Journal of Modern Applied Statistical Methods

Volume 6 | Issue 2 Article 27

11-1-2007

Longitudinal Evaluation of Estimates in an Esablishment Survey After Ration Imputation

Adriana Pérez
University of Louisville

Longitudinal Evaluation of Estimates in an Esablishment Survey After Ration Imputation

Cover Page Footnote

This research was supported by the American Statistical Association (ASA) and the Division of Science Resources Statistics (SRS) of the National Science Foundation through the ASA/SRS-NSF Research Program to the first author in 2005. The ASA/SRS-NSF Research Program fosters collaborative and interdisciplinary research efforts. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The use of NSF data does not imply endorsement of the research methods or conclusions contained in this article. The author would like to thank Jeri M. Mulrow for her collaboration during the planning and implementation of this research. It is important to thank Jeri M. Mulrow and Dr. Lemuel A. Moye for helpful comments in a previous version of this manuscript. I also wish to thank Maria L. Fernandez and Gloria L. Sanchez, for data management and macros programming.

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 630-644

Longitudinal Evaluation of Estimates in an Establishment Survey After Ration Imputation

Adriana Pérez
University of Louisville

Researchers evaluated a ratio imputation technique used at the US Survey of Graduate Students and Postdoctorates in Science and Engineering, which is an annually conducted cross-sectional establishment survey. Standardized bias was used, mean square error and relative bias to appraise this imputation method on point and variance estimates via simulations.

Key words: Total estimate, variance estimation, establishment data, nonresponse, simulations.

Introduction

Nonresponse in establishment surveys is an ongoing problem (Kovar & Whitridge, 1995). The problem of nonresponse affects estimates of survey statistics (Little & Rubin DB, 2002; Rubin, 1987; Kovar, et al., 1995; Ruggles & Joint Economic Committee, 2006; Groves, Dillman, Eltinge, & Little, 2002; Groves, et al., 2004). Many imputation methods used in social, demographic and health science settings have been applied within the economic survey framework and very little information is known about the effect of item nonresponse in establishment surveys (Kovar, et al., 1995; Judkins , 2000; West, Butani, & Witt, 1993). There has been a focus on procedures for reducing measurement error, improving sampling strategies (Lee & Croal, 1989), improving estimators (Sirken & Shimizu, 1999), improving response rates (Chun, 1997), response selection, survey coordination, longitudinal analysis(Ruggles, et al., 2006)' (Schenker, Treiman, & Weidman, 1988; Heeringa & Lepkowski, 1986), or empirical evaluation of imputation methods (West, et al., 1993; Krenzke, Montaquila, & Mohadjer, 2000; Mueller & Butani, 1995) in establishment surveys.

Many imputation methods are available in the literature. Usually, once a dataset has been imputed, analyses are performed treating the imputed values as observed data. This type of analysis could be misleading because variances and covariances may be underestimated (Kovar, et al., 1995). In this article, the effectiveness of a particular ratio imputation method when applied to an item-nonresponse from an establishment survey including a longitudinal perspective on point and variance estimates is evaluated.

There are a variety of techniques for variance estimation for complex surveys (Wolter, 1985) and few of them incorporate the effect of imputation in their estimation (Shao & Sitter, 1996; Shao & Steel, 1999; Shao, 2002). Most of the time imputation methods in a survey are implemented without theoretical development of the methods (Shao, 2002). Simulation studies make it possible to evaluate and compare estimation techniques in national surveys in any country (U.S.Department of Education.National Center for Education Statistics., 2001). Pseudo-universes from survey

Adriana Pérez is Associate Professor of Biostatistics at the University of Louisville, School of Public Health and Information Sciences, Department of Bioinformatics and Biostatistics. This research was conducted when Dr. Pérez was at the University of Texas Health Science Center at Houston. Her research interests are in statistical methods to handle missing data; statistical methods for epidemiology research; design, conduct and analysis of multi-center clinical trials; sample size estimation and modeling strategies. Email her at adriana.perez@louisville.edu

data can be used instead of national universes (i.e., census data) which are not usually available for simulation studies. Pseudo universes permit a comparison of techniques and sample according to a plan of interest, maintaining the distributions of the variables of interest. Simulations from a pseudo universe can provide estimates of interest and give detailed insight of the estimator performance.

It is the researcher's interest to study the effect on the point and variance estimates of the current imputation plan conducted in the Graduate Students and Postdoctorates in Science and Engineering (GSS)(NSF-NIH, 2005). One of the challenging aspects of any simulation is the creation of an artificial population similar to the one investigated. There are two approaches to create a finite population universe(Katzoff, Jones, & Curtin, 1988; Bernaards, Belin, & Schafer, 2006; Schafer, et al., 1996). One is to create pseudo-random values from an actual multivariate probability model, also known as a hypothetical population. The second is to use an actual large data set to reflect the target population and to define population parameters of interest, also known as a pseudo-universe. Use of a specific probability model is a limitation in the creation of a hypothetical population(Schafer, et al., 1996). Therefore, a pseudo universe was created to impose realistic missing data patterns.

The following describes the generation of the pseudo universe and simulations which allow: (i) appraise the longitudinal missing data patterns in GSS between 1999-2001; (ii) evaluate the effect of current imputation methods in this survey on estimates for different missing data mechanism assumptions in GSS; (iii) assess precision and accuracy measures in the total, and corresponding variance estimates in GSS. In following sections, the GSS survey will be described, the current imputation method, the methodology to evaluate the effect of the current imputation method and the results and conclusions, respectively.

