value (Rubin 1987). Mean substitution assigns the same value to all missing observations, resulting in a more peaked distribution with an artificially reduced variance. This subsequently results in narrower confidence intervals, smaller p-values for hypothesis tests, and inflated Type I error rates, because the substitute value is treated as known without error and does not reflect the true variability in the data. An extension of the mean substitution method is to predict the observation based on a regression model using nonmissing data. However, this method also results in the same observation conditioned on a given set of values for the explanatory variables, thus, also yielding invalid, reduced variances.

Multiple imputation has been developed to circumvent these problems by creating a set of substitute values for each missing value (Rubin 1987, 1996). Usually five such sets are recommended. In the case of logistic multiple imputation, which is described in this paper, the replacement observations are drawn from the posterior predicted distribution resulting from a logistic model fitted to a set of covariates in the sample. This method helps to maintain the natural variability in the data that otherwise would be lost by the methods previously discussed. Then the analyst computes the estimates on each of these data sets separately, and these estimates are then combined to arrive at estimates and variances that better reflect the variability in the population. The new estimates and analyses may be any of the usual types performed on survey data, such as the estimation of means, percentages, correlations, and regression models.

When evaluating the potential problem of item nonresponse, it is important to determine the type of randomness that the data exhibit. Two assumptions must be considered: the missing completely at random (MCAR) assumption and the missing at random (MAR) assumption.

The MCAR assumption implies that a missing value is missing randomly and does not depend on its value or that of any other variable. If this assumption is valid, the missing value could be deleted from the analysis without

any effect on the bias; however, with a reduced sample size, the variance will be increased. The MCAR assumption could be tested by separating the data into a missing group and a nonmissing group, and then testing for differences in the other continuous variables by means of two-sample t-tests. Categorical data can be tested by the Rao-Scott chi-square homogeneity test (Rao and Scott 1987). In addition, a logistic model could be fitted to both groups of data and odds ratios and confidence intervals computed and assessed. Little (1988) gives a more formal test.

When the MCAR assumption is not valid, then multiple imputation may be used. However, it requires the MAR assumption, which states that the probability that an observation is missing can depend on the values of the other variables but not on the value of the missing variable itself. Thus, the missing values are not distributed randomly across the data set but are random within an unknown subsample of the data. The MAR assumption cannot be tested with data, but it becomes more plausible as more variables are incorporated into the imputation model (van Buuren and others 1999). MAR is quite common compared to MCAR and suggests the use of multiple imputation. The MCAR is a more restricted assumption and is a special case of MAR. A more difficult problem is non-ignorable missingness; the data are neither MCAR nor MAR, and the necessary methods for handling this problem are beyond the scope of this paper (Rubin 1987).

A simple example can help delineate the distinction between the MCAR and MAR assumptions. Suppose that, in a survey, the income observations were each assigned a random real number between 0 and 1. The MCAR assumption would be valid if all income observations that received a random number less than, say, 0.1 were deleted. However, the MAR assumption would be met if income was deleted only for the females aged 25 and over who received a random number less than 0.1.

The multiple imputation applied here consists of developing a logistic model for the imputed income variable as a function of a set of dependent variables. When developing such an imputer's model, care must be taken that the variables are the same as the analysts' models. However, this is difficult to do or foresee when multiple imputation is used to create a more complete data set for future use by other analysts. To prevent biases and invalid inferences, the imputer's model should contain as many variables as is practically possible (Rubin 1996). Van Buuren and others (1999) provide guidelines for variable inclusion in the imputer's model. The guidelines cover three kinds of variables: those used by the analysts, those correlated with the imputed variable, and those correlated with the missing pattern of the imputed variable. Obviously, those variables that have a high proportion of missing values should not be included. In addition, the multiple imputation model should include variables related to the sampling design, such as stratification, cluster variables, and sampling weights (Rubin 1996). The usual problem with imputers' models is the inclusion of too few explanatory variables.

## Combining Estimates from Multiple Imputations

Multiple imputation creates $M$ imputed data sets for the imputed variable from each of which estimates and analyses can be performed by any of the standard statistical methods. Each data set consists of the same set of complete nonmissing data plus one of the sets of the imputed values on the missing variable. These are then combined as follows to yield the imputed estimates and variances. Let $\hat{Q}_i$ = the estimated value of parameter $Q$ and $\hat{U}_i$ = the estimated variance of $\hat{Q}_i$ based on the complete data from imputation $i$. The combined imputed estimate for $Q$ is the average of the $M$ complete-data estimates and is defined as

$$\bar{Q}_M = \frac{1}{M}\sum_{i=1}^{M}\hat{Q}_i \tag{1}$$

The total variance for $\bar{Q}_M$ is composed of the within-imputation variance

$$\bar{U}_M = \frac{1}{M}\sum_{i=1}^{M}\hat{U}_i \tag{2}$$

and the between-imputation variance

$$B_M = \frac{1}{M-1}\sum_{1=i}^{M}\left(\hat{Q}_i - \bar{Q}_M\right)^2 \tag{3}$$

and is defined as

$$T_M = \bar{U}_M + \left(\frac{M+1}{M}\right)B_M \tag{4}$$

The proportion of the total variance that is due to between-imputation variability is

$$\hat{\gamma}_M = \left( \frac{M+1}{M} \right) \frac{B_M}{T_M} \tag{5}$$

which is approximately the proportion of information that is missing about $Q$ due to nonresponse. Tests and confidence intervals for the parameter $Q$ can be based on the t-distribution where

$$\frac{Q - \bar{Q}_M}{\sqrt{T_M}} \sim t_{\upsilon_M} \tag{6}$$

with degrees of freedom

$$\upsilon_M = \frac{\left(M-1\right)}{\hat{\gamma}_M^2} \tag{7}$$

The relative increase in variance due to nonresponse is defined as

$$r = \left( \frac{M+1}{M} \right) \frac{B_M}{\bar{U}_M} \tag{8}$$

The fraction of missing information about $Q$ is

$$\hat{\lambda} = \frac{r + 2/\left(\upsilon_M + 3\right)}{\left(r+1\right)} \tag{9}$$

The relative efficiency of using $M$ imputation instead of an infinite number is

$$RE = \left( 1 + \frac{\hat{\lambda}}{M} \right)^{-1} \tag{10}$$

The number of imputations needed could be determined by replicating sets of $M$ imputations and determining when the estimates are stable (Horton and Lipsitz 2001).

