Sensitivity of Inference under Imputation: An Empirical Study

Jeffrey M. Gonzalez[1*] John L. Eltinge[1]
1 Office of Survey Methods Research, U.S. Bureau of Labor Statistics

Abstract

Item nonresponse, a common occurrence in many surveys, happens when a respondent fails to provide a response for a survey question. Imputation models can be used to fill in the item missing information with plausible values. These models are built on assumptions about the nature of the missing information. Varying the assumptions on the imputation model would likely change the imputed value. If the primary inferential goal was point prediction of the missing value, then an undesirable result of the imputation procedure would be bias or variation in the imputed values. Oftentimes, however, the main analytic goal is estimation of aggregate values, such as population means. Thus, variation in the individual imputed values is of lesser importance while bias or variation in the final population estimate moves to the forefront. Therefore, we examine, to what extent, if any, the imputation model assumptions affect the estimation of these aggregate values.

To investigate the sensitivity of inferences when using imputation models built on different assumptions, we provide a simulation study with historical data from the U.S. Consumer Expenditure Interview Survey (CE). The CE allows an in-depth consideration of the impact of two features on this potential sensitivity. They are (1) the panel survey design and (2) the characteristic of interest – expenditures. The imputation models should account for these special features of the CE. Thus, we develop several imputation models for imputing the missing data on a variety of expenditures. These expenditures tend to vary in their dollar amount and proportion of respondents with true zero-dollar expenses. We then calculate and compare estimates of population means based on the imputed data. Finally, we comment on imputation model specification and implementation feasibility.

Key Words: Missing data, Multi-stage imputation, Panel survey, Regression imputation, Zero-inflated distribution

1 Introduction

Item nonresponse, a common feature of many surveys, occurs when respondent fails to provide a response for a survey question. Failure to provide a response can happen for a variety of reasons. For example, a respondent may not remember or may not want to divulge sensitive information so he/she does not provide a response to a particular question. In addition, a survey questionnaire designer may design the questionnaire in such a way that a particular respondent is never asked a particular question; thus, never having the opportunity to provide a response. We refer to these situations as "unplanned" and "planned" missingness, respectively. In the situation of "planned" missingness the survey designer has the ability to control the missing data mechanism, i.e., because he/she determines which respondents are administered particular questions. In contrast, the survey designer has essentially no control over the missing data mechanism in the case of "unplanned" missingness. The focus of the research presented here will be on "planned" missing

*Gonzalez.Jeffrey@bls.gov

data; however, extensions of this research can be developed to handle situations with both "planned" and "unplanned" missing data

In both situations, it is desirable to draw inferences about the survey items subject to missing data. So, it is desirable to handle the missing data in a way that maintains the integrity of the target population that the analyst is attempting to characterize. There are four broad categories, which are not mutually exclusive, of missing data methods (Little and Rubin, 2002, p. 19–20). First, there are procedures that are based only on completely recorded units. Under these procedures, units with incomplete data are discarded and the remaining subset of units (i.e., only those with complete data) are analyzed. Although these procedures are relatively easy to carry out, they may lead to serious biases when drawing inferences. The second class of procedures is termed weighting procedures. Weighting procedures alter the design weights in an attempt to account for nonresponse as if it was part of the sample design. The adjustment usually comes in the form a multiplicative factor applied to the design weight; in applications, these adjustments are mostly used to compensate for unit and wave nonresponse. The third category of missing data methods is imputation-based procedures, which will be the focus of our research. These methods involve filling in the missing values and then analyzing the completed data set using standard procedures. The final set of procedures, model-based procedures, defines a broad class of procedures in which the analyst defines a model for the observed data and bases inferences on the likelihood or posterior distribution under that model. For these procedures, parameters may be estimated using methods such as maximum likelihood.