The Survey of Graduate Students and Postdoctorates in Science and Engineering (GSS)

One of the current surveys conducted at the Division of Science Resources Statistics (SRS) of the National Science Foundation (NSF) is the NSF-NIH (National Institutes of Health) survey of Graduate Students and Postdoctorates in Science and Engineering (GSS) (NSF-NIH, 2005). This survey (i) measures academic department level information on all U.S. institutions offering graduate programs (masters or PhD degrees) in science, engineering, or health selected field; (ii) provides a description of graduate science and engineering (S&E) student's enrollment in US institutions; and (iii) assesses trends in financial support patterns and shifts in graduate enrollment and postdoctoral appointments.

This cross-sectional establishment survey is conducted annually (NSF-NIH, 2005). Reports from this survey are presented in current year and historical data, setting up a longitudinal structure (National Science Foundation & Division of Science Resources Statistics (SRS), 2005). Total estimates for domains and sub-domains are reported (National Science Foundation & Division of Science Resources Statistics, 2006). Each year, a ratio imputation technique is used to handle item nonresponse based on inflator/deflator factors (NSF-NIH, 2005). For a particular year, these inflator/deflator factors are computed from the current year observed data in combination with previous year observed and imputed data (Morgan M & ORC Macro, 2004). Replacing missing data in the current year with previous year data is an imputation method known in longitudinal human population studies as the last observation carry-forward (LOCF). This imputation method is modified in the GSS by the use of inflator/deflator factors as adjustments when replacing current cycle missing data with adjusted previous cycle data.

Simulations conducted with LOCF, in longitudinal human population studies, indicates that LOCF produces biased estimates for all three types of missing data mechanisms (Missing completely at random (MCAR), Missing at random (MAR) or Missing not at random (MNAR)) and LOCF produces the smallest standard errors that are biased downward (Gadbury, Coffey, & Allison, 2005). For these reasons, evaluation of the current GSS imputation plan is needed.

Imputation at the Graduate Student Survey

The department within an academic institution is the unit of interest of this survey for imputation purposes. This imputation methodology is presented for four variables used in this research only, but can be generalizable to the rest of the variables within this survey.

Creation of inflator/deflator factors

Departments that provided full or partial information about total full-time students, total part-time students, total postdoctorates and total other non-faculty research staff are used for creation of these factors. Specifically, in this study, total full-time students and total part-time students were used. Inflator/deflator factors are computed by highest institutional degree level (doctorate and master's) and by department type (e.g. Biology, Physics, etc.). For a particular variable of interest (Y_k), its sum is computed by institutional highest degree level and department type. Then factors are computed by dividing the sum of the variable from the current (t) year by the corresponding sum of the variable from the previous year (t-1). These inflator/deflator factors ($\hat{\psi}_{k_t}$) in mathematical terms are calculated for the k^th^ variable and year t.

$$\hat{\psi}_{k_t} = \sum_{j=1}^{r} Y_{jk_t} \Big/ \sum_{j=1}^{r} Y_{jk_{(t-1)}} \qquad (1)$$

r identifies the maximum number of departments in the same institutional degree level and departmental type that provided a variable value Y_k in both years t and $t-1$. Any computed factor less than 0.85 or greater than 1.15 is set to 1 for imputation purposes. In mathematical terms:

$$\hat{\phi}_{k_t} = \begin{cases} \hat{\psi}_{k_t} & \text{if } 0.85 < \hat{\psi}_{k_t} < 1.15 \\ 1 & \text{otherwise} \end{cases} \qquad (2)$$

Using inflator/deflator factors to impute total full- time students all sources of funding and total part-time students of all races

Departments with missing information in total full- time students and/or total part-time students are imputed using equation 3. The imputation value for a particular variable in the current year is obtained by applying $\hat{\phi}_{k_t}$ to previous year information for that variable. This is done at each department institutional level (i.e., MS or PhD) and department type (i.e., Biology, Physics, etc).

$$\hat{Y}_{I(ik_t)} = \hat{\phi}_{k_t} * Y_{ik_{(t-1)}} \qquad (3)$$

i identifies a particular department, k identifies the variable, t identifies the year, $\hat{\phi}_{k_t}$ identifies the inflator/deflator factors, $Y_{ik_{(t-1)}}$ is the k^th^ variable value for department i, in year $t-1$; and $\hat{Y}_{I(ik_t)}$ is the imputed value of the k^th^ variable for department i at year t.

Subsequently, imputed values for total full-time students (from equation 3) are used to impute variables regarding full-time students by: source and mechanism of support. Similarly, imputed values for total part-time students (from Equation 3) are used to impute variables regarding number of part-time students by sex and their distribution by US nationals/permanent residents or foreign students.