## Methods

### Income Nonresponse

Family income was asked of all interviewees in each of the eight annual surveys from 1999 to 2007. The survey was conducted via a random digit dialing system, and the questionnaire was administered to any person who answered the telephone and was 16 years or older. Adults were uncomfortable answering the income question, as were respondents aged 16-19. As well, because of the probability of juveniles answering the income question inaccurately, all interviews of those 16 to 19 years old were deleted from this imputation project.

In the early part of the 1999 survey, if respondents refused to divulge actual income, they were then asked to specify the income range (of 11 classes) that their income fell into. Later, in the 1999 to 2003 surveys, respondents were given the option (by means of a screener question) of providing an actual income figure or an income range. Many respondents preferred to provide an income range, which resulted in reduction of the percent response for actual income from 56.7 percent in 1999 to about 32.1 percent for 2000 to 2003 (table 1). When given the option to provide actual income or income range, the interviewees were distributed about equally between the two choices. The total nonresponse rate for both of these income questions was generally in the 30 to 37 percent range when both actual and income range were used in the survey questions. With the hope of decreasing the nonrersponse rate for income, only income range was asked in the 2005 to 2007 surveys. The screener question was stopped at the end of the 2003 survey, after which the annual nonresponse rate dropped by 9 percent, to 23.1 percent, in future survey years (table 1).

### Relationship of Income Nonresponse to Demographic Variables

The effect of nonresponse is important if the nonresponse percent is related to demographic variables because this indicates that certain segments of the population are not represented in the sample in proportion to what they are in the general population, resulting in biased estimates. A simple univariate method to investigate this effect is to separate the data into two groups, one for those that responded and the other for those that did not. The Rao-Scott chi-square test for survey data was performed to test the homogeneity hypothesis that the proportion of missing income observations was the same for all groups within a demographic variable (table 2). All demographic variables

**Table 1—The percent of interviewees who provided their actual income or income range (11 classes) and the total nonresponse percent for each of the 8 interview years (unweighted)**

| Year | Number of interviews | Income screener asked[a] | Actual income response | Income range response[b] | Total response | Total nonresponse |
|---|---|---|---|---|---|---|
| | | | *percent* | | | |
| 1999 | 6,448 | Yes | 56.7 | 12.6 | 69.3 | 30.7 |
| 2000 | 19,941 | Yes | 28.7 | 36.6 | 65.3 | 34.7 |
| 2001 | 22,675 | Yes | 32.6 | 29.9 | 62.6 | 37.4 |
| 2002 | 9,404 | Yes | 37.1 | 26.9 | 64.0 | 36.0 |
| 2003 | 9,714 | Yes | 32.8 | 35.3 | 68.1 | 31.9 |
| 2005 | 7,158 | No | - | 77.6 | 77.6 | 22.4 |
| 2006 | 5,298 | No | - | 75.0 | 75.0 | 25.0 |
| 2007 | 5,371 | No | - | 78.0 | 78.0 | 22.0 |

[a] The income screener gave the interviewee the option of reporting income by either divulging actual income or indicating an income range (11 classes). In 1999, some interviewees were first asked to report their actual income, and if refused, were then asked to indicate an income range (11 classes).

[b] The percent response for the income range question does not include interviewees who also answered the actual income question. This step eliminated double counting in these two percents, resulting in a valid total response percent.

showed a significant difference. The youngest and oldest age classes had a much higher rate of nonresponse than the middle-age classes. Females had a higher nonresponse rate than males. Blacks and Hispanics had a higher nonresponse rate than Whites. The income nonresponse rates of Native Americans and Asians fell between that of Blacks and Whites. There was a decreasing trend in nonresponse as education level increased, with twice the nonresponse rate from those without a high school degree as from those with a post-graduate education. Although residence in a metropolitan county was significant, the nonresponse rates were very similar and the statistical significance was probably due mainly to large statistical power resulting from large sample size. A similar situation existed for census region of the interviewee. The unemployed also exhibited much greater nonresponse than those who were employed. Thus, the results on these demographic variables indicate that nonresponse is not a random phenomenon over the sample.

An alternate multivariate method to test the effect of nonresponse is to fit a logistic model where the response variable for an individual is 1 if it is a nonrespondent and 0 if a respondent; the demographic variables are then used as explanatory variables. Odds ratio estimates, which represent the odds of one class not responding with respect to another, are then obtained along with Wald confidence intervals. The logistic model with the demographic variables had Wald chi-squares that were all significant with $p<0.0001$ except for *metro* which was $p=0.2675$. The odds ratios corroborate the conclusions reached by the Rao-Scott chi-square tests (table 3). The odds ratios for the youngest and oldest age classes were both significantly different from 1.0 when compared to the middle classes, indicating that they were more likely to not respond. Females were 1.37 times or 37 percent more likely than males to not answer the income question in the survey. The *ethrace* variable showed that Blacks, Asians, and Hispanics were more likely to not respond than Whites and Native Americans. Education again showed a decreasing rate of nonresponse as education level increased. A metropolitan resident was equally likely to exhibit nonresponse as a nonmetropolitan. The odds ratios for census region were close to 1.0. However, the Northeast was more likely to exhibit nonresponse than the other three regions, while the West was less likely to exhibit nonresponse than the other regions. The unemployed were more likely to not respond than the employed.

The results from both the univariate Rao-Scott chi-square tests and the multivariate logistic model substantiate that there is a relationship between nonresponse and most of the demographic variables. Thus, the observations in the survey that contain a value for income are not a random sample of the total data set and the MCAR assumption is rejected. Obviously, any analyses that use income could lead to possible biases and larger variances.