All of these methods tend to rely on some model or assumption about the nature of the missing information. In addition, these methods may also incorporate assumptions about the correlation structure among the various survey items. Varying the assumptions or specifications of features of the models usually changes the imputed value. If the primary inferential goal was point prediction of the missing value, then an undesirable result of the imputation procedure would be bias or variation in the imputed values. Frequently, however, the main analytic goal is estimation of aggregate values, such as population means; thus, predictions of the individual values are of lesser importance while concern over bias or variation in the final population estimator moves to the forefront. Therefore, it is useful to examine to what extent, if any, the assumptions of the imputation models affect the estimation of these aggregate values. This paper accomplishes that goal by conducting a series of simulation studies using historical data from the U.S. Consumer Expenditure Quarterly Interview Survey (CE). In each simulation, missing values are filled using imputation models based on different assumptions and specifications. We then calculate and compare population mean estimates under each set of imputed data and explore the potential variation among the results.

This paper is organized as follows. In the next section, we describe a research endeavor motivating this investigation. In Section 3, we make some general comments about imputation-based procedures and regression imputation. Section 4 discusses the setup and results for three simulation scenarios. The final section concludes with a few summary points and identifies areas of future research.

2 Motivating Application

The impetus for this simulation study is the current redesign effort underway for the CE. The CE is an ongoing national household panel survey sponsored by the U.S. Bureau of Labor Statistics (BLS). It is designed to collect information on the spending habits of consumers living in the U.S. The data from the survey are also used in the calculation of the Consumer Price Index (CPI), one of the nation's leading economic indicators.

The survey presents a number of challenges for both the interviewers and the respondents. First, the interview is long. Depending on the amount and type of expenditures reported, it takes, on average, 65 minutes to complete (BLS *Handbook of Methods*, 2007). Second, the questions are detailed and impose a substantial cognitive burden on respondents. In fact, the respondent is asked to report information (e.g., what was purchased, the amount of the purchase, and when it was purchased) for about 60 to 70 percent of their household's expenses made in the previous three months. These generally include significant purchases,

such as those for property, automobiles, and other large durable goods, as well as mortgage/rent and utility bill payments. Even though these are thought to be the types of expenditures that respondents can easily recount over a three-month period or longer, the nature of the reporting task may still pose problems for some respondents and thus undermine the quality of the data collected. Finally, there is also concern over declining unit response rates. Although this trend is not unique to the CE (de Leeuw and de Heer, 2002), the unit response rate for the CE was about 80 percent in the early 2000s, but now hovers around 74 percent (BLS *Handbook of Methods*, 2007). The Office of Management and Budget (OMB) has emphasized that every attempt should be made to achieve and maintain an acceptable unit response rate (OMB *Standards and Guidelines for Statistical Surveys*, 2006).

One approach to potentially reducing respondent burden while improving data quality and the unit response rate is to devise an alternative data collection strategy. Researchers within the BLS are currently investigating the use of matrix sampling or split questionnaire designs (Gonzalez and Eltinge, 2007a; 2007b; 2008). Loosely defined, split questionnaire methods involve dividing a lengthy survey into subsets of questions and then administering each subset to subsamples of a full, initial sample. These types of designs reduce the total number of questions administered to each respondent. Because of this reduction, they are thought to decrease respondent burden and improve data quality (see Herzog and Bachman [1981], Johnson *et al.* [1974], and others for a discussion on the relationship between survey length and survey data quality). However, a consequence of administering a split questionnaire to sample members is that it creates missing data for the questions not asked. Since, in the case of the CE, population estimates of average expenses made for particular items are still desired for all survey questions, regardless of whether or not they were asked of a particular respondent, it is necessary to investigate the utility of split questionnaire methods in combination with the missing data methods previously described to achieve this goal.

There are special features of the CE that we devote particular attention to in our simulation study. They are the (1) panel survey design and (2) the characteristic of interest – expenditures. The CE employs a rotating panel survey design in which a particular consumer unit (CU), which can be thought of as being equivalent to a household, is interviewed once every quarter for five calendar quarters. Although the initial interview is a bounding interview and the data are not used in any official published estimates, the same types of expenditures, in general, are asked in all five interviews. Under a possible implementation of split questionnaire designs, the first interview could remain as it is, but the second (and possibly, third through fifth) could be subject to "planned" missingness. Thus, in our evaluation, we explore the effectiveness of incorporating this information into the imputation models.