Using inflator/deflator factors to impute total part-time students

The imputed value for the total of female part-time students is computed using the same percentage as reported in the previous year on the imputed value from the total part-time students in the current year. Equation 4 shows this in mathematical terms.

$$\hat{Y}_{I(ij_t)} = \hat{Y}_{i_t} \left(Y_{ij_{(t-1)}} \Big/ Y_{i_{(t-1)}} \right) \qquad (4)$$

Where i identifies a particular department, t identifies the year, j identifies women, $\hat{Y}_{i_t}$ represents the observed or imputed value of the total part-time students enrolled for a particular department i at year t, $Y_{ij_{(t-1)}}$ represents the observed value of the total part-time women students for year $t-1$ at particular department i, $Y_{i_{(t-1)}}$ represents the observed value of the total part-time students for year $t-1$ at a particular department i, and $\hat{Y}_{I(ijt)}$ represents the imputed value of the total part-time women at year t at particular department i.

The imputed value for the total of male part-time students is calculated as the difference between the total part-time students in the current year (observed or imputed) and the observed or imputed value for the total of female part-time students.

Methodology

The purpose is to evaluate the longitudinal effect of imputation on estimates in the GSS, data from the years 1998-2002 (the most recent data through 2005). The GSS survey in 1998 contained 639 variables and 11686 departments and in 2002 contained 639 variables and 12126 departments. Overall, 15379 departments reported any information on the GSS data from the years 1998--2002. The first four variables imputed in this survey were selected for analysis in this research: total full-time graduate students all sources of support, total part-time students of all races, total part-time male students of all races and total part-time female students of all races. To evaluate the effect of the imputation method within this survey a simulation study with a pseudo universe from this survey was conducted.

Generation of the pseudo universe 1998-2002

A dataset called "Observed 1998-2002" which mainly excluded departments with unit nonresponse between years 1998-2002 was created. If a department reported any missing value for any of the variables of interest in year 1998 and 2002 they were excluded. This is because stable departments were to be used, to exclude new programs (i.e., if a department created a new master or doctoral program in 2002 then previous years would not have reported any information and missing values would appear in the longitudinal structure), and to exclude non current programs. If departments provided information in years 1998 and 2002 this indicated continuity of the master's or PhD degree program at that institution. In summary, one department was excluded because it did not report the type of academic institution (neither school under which this department was associated, nor public or private nor which institutional highest degree is granted). Departments with unit nonresponse were excluded for each year as follows: 3693 within 1998, 936 within 1999, 514 within 2000, 755 within 2001 respectively and 610 within 2002. It was assumed that these departments with unit nonresponse were not stable. Furthermore, the study excluded 328 departments without students enrolled either full-time or part-time in 1998 in any of the four variables of interest, which indicated historically unstable enrollment in that program. This dataset Observed 1998-2002 contained 8542 out of 15379 departments with item nonresponse between the years 1999 and 2001. Using this dataset the researchers generated the longitudinal distributional patterns of missing data in years 1999-2001.

After this, the researchers generated a pseudo universe from this survey by removing any department with missing data in our variables of interest from years 1999-2001. Researchers excluded 685 departments because they did not report full time students for at least one of these years. Forty-five departments that did not report part-time students for at least one of these years were deleted. Furthermore, 127 departments that did not report part time male students for at least one of these years were excluded.

This complete dataset was called and used as Pseudo Universe 1998-2002 and contains 7685 departments with complete information on all these variables. This pseudo universe was used to develop and evaluate the imputation methods used in GSS for the variable totals and their corresponding variability measures. Total estimates coming from this pseudo-universe were treated as parameter values from this pseudo universe. This is notated θ_{k_t} as the total estimate of the k^{th}-variable of interest for years 1999 to 2001. These parameter values were used for comparison purposes in evaluation the GSS imputation methods.

Simulation of mechanisms of missingness

Two missingness mechanisms to evaluate the imputation methods at GSS were

Table 1: Actual percentages of missing values in dataset “Observed 1998-2002”

Year	N	Full time students all sources of support	Part Time students of all races	Part time male students of all races	Part time female students of all races
1999	11832	1.49	1.58	3.26	3.25
2000	11899	1.58	1.60	1.99	1.99
2001	11968	3.53	3.77	4.12	4.12

explored. The first approach was to create an MCAR mechanism.

Actual percentages of missing values were imposed within “Pseudo Universe 1998-2002” on within Pseudo Universe 1998-2002 on each Y_{k_t} independently of any variable in the system. Table 1 illustrates the “Actual” percentage of missing data observed in years 1999—2001 for the four variables of interest in this survey and these percentages were used for creating the MCAR mechanism for evaluation purposes. Our “MCAR dataset” contains these “Actual” percentages imposed randomly as missing. As you may notice these percentages are not high and it will be desired to evaluate the imputation method with this low percentages of missingness.

It was assumed that the occurrence of missing values at the GSS survey is MAR. Under this assumption, the second approach was to impose the Actual percentages of missing values with the same longitudinal distributional patterns of missing data in years 1999-2001 from Observed 1998-2002 within Pseudo Universe 1998-2002. Table 2 shows the observed longitudinal patterns of missing values for these variables, where 0 represents data was missing and 1 represents data was observed.

For the purposes of understanding the effect of the imputation method with increased percentages of missing values, in simulations, the researchers increased these observed longitudinal distributional patterns of missing values from the Observed 1998-2002 in 25%, 50%, 75% and 100% (data not shown) within Pseudo Universe 1998-2002.

Parameter estimation

These datasets, with imposed missing values, can be used to examine many quantities of interest. The total and its corresponding variance estimate were examined for each year. Many other parameters were included in simulations but are not reported for brevity and the research primarily presents the results under the MAR mechanism. Each one of these missingness mechanisms were replicated one thousand times.