**Table 2—The percent missing for the income variable for each of the demographic variables (weighted)**

| Demographic variable (Rao-Scott $\chi^2$) | Total sample size | Percent missing | Standard error | 95 percent confidence interval |
|---|---|---|---|---|
| **AGECAT**[a] (1,023) | | | | |
| 2=20-24 | 5,595 | 41.9 | 0.77 | 40.4, 43.4 |
| 3=25-34 | 14,892 | 29.7 | 0.60 | 28.5, 30.9 |
| 4=35-44 | 18,014 | 28.6 | 0.46 | 27.7, 29.5 |
| 5=45-54 | 18,463 | 30.9 | 0.50 | 29.9, 31.9 |
| 6=55-64 | 13,453 | 36.8 | 0.65 | 35.5, 38.0 |
| 7=65+ | 14,036 | 49.3 | 0.69 | 47.9, 50.6 |
| **SEX** (301) | | | | |
| 0=Female | 48,190 | 40.4 | 0.33 | 39.8, 41.1 |
| 1=Male | 37,382 | 31.6 | 0.38 | 30.9, 32.4 |
| **ETHRACE**[b] (181) | | | | |
| 1=White[c] | 71,039 | 33.2 | 0.21 | 32.8, 33.7 |
| 2=Black[c] | 5,404 | 40.0 | 0.90 | 38.3, 41.8 |
| 3=Native American[c] | 1,107 | 36.1 | 2.18 | 31.8, 40.4 |
| 4=Asian[c] | 1,389 | 36.0 | 1.80 | 32.4, 39.5 |
| 5=Hispanic | 4,809 | 44.0 | 1.06 | 41.9, 46.1 |
| **ED** (1,361) | | | | |
| 1=Less than high school | 4,913 | 52.0 | 1.01 | 50.0, 54.0 |
| 2=High school graduate | 22,018 | 39.2 | 0.39 | 38.4, 39.9 |
| 3=Some college or technical school | 26,021 | 30.2 | 0.33 | 29.6, 30.9 |
| 4=Bachelor's degree | 19,320 | 27.3 | 0.37 | 26.6, 28.0 |
| 5=Post-graduate degree | 12,319 | 26.0 | 0.49 | 25.0, 27.0 |
| **METRO** (9.6) | | | | |
| 0=No | 27,297 | 37.7 | 0.45 | 36.8, 38.6 |
| 1=Yes | 58,712 | 36.0 | 0.29 | 35.5, 36.6 |
| **CREGION** (38.1) | | | | |
| 1=Northeast | 15,643 | 38.3 | 0.55 | 37.3, 39.4 |
| 2=Midwest | 21,299 | 35.5 | 0.47 | 34.6, 36.4 |
| 3=South | 29,823 | 37.3 | 0.43 | 36.5, 38.2 |
| 4=West | 19,244 | 34.2 | 0.54 | 33.2, 35.3 |
| **EMPLOY** (891) | | | | |
| 0=No | 27,923 | 46.1 | 0.47 | 45.2, 47.1 |
| 1=Yes | 57,650 | 30.3 | 0.27 | 29.8, 30.9 |

[a] AGECAT=1 was age 16 to 19 years which was eliminated from the imputation process.

[b] ETHRACE 1 and 4 were combined and 2, 3, and 5 were combined for the imputation process.

[c] NonHispanic.

Note: The Rao-Scott chi-square tests for homogeneity of the percent missing were all significant with $p < 0.0001$ except for METRO which was 0.0019.

**Table 3—The odds ratio from a logistic model (weighted) where the probability of a nonresponse is modeled**

| Demographic variable | Reference class | Reference class | Reference class | Reference class | Reference class | Reference class |
|---|---|---|---|---|---|---|
| **AGECAT** | 20-24 | 25-34 | 35-44 | 45-54 | 55-64 | 65+ |
| 20-24 | - | 1.74* | 1.79* | 1.61* | 1.39* | 1.02 |
| 25-34 | 0.57* | - | 1.03 | 0.92* | 0.80* | 0.59* |
| 35-44 | 0.56* | 0.97 | - | 0.90* | 0.78* | 0.57* |
| 45-54 | 0.62* | 1.08* | 1.12* | - | 0.87* | 0.64* |
| 55-64 | 0.72* | 1.25* | 1.29* | 1.15* | - | 0.74* |
| 65+ | 0.98 | 1.70* | 1.75* | 1.57* | 1.36* | - |
| **SEX** | Female | Male | | | | |
| Female | - | 1.37* | | | | |
| Male | 0.73* | - | | | | |
| **ETHRACE** | White | Black | Native Am. | Asian | Hispanic | |
| White | - | 0.82* | 0.97 | 0.71* | 0.68* | |
| Black | 1.22* | - | 1.19 | 0.87* | 0.84* | |
| Native Am. | 1.03 | 0.84 | - | 0.73* | 0.71* | |
| Asian | 1.41* | 1.15 | 1.37* | - | 0.96 | |
| Hispanic | 1.46* | 1.20* | 1.42* | 1.04 | - | |
| **ED** | No HS | HS | Some College | BS | Post-grad | |
| No HS | - | 1.42* | 2.04* | 2.25* | 2.47* | |
| HS | 0.71* | - | 1.44* | 1.59* | 1.74* | |
| Some College | 0.49* | 0.69* | - | 1.10* | 1.21* | |
| Bachelor's | 0.43* | 0.63* | 0.91* | - | 1.10* | |
| Post-grad | 0.41* | 0.57* | 0.83* | 0.91* | - | |
| **METRO** | No | Yes | | | | |
| No | - | 1.01 | | | | |
| Yes | 0.99 | - | | | | |
| **CREGION** | Northeast | Midwest | South | West | | |
| Northeast | - | 1.14* | 1.14* | 1.26* | | |
| Midwest | 0.88* | - | 1.00 | 1.11* | | |
| South | 0.88* | 1.00 | - | 1.11* | | |
| West | 0.79* | 0.90* | 0.90* | - | | |
| **EMPLOY** | No | Yes | | | | |
| No | - | 1.35* | | | | |
| Yes | 0.74* | - | | | | |

Note: The columns represent the reference class for each of the rows within a given demographic variable. An asterisk (*) represents a significant odds ratio from 1.0 using the Wald chi-square and an alpha of 0.05. The odds ratios in the upper diagonal for a demographic variable are the reciprocal of those in the lower diagonal.

### Explanatory Variables for the Income Model

The logistic model developed for imputation includes the response variable for income group, *incgrp*, which was formulated as seven income groups derived from the interviewees' responses to the actual income question or the 11 classes of income asked during the survey. The seven groups were sufficient for *incgrp* defined as 1=less than 15,000; 2=15,000 to 24,999; 3=25,000 to 49,000; 4=50,000 to 74,999; 5=75,000 to 99,999; 6=100,000 to 149,999; and 7=150,000 and over. Overall, *incgrp* had a 32.5 percent nonresponse rate.