Applying the Imputation Method

Inflator/deflator factors for year 1999 were computed using the observed data from the 1998 Pseudo Universe 1998-2002. The ratio imputation methods described in equations 3 and 4 were applied for missing values in year 1999. Then, an imputed and complete 1999 dataset was reached. Similarly, the researchers continued to generate the inflator/deflator factors and to impute missing values in years 2000 and 2001. This procedure produced an Observed and imputed longitudinal 1999-2001 dataset. Cross-sectional 1999-2001 total estimates and their corresponding variances were computed. Estimates after imputation are notated as $\hat{\theta}_{AI_{k_t}}$ for each k th variable on years 1999--2001.

Evaluation criteria

The performance of the GSS imputation method by the following quantities in years 1999-2001 were evaluated. First, the bias of the total and the variance estimates after imputation of the simulations are described in Equations 5 and 6, respectively.

Table 2. Percentages missing values for each pattern in dataset “Observed 1998-2002”.

Pattern	Full time students all sources of support			Part Time students of all races			Part time male students of all races			Part time female students of all races			%
	1999	2000	2001	1999	2000	2001	1999	2000	2001	1999	2000	2001	
1	0	0	0	0	0	0	0	0	0	0	0	0	0.05
2	0	0	1	0	0	1	0	0	1	0	0	1	0.08
3	0	1	0	0	1	0	0	1	0	0	1	0	0.01
4	0	1	1	0	0	1	0	0	1	0	0	1	0.01
5	0	1	1	0	1	1	0	0	1	0	0	1	0.01
6	0	1	1	0	1	1	0	1	1	0	1	1	0.89
7	0	1	1	0	1	1	0	1	1	1	1	1	0.01
8	0	1	1	1	1	1	1	1	1	1	1	1	0.06
9	1	0	0	1	0	0	1	0	0	1	0	0	0.52
10	1	0	0	1	0	1	1	0	1	1	0	1	0.01
11	1	0	1	1	0	0	1	0	0	1	0	0	0.01
12	1	0	1	1	0	1	1	0	1	1	0	1	0.89
13	1	0	1	1	1	1	1	1	1	1	1	1	0.01
14	1	1	0	1	1	0	0	1	0	0	1	0	0.02
15	1	1	0	1	0	0	1	0	0	1	0	0	0.01
16	1	1	0	1	1	0	1	0	0	1	0	0	0.04
17	1	1	0	1	1	0	1	1	0	1	1	0	2.60
18	1	1	0	1	1	1	1	1	1	1	1	1	0.01
19	1	1	1	0	1	1	0	1	1	0	1	1	0.12
20	1	1	1	1	1	1	0	0	0	0	0	0	0.12
21	1	1	1	1	1	1	0	0	1	0	0	1	0.05
22	1	1	1	1	1	1	0	1	0	0	1	0	0.01
23	1	1	1	1	1	1	0	1	1	0	1	1	0.74
24	1	1	1	1	1	1	1	0	0	1	0	0	0.01
25	1	1	1	1	1	1	1	0	1	1	0	1	0.21
26	1	1	1	1	1	0	1	1	0	1	1	0	0.07
27	1	1	1	1	1	1	1	1	0	1	1	0	0.18
28	1	1	1	1	1	1	1	1	1	1	1	1	93.26

$$\text{Bias of the total estimate}_{k_t} = E\left(\hat{\theta}_{AI_{k_t}}\right) - \theta_{k_t} \tag{5}$$

$$\text{Bias\_variance\_total}_{k_t} = E\left(\text{Var}(\hat{\theta}_{AI_{k_t}})\right) - \sigma^2_{k_t} \tag{6}$$

where $\sigma^2_{k_t}$ identifies the population variance among the data [$\text{Var}(Y_{k_t})$] within the "Pseudo Universe 1998-2002" and not the variance of the mean estimates. $\text{Var}(\hat{\theta}_{AI_{k_t}})$ identifies the estimated variance after imputation. Second, given that the raw bias can be misleading the standardized bias of the total estimate using equation 7 was computed. A standardized bias of less of 50% in both directions should be considered practically insignificant. Third, the mean square error (MSE) for the total and the variance estimates are described in equations 8 and 9.

$$\text{Standardized Bias}_{k_t} = 100 * \frac{E\left(\hat{\theta}_{AI_{k_t}}\right) - \theta_{k_t}}{SE(\hat{\theta}_{AI_{K_t}})} \tag{7}$$

$$\text{MSE(Total)} = \left(\frac{1}{1000}\right)\sum_{m=1}^{1000}(\hat{\theta}_{AI_{k_t}} - \theta_{k_t})^2 \tag{8}$$

$$\text{MSE(Var)} = \left(\frac{1}{1000}\right)\sum_{m=1}^{1000}\left(\text{Var}(\hat{\theta}_{AI_{k_t}}) - \sigma^2_{k_t}\right)^2 \tag{9}$$

Fourth, the average relative bias of the total and the variance estimates are described in equations 10 and 11. These average relative biases measure the average magnitude of over or under estimation of the imputation method compared with the true value. Finally, the average relative stability of the variance is described in equation 12.

Relative Bias of the total =

$$\left(\frac{1}{1000}\right)\sum_{m=1}^{1000}\frac{(\hat{\theta}_{AI_{k_t}} - \theta_{k_t})}{\theta_{k_t}} \tag{10}$$

Relative Bias of the variance =

$$\left(\frac{1}{10000}\right)\sum_{m=1}^{1000}\left(\frac{\text{Var}(\hat{\theta}_{AI_{k_t}}) - \sigma^2_{k_t}}{\sigma^2_{k_t}}\right) \tag{11}$$

Relative Stability =

$$\frac{\left[\left(\frac{1}{1000}\right)\sum_{m=1}^{1000}\left(\text{Var}(\hat{\theta}_{AI_{k_t}}) - \text{MSE(Var)}\right)^2\right]^{1/2}}{\text{MSE(Var)}} \tag{12}$$

Results

Table 3 presents the results of the 1000 simulations under MCAR mechanism. The current imputation method underestimates total full-time students and total part-time female students and overestimates part-time students and part-time male students under this mechanism. The underestimation or overestimation of these variables increased yearly from 1999 to 2001. The standardized biases were larger than 50% for many of the variables of interest.

Results of simulations under the MAR mechanism are presented in Tables 4-7. Table 4 shows the results from the evaluation criteria for the imputation method on full-time students all sources of funding. The relative bias of the total estimate of full-time students indicates a 10% underestimation for years 2000 and 2001 with the current amount of missing values. If the amount of missing values increases then this underestimation increased up to 20% for year 2001. It is interesting to note that this imputation method would overestimate the total estimate of full-time students by 40% if the current patterns of missing values were increased by 100% for the year 2000.

Results from the relative bias of the variance of the total estimate of full-time students across the years indicates overestimation between 10% and 30% for year 1999 for increasing percentages of missing values. This overestimation is also observed for year 2001 with a range of 20% to 70%.

Table 3. Results from 1000 replicates under MCAR

Year	**1999**	**2000**	**2001**
Bias of the total			
Full time students all sources of support	-878	-665	-3,629
Part Time students of all races	1,177	2,115	2,347
Part time male students of all races	617	1,439	2,192
Part time female students of all races	-276	-478	-1,250
Bias of the variance			
Full time students all sources of support	-6	-1	-78
Part Time students of all races	6	5	4
Part time male students of all races	1	4	8
Part time female students of all races	2	1	16
MSE of the total			
Full time students all sources of support	818	1,372	9,166
Part Time students of all races	98	267	329
Part time male students of all races	3	34	2332
Part time female students of all races	7	22	13,175
MSE of the variance			
Full time students all sources of support	1.11E+06	1.11E+06	1.45E+07
Part Time students of all races	1.47E+06	4.68E+06	5.78E+06
Part time male students of all races	3.94E+05	2.12E+06	4.90E+06
Part time female students of all races	9.09E+04	2.67E+05	1.72E+06
Standardized Bias of the variance			
Full time students all sources of support	-149.8	-81.5	-313.9
Part Time students of all races	415.5	460.8	454.0
Part time male students of all races	526.5	653.4	734.2
Part time female students of all races	-227.5	-244.2	-313.9

Table 4. Results from 1000 replicates under MAR for full time students

Year	Actual%	25%	50%	75%
Bias of the total				
1999	-10812	-15523	-15156	-19867
2000	-12994	-22657	-13833	-19737
2001	-29314	-46229	-43331	-28207
Bias of the variance				
1999	9.2E+09	2.1E+10	1.2E+10	3.0E+10
2000	-2.1E+10	-3.1E+10	-2.4E+10	-2.8E+10
2001	5.4E+10	2.7E+10	7.2E+10	1.0E+11
Standardized bias				
1999	-11.9	-15.3	-14.1	-16.6
2000	-10.8	-16.5	-9.1	-12.3
2001	-15.3	-21.1	-18.9	-11.3
MSE of the total				
1999	8.4E+06	1.0E+07	1.2E+07	1.5E+07
2000	1.5E+07	1.9E+07	2.3E+07	2.6E+07
2001	3.8E+07	5.0E+07	5.4E+07	6.3E+07
MSE of the variance				
1999	5.3E+19	8.4E+19	7.6E+19	1.1E+20
2000	8.8E+19	1.2E+20	1.4E+20	1.5E+20
2001	4.6E+20	5.6E+20	6.2E+20	7.6E+20
Relative Bias of the total				
1999	0.0	-0.1	-0.1	-0.1
2000	-0.1	-0.1	-0.1	-0.1
2001	-0.1	-0.2	-0.2	-0.1
Relative Bias of the Variance				
1999	0.1	0.2	0.1	0.2
2000	-0.2	-0.2	-0.2	-0.2
2001	0.3	0.2	0.5	0.7
Relative Stability of the variance				
1999	1.0	0.6	0.7	0.5
2000	0.6	0.5	0.4	0.4
2001	0.1	0.1	0.1	0.1