The imputation process required the identification of relevant explanatory variables used to develop a logistic model for the ordinal *incgrp* variable. The imputation approach used a complete set of data with no missing values for any of the explanatory variables. This simplified the imputation by requiring only the imputation of *incgrp* instead of any missing values for the explanatory variables. Although it was possible to impute these explanatory variables in a sequential manner (Raghunathan and others 2001), the increased complexity was considered unnecessary because the percent missing was quite small. Only demographic variables with less than 3 percent missing were selected as candidates for the logistic model which included *agecat*, *sex*, *ethrace*, *ed*, *metro*, *cregion*, and *employ*. These were similar to the variables used by Schenker and others (2006) for their income imputation for the National Health Interview Survey. The binary variable *born in USA* (no, yes) also met the 3 percent missing criterion and was used in the initial modeling efforts. However, it had very few observations in the "no" category which resulted in a sparse data matrix and, thus, it was eliminated. Several other binary variables such as *student*, *retired*, and *homemaker* were also considered as were continuous variables such as *hours worked*, *family size*, *number of children under 6 years old*, and *number of children under 16 years old*. Although some of these were found to be important explanatory variables in preliminary modeling efforts, their nonresponse rates were over 30 percent and inclusion would greatly reduce the number of cases for modeling purposes.

Also examined was another set of explanatory variables pertaining to outdoor recreation activity participation. Future use of the imputed data would be for developing outdoor recreation activity participation and consumption models. Thus, since the usefulness of imputation is dependent on the imputer's model being the same as the analyst's model, explanatory variables that pertained to these activities and that were conceivably related to income were selected from the survey (table 4). All these variables were binary (0, 1), reflecting whether a person participated in the specific activity or not. Many of the participation questions were asked only in certain versions of the survey, resulting in many missing values. Hence, they were combined into one of three recreation activity participation indices. Participation *Index1* consisted of 11 passive participation variables that require little physical exertion, usually close to home, and at low cost. Participation *Index2* was based on 19 active, non-motorized activities that require a moderate amount of physical exertion, considerable commute to the place of the activity, and moderate expense. Participation *Index3* contained six variables and was similar to *Index2* except that the activities of *Index3* are motorized. Each index was simply defined as binary where 1 indicated that the interviewee participated in at least one of the activities in the index's group of variables or 0 otherwise.

Survey design should also be considered when developing an imputation model. The NSRE survey is a random digit dialing survey design, and, thus, there is no need to account for stratification or clustering as is typical in more complex surveys. However, the NSRE provides a *weight* variable that adjusts the sampling fraction in the survey to better represent the non-institutionalized population (as defined by the U.S. Census Bureau). This *weight* was used as a sampling weight and as an explanatory variable for the logistic income model.

### Initial Income Model Development

The cumulative logit model for income was initially developed based on data that had complete case responses for the income variable *incgrp* and the seven demographic variables. The predictive ability of the final model was assessed with the max-rescaled $R^2$ (Nagelkerke 1991) and several indices of rank correlation. The modeling process contained three stages.

The first stage considered all demographic variables and their two-way interactions as potential candidates for the model using a stepwise approach. The objective was to determine which variables and interactions were important in the modeling of *incgrp*. To preserve model hierarchy, the demographic variables were forced into the model regardless of their significance levels, while the interactions were selected with the stepwise method using PROC LOGISTIC (SAS Institute Inc. 2004) with both selection and deletion criteria set at 0.05.

**Table 4—Variables (unweighted) used in the passive participation INDEX1; active, non-motorized participation INDEX2; and active, motorized participation INDEX3 (n=86,009)**

| Variable | Definition | Percent missing |
|---|---|---|
| **INDEX1** | Participates in at least one of the passive recreation participation variables | 1.7 |
| BEACH | Visit a beach | 6.8 |
| BIRD | View and/or photograph birds | 4.7 |
| FAM | Family gathering | 22.9 |
| FV | View and/or photograph fish, etc. | 4.0 |
| GATHERMB | Gather mushrooms, berries, etc. | 19.3 |
| OV | View and/or photograph flowers, etc. | 5.1 |
| PICNIC | Picnicking | 16.0 |
| SCENERY | View and/or photograph natural scenery | 5.1 |
| SWIMP | Swimming in an outdoor pool | 52.2 |
| WALK | Walk for pleasure | 13.6 |
| WV | View and/or photograph other wildlife | 5.3 |
| **INDEX2** | Participates in at least one of the active, non-motorized recreation participation variables | 1.7 |
| BACPAC | Backpacking | 14.9 |
| BOATF | Rafting | 6.6 |
| CAMPRI | Primitive camping | 23.5 |
| CANOE | Canoeing | 6.6 |
| CAVE | Caving | 79.6 |
| CSKI | Cross-country skiing | 21.3 |
| DEVCAM | Developed camping | 17.0 |
| DSKI | Downhill skiing | 21.3 |
| FISH | Fishing | 1.8 |
| HIKE | Day hiking | 15.7 |
| HORSETRL | Horseback riding on trails | 22.0 |
| HUNT | Hunting | 1.8 |
| KAYAK | Kayaking | 6.6 |
| MC | Mountain climbing | 81.6 |
| MTNBIKE | Mountain biking | 18.6 |
| RC | Rock climbing | 79.6 |
| SNOBORD | Snowboarding | 21.3 |
| SWIM | Swimming in lakes, ponds, etc. | 6.6 |
| WILDERN | Visit a wilderness | 15.8 |
| **INDEX3** | Participates in at least one of the active, motorized recreation participation variables | 1.7 |
| BOATM | Motorboating | 6.6 |
| BOATWS | Waterskiing | 16.9 |
| DRIVING | Driving for pleasure | 25.9 |
| MV | Drive off-road | 23.2 |
| SITSEE | Sightseeing | 29.6 |
| SNMOB | Snowmobiling | 21.3 |

A few complicating issues surfaced during stage 1 that had to be resolved before progressing. First, as mentioned previously, the age group 16–19 was not used because the validity of income obtained from juveniles was thought questionable. Second, the *ethrace* variable was reduced from five classes to two due to small sample size for the Native American and Asian classes. The White and Asian were combined into *ethrace*=1 and the Black, Native American, and Hispanic were combined into *ethrace*=2. It was felt that the income pattern was similar within these two new composite classes. Third, a *weight* variable was included in the logistic modeling process as a sampling weight but not included as an explanatory variable in the model at this stage. Fourth, to avoid having to adjust the categorical income group endpoints each year according to the Consumer Price Index for annual income inflation adjustments, a separate model was fit each year which accounts for annual inflation through the specific annual parameter estimates for each model. Fifth, the selection of important interactions was based on the criterion that an interaction had to be selected at least three times among the eight annual models (1999 to 2007) for the interaction to be included in subsequent model development. The overall goal was to use the same variables for each annual model.

Stage 2 was built upon the results of stage 1 where it was found that most of the interactions were selected as significant in at least three of the eight annual models. This meant that all interactions satisfied the criterion presented in the previous paragraph. Thus, stage 2 began by keeping all seven demographic variables and all 28 two-way interactions in the model for evaluation. The max-rescaled $R^2$'s were improved very little compared to the models developed under stage 1, which did not always contain all the interactions. Thus, to simplify the models and prevent over-fitting problems, some interactions that were not significant in some models, or that entered the stepwise process only near the final step, were eliminated from all models. The philosophy was to have identical models with a consistent structure for each year.

The third stage further developed the model by incorporating the three activity participation indices with their two-way interactions into the final stage 2 model. In addition, the *weight* variable was included as an explanatory variable in the model as well as a sampling weight variable as used in the previous two stages. The three indices' interactions were subsequently removed because they were usually not significant. Although the inclusion of the indices improved the max-rescaled $R^2$ slightly, usually by 1 to 2 percentage points, they were included for their value in modeling activity participation so that the analyst's model is compatible with the imputer's model. The final logistic income model was defined as

*logit(incgrp)=f(intercept agecat sex ethrace ed metro cregion employ agecat*sex agecat*ethrace agecat*ed agecat*employ sex*employ sex*ed ethrace*ed ethrace*cregion ed*cregion ed*employ index1 index2 index3 weight*)

The final model was evaluated on several fit statistics. The max-rescaled $R^2$ for all years ranged from 0.30 to 0.44 with the later years being the highest. All annual models had percent concordant slightly above 70 percent and percent discordant slightly below 29 percent. The rank correlation statistics were quite uniform for the eight annual models. Somers' D ranged from 0.417 to 0.473, Goodman-Kruskal gamma ranged from 0.420 to 0.475, Kendall's tau-a ranged from 0.327 to 0.391, and c ranged from 0.709 to 0.737. The percent of observations predicted in the observed income class was 34.6 percent and the percent predicted within one income class was 72.3 percent. The actual imputation process including the fitting of the final logistic model and creation of the imputed data was performed by using PROC MI and PROC MIANALYZE (SAS Institute Inc. 2004).

## Results

### Evaluation of the Imputation Model

**Background**—To test functioning and validity of the imputation model, a subset population of the NSRE data was developed to include n=54,895 observations that had complete responses for *incgrp* and all the independent variables in the logistic model. The exact income distribution could then be obtained and considered as the known (true) seven income groups in *incgrp* for this subset population because this income distribution was computed from data with no missing income observations. Scenarios were then developed that consisted of different types of missingness imposed on only the income variable. For each scenario, the logistic income model was refitted and used to perform the imputation, and the imputed income distribution obtained and compared to the complete data income distribution. The same model form developed previously was used but the estimated parameters varied for each scenario because they were based on different data depending on the missingness pattern for each scenario. Each scenario was evaluated under no imputation ($M$=0), single imputation ($M$=1), and multiple imputation ($M$=5, 10, 20, 100). The comparisons to the complete data were

performed for each year and all years combined, but only presented for the latter because the yearly results were very similar. Although this evaluation cannot provide irrefutable evidence that the imputation process is functioning appropriately, it can ensure that the form of the logistic model is stable and the parameter estimates are reasonable under different patterns of missingness.

**Missing Completely at Random: Scenario 1**—The first scenario tested was to determine if the imputation process functions adequately when income observations are missing completely at random (which is analogous to the MCAR assumption). For this scenario, 30 percent of the income observations were randomly selected and assigned a missing value. The results indicate that the estimated income distributions based on no imputation (*M*=0), single imputation (*M*=1), and multiple imputations (*M*=5, 10, 20, 100) were very close to that of the complete data (table 5). These results are as expected because the data were randomly deleted, which satisfies the MCAR assumption. The standard errors for no imputation (*M*=0) are inflated due to reduced sample size resulting from the 30 percent missing income data. The single imputation (*M*=1) standard errors are very close to that of the complete data, and, hence, do not reflect the uncertainty due to the 30 percent imputed income observations. The standard errors for the multiple imputation estimates (*M*=5, 10, 20, 100) are larger and reflect the increased variability due to imputation. All these results are as expected and, thus, the imputation model and process appears to be reasonably valid, based on this scenario.

**Missing at Random: Scenario 2**—To further test the imputation process, the MAR assumption was imposed on the complete income data. This test is identical to that previously done with the MCAR scenario except that, in this test, 30 percent of the respondents with a high school degree or less (*ed* = 1 or 2), or who were 20 to 34 years old (*agecat* = 2 or 3) or 65 years or older (*agecat* = 7), now had their responses about their income changed to missing. This change was achieved by first assigning a random number between 0 and 1 to all observations in this group and then changing the income response to missing if the random number was less than 0.3. These respondents usually report having a lower income than the general population, and, thus, this scenario should result in the no imputation income distribution (*M*=0) having lower percents in the lower income classes and, consequently, more in the upper classes. A proper imputation process should correct for this discrepancy and yield income distributions similar to the complete data. This scenario is similar to the pattern of missingness originally detected in the NSRE data that indicated respondents in demographic classes characterized by lower incomes had a higher rate of missingness (tables 2, 3).

Under this scenario, the no imputation estimates (*M*=0) are biased as expected (table 5). The percent income group 1 was estimated to be 11.2 percent compared to 12.5 percent for the complete data, or 10.4 percent under estimate. This negative bias continued but diminished through income group 3. The bias became positive for income group 4 to 7, increasing to positive 11.6 percent for the highest income. This bias is as expected because the scenario changed 30 percent of the respondents, many of whom reported they were in the lower income classes, to missing, which removed these respondents from the analysis. All imputation estimates (*M*=1, 5, 10, 20, 100) are very close to those of the complete data. The standard errors for the no imputation estimates (*M*=0) are higher than the complete data because of a reduction in sample size. The standard errors for the single imputation (*M*=1) are similar to the complete data and do not reflect the uncertainty due to imputation. However, those standard errors for the multiple imputation estimates (*M*=5, 10, 20, 100) are larger, reflecting the added uncertainty due to the imputation.