Table 5. Results from 1000 replicates under MAR for part time students

Year	Actual%	25%	50%	75%
Bias of the total				
1999	877	2366	4199	5853
2000	17076	26118	26247	37098
2001	-9815	-20641	-17052	-27822
Bias of the variance				
1999	-7.1E+09	-4.7E+09	-4.8E+09	3.2E+09
2000	-2.7E+10	-2.2E+10	-2.6E+10	-1.7E+10
2001	-1.9E+10	-3.0E+10	-1.4E+10	-1.9E+10
Standardized bias				
1999	1.3	3.0	4.8	6.2
2000	16.5	23.7	20.9	28.0
2001	-5.8	-11.2	-8.3	-12.8
MSE of the total				
1999	4.5E+06	6.3E+06	7.7E+06	9.0E+06
2000	1.1E+07	1.3E+07	1.6E+07	1.9E+07
2001	2.8E+07	3.4E+07	4.3E+07	4.8E+07
MSE of the variance				
1999	1.6E+19	2.0E+19	2.2E+19	3.0E+19
2000	5.9E+19	6.4E+19	8.4E+19	7.3E+19
2001	1.8E+20	2.3E+20	2.9E+20	3.2E+20
Relative Bias of the total				
1999	0.0	0.0	0.0	0.1
2000	0.2	0.3	0.3	0.4
2001	-0.1	-0.2	-0.2	-0.3
Relative Bias of the Variance				
1999	0.1	0.2	0.1	0.2
2000	-0.2	-0.2	-0.2	-0.2
2001	0.3	0.2	0.5	0.5
Relative Stability of the variance				
1999	1.0	0.8	0.7	0.5
2000	0.3	0.3	0.2	0.2
2001	0.1	0.1	0.1	0.1

Results from the relative bias of the variance in year 2000 indicate that this imputation method underestimates the variance of the total estimate of full-time students from 20% to 40% depending on the amount of missingness. The MSE of the total and the variance of full-time students using the current imputation method at GSS is large. The MSE of the variance increases for each year of increase and as expected if the percentage of missing values increases then the MSE of the variance will increase. The average relative stability of the variance of the total estimate of full-time students decreases noticeably for each one year increase. This behavior is consistently observed across increasing percentages of missing values.

Table 5 shows the results from the evaluation criteria for the imputation method on part-time students of all races. The relative bias of the total estimate of part-time students indicates a 20% overestimation for year 2000 and a 10% underestimation for year 2001 with the current amount of missing values. If the amount of missing values increases then this

Table 6. Results from 1000 replicates under MAR for part time male students

Year	Actual%	25%	50%	75%
Bias of the total				
1999	40137	51315	61580	75489
2000	53259	66665	77239	92581
2001	40137	51358	66098	76572
Bias of the variance				
1999	9.1E+09	1.2E+10	1.4E+10	2.3E+10
2000	9.1E+09	1.4E+10	1.9E+10	2.8E+10
2001	8.5E+09	8.2E+09	2.0E+10	2.6E+10
Standardized bias				
1999	85.5	94.0	102.6	113.4
2000	75.8	91.0	92.6	107.8
2001	40.0	42.8	49.0	52.4
MSE of the total				
1999	3.8E+06	5.6E+06	7.4E+06	1.0E+07
2000	7.8E+06	9.8E+06	1.3E+07	1.6E+07
2001	1.4E+07	1.7E+07	2.3E+07	2.7E+07
MSE of the variance				
1999	3.5E+18	5.1E+18	5.6E+18	7.8E+18
2000	1.8E+19	1.7E+19	2.4E+19	2.0E+19
2001	4.1E+19	5.5E+19	6.6E+19	7.7E+19
Relative Bias of the total				
1999	0.8	1.0	1.3	1.5
2000	1.1	1.4	1.6	2.0
2001	0.9	1.1	1.4	1.6
Relative Bias of the Variance				
1999	0.5	0.7	0.9	1.4
2000	0.6	0.9	1.1	1.7
2001	0.5	0.5	1.2	1.6
Relative Stability of the variance				
1999	1.0	0.7	0.6	0.5
2000	0.2	0.2	0.1	0.2
2001	0.1	0.1	0.1	0.0

overestimation increases up to 40% for year 2000 and the underestimation will decrease by at least 20% for year 2001. Results from the relative bias of the variance of the total estimate of part-time students across years indicates increased underestimation for increased year and this behavior seems to follow a U shape for increasing percentages of missing values. Findings about the MSE for the total and variance of full-time students are similar than for part-time students as well as regarding the average relative stability of the variance.

Table 6 shows the results from the evaluation criteria for the imputation method on part-time male students of all races. The relative bias of the total estimate of part-time male students with the current amount of missing values indicates 80%, 110% and 90% overestimation for years 1999, 2000 and 2001, respectively.