**Missing at Random: Scenario 3**—This scenario consisted of deleting 30 percent of the incomes from respondents with at least a bachelor's degree (*ed* = 4, 5) and who were also 35 to 64 years old (*agecat* = 4, 5 , 6). This scenario should result in a higher percentage of respondents in the lower income classes and a lower percentage in the upper income classes for the no imputation estimates (*M*=0). Under this scenario, the no imputation estimates (*M*=0) were again biased (table 5). The percent income group 1 was estimated to be 13.0 percent compared to 12.5 percent for the complete data, or 4 percent over estimate. This positive bias continued but diminished through income group 3. The bias became negative for income group 4 to 7, increasing to negative 11.6 percent for the highest income. Imputation corrected these biases across the total income distribution. Single imputation (*M*=1) provided estimates almost identical to multiple imputation (*M*=5, 10, 20, 100) estimates, which were all very close to the complete data income estimates. The standard errors were very similar among the imputations, reflecting properties that were expected based on imputation theory. Generally, the no imputation (*M*=0) standard errors were slightly larger than the complete data, mainly for the lower income groups. The single imputation (*M*=1) had standard errors slightly smaller than the multiple imputations.

**Table 5—Evaluation of the imputed income distribution**

| Income class | Complete data | *M*=0 | *M*=1 | *M*=5 | *M*=10 | *M*=20 | *M*=100 |
|---|---|---|---|---|---|---|---|
| | *percent* | | | | | | |
| | *(standard error)* | | | | | | |
| Missing Completely at Random: Scenario 1 | | | | | | | |
| 1 | 12.5 (0.28) | 12.7 (0.31) | 12.7 (0.26) | 12.6 (0.27) | 12.6 (0.31) | 12.7 (0.30) | 12.7 (0.31) |
| 2 | 13.2 (0.23) | 13.2 (0.26) | 12.9 (0.22) | 13.1 (0.25) | 13.1 (0.26) | 13.1 (0.27) | 13.1 (0.25) |
| 3 | 31.7 (0.28) | 31.7 (0.33) | 31.6 (0.28) | 31.7 (0.31) | 31.7 (0.37) | 31.7 (0.31) | 31.7 (0.32) |
| 4 | 20.2 (0.21) | 19.9 (0.25) | 20.2 (0.22) | 20.0 (0.25) | 20.1 (0.24) | 20.1 (0.27) | 20.0 (0.25) |
| 5 | 10.4 (0.15) | 10.4 (0.18) | 10.6 (0.15) | 10.4 (0.17) | 10.4 (0.24) | 10.5 (0.18) | 10.5 (0.19) |
| 6 | 7.7 (0.14) | 7.8 (0.17) | 7.9 (0.14) | 7.8 (0.17) | 7.7 (0.15) | 7.8 (0.17) | 7.8 (0.17) |
| 7 | 4.3 (0.10) | 4.3 (0.13) | 4.1 (0.10) | 4.3 (014) | 4.2 (0.12) | 4.2 (0.13) | 4.2 (0.13) |
| Missing at Random: Scenario 2 | | | | | | | |
| 1 | 12.5 (0.28) | 11.2 (0.28) | 12.3 (0.25) | 12.4 (0.30) | 12.4 (0.30) | 12.3 (0.29) | 12.3 (0.29) |
| 2 | 13.2 (0.23) | 12.3 (0.24) | 12.9 (0.22) | 13.0 (0.24) | 13.0 (0.24) | 13.1 (0.25) | 13.0 (0.26) |
| 3 | 31.7 (0.28) | 30.9 (0.30) | 31.8 (0.28) | 31.6 (0.30) | 31.6 (0.30) | 31.6 (0.30) | 31.6 (0.32) |
| 4 | 20.2 (0.21) | 21.0 (0.24) | 20.2 (0.22) | 20.2 (0.27) | 20.3 (0.25) | 20.3 (0.24) | 20.3 (0.24) |
| 5 | 10.4 (0.15) | 11.3 (0.17) | 10.7 (0.16) | 10.5 (0.21) | 10.5 (0.18) | 10.5 (0.17) | 10.5 (0.17) |
| 6 | 7.7 (0.14) | 8.5 (0.16) | 7.7 (0.14) | 7.8 (0.15) | 7.8 (0.16) | 7.8 (0.17) | 7.8 (0.17) |
| 7 | 4.3 (0.10) | 4.8 (0.12) | 4.4 (0.15) | 4.4 (0.10) | 4.3 (0.15) | 4.3 (0.13) | 4.3 (0.15) |
| Missing at Random: Scenario 3 | | | | | | | |
| 1 | 12.5 (0.28) | 13.0 (0.29) | 12.5 (0.28) | 12.5 (0.28) | 12.4 (0.28) | 12.4 (0.28) | 12.4 (0.28) |
| 2 | 13.2 (0.23) | 13.7 (0.24) | 13.2 (0.23) | 13.2 (0.23) | 13.2 (0.23) | 13.2 (0.23) | 13.2 (0.23) |
| 3 | 31.7 (0.28) | 32.4 (0.29) | 31.8 (0.28) | 31.7 (0.28) | 31.7 (0.28) | 31.7 (0.28) | 31.7 (0.28) |
| 4 | 20.2 (0.21) | 20.0 (0.22) | 20.2 (0.21) | 20.3 (0.21) | 20.3 (0.21) | 20.3 (0.22) | 20.3 (0.22) |
| 5 | 10.4 (0.15) | 10.0 (0.15) | 10.4 (0.15) | 10.5 (0.17) | 10.5 (0.15) | 10.5 (0.16) | 10.5 (0.16) |
| 6 | 7.7 (0.14) | 7.1 (0.14) | 7.7 (0.14) | 7.7 (0.15) | 7.6 (0.14) | 7.7 (0.14) | 7.7 (0.14) |
| 7 | 4.3 (0.10) | 3.8 (0.10) | 4.3 (0.10) | 4.2 (0.11) | 4.3 (0.11) | 4.2 (0.11) | 4.2 (0.11) |

Note: Scenario 1 deletes 30 percent of income observations completely at random. Scenario 2 deletes 30 percent of income observations for respondents with a high school education or less (*ed* = 1 or 2) or aged 20 to 34 or 65+ (*agecat* = 2, 3 or 7). Scenario 3 deletes 30 percent of income observations for respondents with at least a bachelor's degree (*ed* = 4 or 5) and who are between 35 and 64 years old (*agecat* = 4, 5 or 6).