Table 7. Results from 1000 replicates under MAR for part time female students

Year	Actual%	25%	50%	75%
Bias of the total				
1999	-39260	-48949	-57381	-69636
2000	-36183	-40547	-50992	-55483
2001	-829684	-847999	-859150	-880394
Bias of the variance				
1999	-4.8E+09	-4.8E+09	-3.9E+09	-4.9E+09
2000	-1.8E+10	-1.9E+10	-2.5E+10	-2.3E+10
2001	-2.3E+10	-2.9E+10	-3.3E+10	-4.3E+10
Standardized bias				
1999	-87.2	-94.6	-105.9	-114.1
2000	-61.6	-62.5	-70.1	-71.0
2001	-924.3	-842.8	-768.8	-744.2
MSE of the total				
1999	3.6E+06	5.1E+06	6.2E+06	8.6E+06
2000	4.8E+06	5.9E+06	7.9E+06	9.2E+06
2001	7.0E+08	7.3E+08	7.5E+08	7.9E+08
MSE of the variance				
1999	2.9E+18	3.2E+18	3.6E+18	4.7E+18
2000	4.1E+18	5.3E+18	7.5E+18	8.8E+18
2001	1.1E+19	1.4E+19	1.9E+19	2.1E+19
Relative Bias of the total				
1999	-0.9	-1.1	-1.2	-1.5
2000	-0.8	-0.9	-1.1	-1.2
2001	-18.1	-18.5	-18.8	-19.2
Relative Bias of the Variance				
1999	-0.3	-0.3	-0.2	-0.3
2000	-1.1	-1.1	-1.5	-1.4
2001	-1.3	-1.7	-1.9	-2.5
Relative Stability of the variance				
1999	1.0	0.9	0.8	0.6
2000	0.7	0.6	0.4	0.3
2001	0.3	0.2	0.2	0.1

As expected if the amount of missing values increases then this overestimation increases. Results from the relative bias of the variance of the total estimate of part-time male students across years indicates overestimation above 50% and increases for increasing percentages of missing values. Findings about the MSE for the total and variance of full-time students are equal for part-time male students and the average time male students as well as regarding the average relative stability of the variance.

Table 7 shows the results from the evaluation criteria for the imputation method on part-time female students of all races. The relative bias of the total estimate of part-time female students, with the current missing values, indicates a 90%, 80% and 1813% underestimation for 1999, 2000, and 2001.

If the amount of missing values increases then this underestimation increases as well. Results from the relative bias of the variance of the total estimate of part-time female students across years indicates underestimation between 20% and 30% for year 1999 for increasing percentages of missing values. This underestimation is also observed for years 2001 and 2002 with a range from 110% to 270%. Findings about the MSE for the total and variance of full-time students are equal for part-time female students as well as regarding the average relative stability of the variance.

Conclusion

Overall, the bias and the MSE of the total and the variance estimates are not acceptable under the MCAR mechanism. Our findings under MCAR in this establishment survey are consistent with the literature in human populations where you will expect a higher underestimation or overestimation for increasing percentage of missing values in a variable including the increase as a year passes by.

Overall, the bias of the total estimates for full-time students and part-time students are acceptable under the MAR mechanism. This is because although the estimates across years and for different percentages of increase of current missing values are biased, the standardized biases are less than -50% which means that this bias is practically insignificant. On the contrary, the bias of the total estimates for part-time male and female students are not acceptable under the MAR mechanism using similar criteria of the standardized bias which surpass 50% in either direction for any percentage increase of missing values.

The results of overestimation for 1999 and 2001 using the relative bias of the variance of the total estimate of full-time students and its underestimation in 2000 with this imputation method are in agreement with previous descriptions of variance estimate behaviors after imputation in human population surveys, where imputation methods underestimate or overestimate depending on the variability of the variable. Most of the time it is expected to provide an underestimation of this variance estimate and this is shown in many of the variables chosen for this research.

The MSE incorporates two components, one measuring the variability of the estimator (precision) and the other measuring its bias (accuracy). Overall, the estimators generated with the current imputation method in GSS do not have good MSE properties because they do not have small combined variance and bias.

The findings regarding the variance estimates using the current imputation methods in this establishment survey for the variables chosen are in agreement with findings with many imputation methods for human population surveys where priority and challenges need to be overcome for improving variance estimates in surveys. The noticeable decrease in the average relative stability of the variance of the total estimates of the variables of interest warrants consideration.

There were many limitations to this study. The chosen pseudo universe represents a best case scenario where departments are fully compliant and provided full information. Furthermore, sampling did not come from this finite population to test the imputation method in full when a sample is selected instead of using the entire population. The entire population was used, which is the best case scenario, being fully efficient in the scenarios regarding the imputation method. It is expected that by selecting different sample sizes will provide worst results than the ones presented here. Also, a good scenario where the current percentages of missing values do not seem very high for each cross-sectional year was used. However, the findings are overwhelming in the large effects that the current GSS imputation method affects the bias of the total estimates of part-time males and females and overall variance estimates. Another limitation is that this study only handles the issue of item-nonresponse when unit non-response was excluded from this research. The results limitation as a best case scenario warrants consideration because worse results would be expected under worse conditions than those presented here.

Currently NSF publishes total estimates from this survey without reporting any variance estimate. Careful attention is needed for those variables where standardized biases are larger than -50% as well as how to improve the stability of the variance decreasing for increasing percentages of missingness in the cross-sectional and longitudinal setting. Minor discrepancies were observed in the bias and MSE estimates when the unit of analysis is establishments instead of individuals. Further research is needed to identify statistical methods to handle the missing data from this survey and to evaluate this method under a missing not at random mechanism.

Acknowledgments

This research was supported by the American Statistical Association (ASA) and the Division of Science Resources Statistics (SRS) of the National Science Foundation through the ASA/SRS-NSF Research Program to the first author in 2005. The ASA/SRS-NSF Research Program fosters collaborative and interdisciplinary research efforts. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The use of NSF data does not imply endorsement of the research methods or conclusions contained in this article. The author would like to thank Jeri M. Mulrow for her collaboration during the planning and implementation of this research. It is important to thank Jeri M. Mulrow and Dr. Lemuel A. Moye for helpful comments in a previous version of this manuscript. I also wish to thank Maria L. Fernandez and Gloria L. Sanchez, for data management and macros programming.

References

Bernaards, C. A., Belin, T. R., & Schafer, J. L. (2006). Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Stat.Med, Epub ahead of print*.

Chun YI (1997). Nonresponse follow-up in establishment surveys: A split-half experiment. In V. 2. Alexandria (Ed.), (pp. 988-993). American Statistical Association: Proceedings of the Section on Survey Research Methods.

Gadbury G.L, Coffey C.S, & Allison D.B. (2005). Modern statistical methods for handling repeated measurements in obesity trial data: beyond LOCF. Obesity Reviews 4[3], 175-184.

Groves RM, Dillman DA, Eltinge JL, & Little RJA (2002). *Survey Nonresponse*. New York: John Wiley & Sons.

Groves RM, Fowlwe F.J.Jr, Couper

M.P, Lepkowski J.M, Singer E, & Tourangeau R (2004). *Survey Methodology*. (1 ed.) New Jersey: John Wiley & Sons, Inc.

Heeringa SG & Lepkowski JM (1986). Longitudinal imputation for the SIPP. In (pp. 206-210). Proceedings of the Survey Research Methods Section: American Statistical Association.

Judkins DR (2000). Discussion. Session 44: New Developments in Imputation of Business Survey Data. In Alexandria, VA 22314: American Statistical Association.

Katzoff MJ, Jones GK, & Curtin LR (1988). A general system for the empirical evaluation of statsitical methods for data from complex surveys. In (pp. 293-297). American Statistical Association: Proceedings of the Section on Survey Research Methods.

Kovar JG & Whitridge (1995). Imputation of Business Survey Data. In Cox BG, Binder DA, Chinnappa BN, Christianson A, Colledge MJ, & Kott PS (Eds.), *Business Survey Methods* (pp. 403-423). New York: John Wiley & Sons, Inc.

Krenzke T, Montaquila J, & Mohadjer L (2000). Accounting for imputation error variance for establishment surveys: an empirical evaluation. In Alexandria, VA 22314: American Statistical Association.

Lee H & Croal J (1989). A simulation study of various estimators which use auxiliary data in an establishment survey. In (pp. 336-341). Alexandria, VA 22314: American Statistical Association.

Little RJA & Rubin DB (2002). *Statistical analysis with missing data*. (Second ed.) New York: Wiley-Interscience.

Morgan M & ORC Macro. (2004). Memorandum: Imputation Plan for Fall 2002 Graduate Student Survey. Burelli J and NSF/SRS.

Mueller K & Butani S (1995). Nonreponse adjustment in certainty strata for an establishment survey. In (pp. 479-484). Proceedings of the Survey Research Methods Section: American Statistical Association.

National Science Foundation & Division of Science Resources Statistics (2006). Graduate students and Postdoctorates in Science and Engineering: Fall 2003. Detailed Statistical Tables.http://www.nsf.gov/statistics/nsf06307/tables.htm#group1 [On-line].

National Science Foundation & Division of Science Resources Statistics (SRS) (2005). Graduate Students and Postdoctorates in S&E: Fall2003.http://www.nsf.gov/statistics/gradpostdoc/ [On-line].

NSF-NIH (2005). Survey of graduate students and postdoctorates in Science and Engineering. The National Science Foundation and the National Institutes of Health [On-line]. Available:http://www.nsf.gov/statistics/survey.cfm

Rubin DB (1987). *Multiple imputation for nonresponse in surveys*. (vols. John Wiley & Sons, Inc) New York: John Wiley & Sons, Inc.

Ruggles P & Joint Economic Committee (2006). *Longitudinal Analysis of Federal Survey Data* (Rep. No. 112). US Department of Commerce. Bureau of the Census.

Schafer JL, Ezzati-Rice TM, Johnson W, Khare M, Little RJA, & Rubin DB (1996). The NHANES III multiple imputation project. In (pp. 28-37). Alexandria, VA: American Statistical Association.

Schenker N, Treiman DJ, & Weidman L (1988). Multiple imputation of industry and occupation codes for public use files. In (pp. 85-92). Proceedings of the Survey Research Methods Section: American Statistical Association.

Shao J (2002). Replication methods for variance estimation in complex surveys with imputed data. In Robert M Groves, Don A Dillman, John L Eltinge, & Robert JA Little (Eds.), *Survey Nonresponse* (pp. 303-314). New York: John Wiley & Sons.

Shao J & Sitter RR (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association, 91,* 1278-1288.

Shao J & Steel P (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association, 94,* 254-265.

Sirken M & Shimizu I (1999). The Horvitz-Thompson Estimator in population based establishment sample surveys. In (pp. 233-237). Alexandria, VA 22314: American Statistical Association.

U.S.Department of Education.National Center for Education Statistics. (2001). *A study of imputation algorithms* (Rep. No. Working paper No. 2001-17 by Ming-xiu Hu and Sameena Salvucci). Washington, DC: Project Officer, Ralph Lee.

West SA, Butani S, & Witt M (1993). Alternative Imputation Methods for Labor Type Data. In Alexandria, VA 22314: American Statistical Association.

Wolter KM (1985). *Introduction to variance estimation*. New York: Springer-Verlag.