The values in parenthesis are standard errors.

### Imputed Income Distribution

To evaluate the imputation for the complete NSRE data, the income distribution was estimated with no imputation, and then with single and multiple imputations. All imputations yielded estimates for the two lowest income groups substantially more than that of no imputation, ranging from 8 to 26 percent higher estimates (table 6). Consequently, the middle and upper imputed income classes were lower when compared to no imputation, with the highest income class estimated at 14 percent lower. The cause for these different estimates helps to illustrate the functioning of the imputation process.

The youngest age class had the highest percent missing income data (table 2). Females also had a higher missing rate. In addition, minorities (Blacks and Hispanics) had the highest percent missing among the ethnic racial classes. Moreover, percent missing had an inverse relationship with level of education. Those respondents who were unemployed had a higher income nonresponse level than those who were employed. These factors led to a greater proportion missing for lower income respondents in the sample. Thus, this result explains why the no imputation estimates (*M*=0) tend to be less for the lower income groups which were underrepresented in the sample with respect to the income variable. Imputation corrects this bias by providing larger estimates for the lower income classes and smaller estimates for the upper income classes (table 6).

The estimates under any of the imputations are virtually the same for all practical purposes; however, the standard errors are different. The single imputation (*M*=1) standard errors are substantially lower than the no imputation (*M*=0) standard errors, because the former do not account for the added variability due to the imputed values. All multiple imputations provide larger standard errors to reflect this variability and, thus, more accurately reflect the underlying variation.

Although all the multiple imputations yielded very similar results, the number of imputations needed to reach convergence appears to be *M*=10, which is numerically superior for several reasons. First, because *M*=1 provides deflated standard errors, confidence intervals will be too narrow and test statistics will show higher levels of significance, increasing the chance of Type 1 errors. Second, there is a practical limit to the amount of generated imputed data that is feasibly carried and maintained in a database, with *M*=10 appearing to be the upper limit. Third, the relative efficiency (*RE*) (equation 10) for the imputations were approximately 0.90 for *M*=5 and 0.97 for *M*=10, increasing to1.0 for *M*=100. Based on the *RE*, it appears that *M*=10 is most appropriate among those evaluated.

**Table 6—Estimated income distribution percent (standard error) for no imputation, single imputation, and multiple imputation for the original NSRE data**

| Income class | *M*=0 | *M*=1 | *M*=5 | *M*=10 | *M*=20 | *M*=100 |
|---|---|---|---|---|---|---|
| | *percent* | | | | | |
| 1 | 12.5<br>(0.28) | 15.6<br>(0.25) | 15.6<br>(0.41) | 15.7<br>(0.31) | 15.7<br>(0.34) | 15.6<br>(0.34) |
| 2 | 13.2<br>(0.23) | 14.3<br>(0.20) | 14.4<br>(0.25) | 14.3<br>(0.27) | 14.3<br>(0.24) | 14.4<br>(0.25) |
| 3 | 31.7<br>(0.28) | 31.7<br>(0.23) | 31.6<br>(0.33) | 31.5<br>(0.29) | 31.6<br>(0.29) | 31.6<br>(0.29) |
| 4 | 20.2<br>(0.21) | 18.8<br>(0.17) | 18.7<br>(0.25) | 18.7<br>(0.21) | 18.7<br>(0.20) | 18.7<br>(0.21) |
| 5 | 10.4<br>(0.15) | 9.2<br>(0.12) | 9.2<br>(0.13) | 9.3<br>(0.13) | 9.3<br>(0.14) | 9.3<br>(0.14) |
| 6 | 7.7<br>(0.14) | 6.7<br>(0.10) | 6.7<br>(0.11) | 6.8<br>(0.12) | 6.8<br>(0.12) | 6.8<br>(0.12) |
| 7 | 4.3<br>(0.10) | 3.7<br>(0.08) | 3.7<br>(0.10) | 3.7<br>(0.09) | 3.7<br>(0.09) | 3.7<br>(0.09) |

## Discussion

The three illustrative scenarios and the imputed income distribution analysis using NSRE data demonstrate that multiple imputation is useful for correctly mitigating missing values, increasing sample size, and removing possible biases often found in household surveys. The pattern of missing NSRE income data was similar to that found in the National Health Interview Survey (Schenker and others 2006) where the income variable annually averaged approximately 30 percent missing from 1997 to 2004. Most of the NSRE demographic variables beside *incgrp* had a low rate of missingness which was similar to what Schenker and others (2006) found in their survey. Many of the demographic variables used for imputation in the National Health Interview Survey were similar to those used here, including age, sex, race/ethnicity, region, and metropolitan.

Caution and check scenarios should be performed before using an operational application of any income imputed data set. If thorough examination of the imputation process is not performed, many problems can become masked and go undetected. In this paper, evaluation was performed on three scenarios for each of the 8 years individually (data not shown) and on all years combined. This encompassed a total of 8(7+7+7)+7+7+7 =189 pairs of percent and standard error estimates for evaluation. In addition, each was estimated with no imputation, single imputation, and multiple imputation, and evaluated for any potential problems that may invalidate the imputation process. It is especially important to perform annual tests (which were done here but not presented) when individual annual logistic models are used for the imputation because combining over years may mask problems occurring for a particular annual logistic model. The problem of a sparse data set which could lead to complete separation or quasi-separation when fitting logistic models can easily go unnoticed, resulting in undefined maximum likelihood estimates of the logistic parameters, thus compromising the validity of the imputed data. The distinction between unit and item nonresponse must be carefully kept in mind when interpreting and evaluating the results of the imputation process. In the NSRE study, the tendency of certain demographic groups to be underrepresented in the survey was corrected by means of the sampling weight. This sampling weight corrects for unit nonresponse where the entire set of survey questions is missing for an individual. When biases were detected for some of the illustrated scenarios provided in this paper, they were due to item nonresponse because unit nonresponse was already corrected by means of the sampling weights.

While the imputation process used in this paper appears sound, there are several possibilities for improvement. For instance, it may be beneficial to perform imputation not only for the income variable but also for all the demographic variables used in the logistic model. Raghunathan and others (2001) have developed a multivariate technique that could be useful for this purpose; the technique involves imputing missing values on several variables, using a sequence of regression models. This would not only enable the use of more missing observations but also address other sources of potential bias.

Another possible improvement is refinement of the income variable, which may lead to improved logistic models for the imputation. The NSRE income variable pertains to income, but it may not be directly associated with the demographic variables of the interviewee, especially in situations when the interviewee is a teenager or adult child who may not be privy to information about family income. This problem was partially corrected in this study by deleting the first age class from the imputation process because it was felt that teenagers could not report income accurately to phone interviewers. The imputation of income based on an ordinal variable with seven groups, as was done in this study, may not be as strong as if income was continuous. Logistic models for ordinal data are more complex to evaluate than regression models for continuous data, and logistic models may exhibit the problem of complete separation or quasi-separation. In addition, the logistic model used here assumes the proportional odds model, which may not be justifiable in some instances. Despite these factors, a higher percent of respondents appear more willing to answer an ordinal class income variable question, as was demonstrated with the NSRE data for 2005 to 2007. Thus, it is unclear if a continuous income variable is actually better in such surveys.

## Conclusion

The problem of missing income values in the NSRE survey was described and evaluated. Methods for detecting departure from the MCAR assumption were performed and indicated a potential bias due to differential nonresponse by certain demographic groups of people. There were about 30 percent missing values for the income variable, most of which were supplemented through imputation. The logistic model was developed on an annual basis to control for inflation from 1999 to 2007. The model appeared to be an adequate predictor with max-rescaled $R^2$ in the low 30s

to 40s. An imputed data set with $M$=10 replications was developed and shown to be most appropriate among those evaluated. The imputation process was evaluated under three scenarios and appeared to function properly.

## Acknowledgments

We thank J.M. Bowker of the U.S. Forest Service and V.R. Leeworthy of the U.S. Department of Commerce (National Oceanic and Atmospheric Administration) for their reviews of an earlier version of this manuscript, as well as the National Survey on Recreation and the Environment (NSRE) for sharing NSRE data. The NSRE is managed by the U.S. Forest Service and the University of Georgia in Athens, and is a cooperative venture between a number of Federal agencies and some States. NSRE data is collected by the Human Dimensions Laboratory at the University of Tennessee in Knoxville, TN. For further details, contact Ken Cordell at kcordell@fs.fed.us.

## Literature Cited

Bowker, J.M.; Lim, S.H.; Cordell, H.K. [and others]. 2008. Wildland fire, risk, and recovery: results of a national survey with regional and racial perspectives. Journal of Forestry. 106(5): 268-276.

Bowker, J.M.; Murphy, D.; Cordell, H.K. [and others]. 2006. Wilderness and primitive area recreation participation and consumption: an examination of demographic and spatial factors. Journal of Agricultural and Applied Economics. 38(2): 317-326.

Horton, N.J.; Lipsitz, S.R. 2001. Multiple imputation in practice: comparison of software packages for regression models with missing variables. The American Statistician. 55: 244-254.

Johnson, C.Y.; Bowker, J.M.; Cordell, H.K. 2001. Outdoor recreation constraints: an examination of race, gender, and rural dwelling. Southern Rural Sociology. 17: 111-133.

Johnson, C.Y.; Bowker, J.M.; Bergstrom, J.C.; Cordell, H.K. 2004. Wilderness values in America: does immigrant status or ethnicity matter? Society and Natural Resources. 17: 611-628.

Little, R.J.A. 1988. A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association. 83: 1198-1202.

Nagelkerke, N.J.D. 1991. A note on a general definition of the coefficient of determination. Biometrica. 78: 691-692.

Raghunathan, T.E.; Lepkowski, J.M.; van Hoewyk, J.; Solenberger, P. 2001. A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology. 27: 85-95.

Rao, J.N.K.; Scott, A.J. 1987. On simple adjustments to chi-square tests with sample survey data. The Annals of Statistics. 15: 385-397.

Rubin, D.B. 1987. Multiple imputation for nonresponse in surveys. New York, NY: John Wiley and Sons. 258 p.

Rubin, D.B. 1996. Multiple imputation after 18+ years. Journal of the American Statistical Association. 91: 473-489.

SAS Institute Inc. 2004. SAS/STAT® 9.1 User's Guide. Cary, NC: SAS Institute Inc. 5121 p.

Schenker, N.; Raghunathan, T.E.; Chiu, P. [and others]. 2006. Multiple imputation of missing income data in the National Health Interview Survey. Journal of the American Statistical Association. 101: 924-933.

van Buuren, S.; Boshuizen, H.C.; Knook, D.L. 1999. Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in Medicine. 18: 681-694.

**Zarnoch, Stanley J.; Cordell, H. Ken; Betz, Carter J.; Bergstrom, John C.** 2010. Multiple imputation: an application to income nonresponse in the National Survey on Recreation and the Environment. Res. Pap. SRS–49. Asheville, NC: U.S. Department of Agriculture Forest Service, Southern Research Station. 15 p.

Multiple imputation is used to create values for missing family income data in the National Survey on Recreation and the Environment. We present an overview of the survey and a description of the missingness pattern for family income and other key variables. We create a logistic model for the multiple imputation process and to impute data sets for family income. We compare results between estimates of the income distribution based on no imputation, single imputation, and multiple imputation. Although the imputation methodology has been applied to the income variable, it is transferable as a general approach to dealing with item nonresponse for other variables in this and other survey studies.

**Keywords**: MAR assumption, MCAR assumption, nonresponse bias, outdoor recreation, single imputation.


The Forest Service, United States Department of Agriculture (USDA), is dedicated to the principle of multiple use management of the Nation's forest resources for sustained yields of wood, water, forage, wildlife, and recreation. Through forestry research, cooperation with the States and private forest owners, and management of the National Forests and National Grasslands, it strives—as directed by Congress—to provide increasingly greater service to a growing Nation.