The second feature that we investigate is related to specific characteristics of the expenditures that the CE is designed to collect. Expenditure data are an interesting case because responses for certain types of expenditures are often equal to zero, and for those respondents who actually incurred an expense, their amounts tend to (approximately) follow a continuous distribution. If we examine the full distribution of an expenditure, then we likely observe a point mass, or spike, at zero, and a continuous distribution on the positive half of the real line. Moreover, expenditures or their characteristics tend to be correlated from one interview to another. If we assume that characteristics of an expense incurred in one quarter are predictive of characteristics of expenses incurred in a subsequent quarter, then we should be able to use the information obtained from the first interview in the models for recovering the missing data in the second interview. As an example, in data collected between 2006 and 2008, 54.8 percent of CUs reported clothing expenditures at the first interview and among those about 81 percent reported a clothing expense at the second interview.[1] We can exploit relationships such as this to recover information not collected in the second interview. Therefore, in our simulations, we explore the usefulness of incorporating this type of information into the imputation models.

3 Imputation-Based Procedures

[1] These calculations were made internally.

As previously mentioned, there are four broad categories of missing data methods. In general, one of the disadvantages of analyses using only complete-cases or weighting adjustments is that they make no use of cases with the variable of interest missing. This may be a mistake because these analyses ignore the potentially rich information contained in the variables without missing information. Thus, we focus on imputation-based procedures because these methods do not ignore that information. Furthermore, imputation is a fairly flexible and general method for handling item missing data (Little and Rubin, 2002). By definition, imputations are means or draws from a predictive distribution of missing values. Thus, the natural first step in imputing missing values is to construct the predictive distribution based on the observed data. Little and Rubin (2002) describe two generic approaches to generating this distribution – explicit and implicit modeling.

On the one hand, explicit modeling bases the predictive distribution on a specified formal statistical model, such as the multivariate normal distribution. On the other hand, implicit modeling focuses on an algorithm or set of procedures to construct the predictive distribution. This algorithm is consistent with one or more implicit models. In both modeling efforts, however, the assumptions need to be carefully assessed to ensure that they are reasonable. By looking at different specifications of imputation models, we can explore to what extent the assumptions are, in fact, reasonable. An example of an explicit modeling method is regression imputation. Regression imputation is a generalization of mean imputation where means from the responding units in the sample are substituted. Specifically, though, regression imputation replaces missing values by predicted values from a regression of the missing item on items observed for the unit. In addition, the regression might include continuous and categorical predictor variables, interactions, and less restrictive parametric forms in order to improve predictions.

Recall that the CE collects data for which responses are often equal to zero, and otherwise (approximately) follow a continuous distribution. In other words, expenditures follow a mixture of two distributions. For such cases, the ability to impute "zero dollars" amounts for some expenditure is a desirable property of any chosen imputation procedure. This is because not all sample members will report "non-zero dollar" amounts for all expenditure information collected. If a simple regression imputation is used, then the imputed expenditure values would likely not reflect the true underlying distribution of expenditure values. To deal with this problem, a multi-step imputation procedure might be implemented. In this situation, the first primary step would be to impute an indicator of the presence (or absence) of a given type of expenditure for a specific CU. Conditional on that indicator, the second major step would be to impute the specific dollar amount for the particular type of expenditure.

Greenlees *et al.* (1982) encountered a somewhat similar problem when imputing missing values for nonrespondents when response propensity was related to the variable being imputed. Their primary objective was to impute missing income information, but they believed that income was related to income response propensity. Basing income imputation only on the respondents would be incorrect if the distribution of income among nonrespondents was, in fact, different from the income distribution of the respondents. Essentially there are two distributions of income, one for respondents and one for nonrespondents. To combat this complication, they developed a multi-step imputation procedure in which they imputed values based on two income distributions - one for the respondents and one for the nonrespondents. By extending their approach and applying it to our research problem, we can appropriately impute both zero and non-zero dollar amounts for the expenditures of interest.

4 Simulations

A hypothetical population was constructed using CE data collected during the time frame of April 2007 to March 2008. We only included sample units that completed the first two consecutive interviews. In other words, they were respondents for both interviews one and two. This resulted in a hypothetical population of $10,412$ units. It is from this hypothetical population that $1,000$ repeated simple random samples (without replacement) of size 2,500 were drawn and missing data were imputed according to each simulation scenario specified below. Although the CE collects data on a broad set of expenditure categories, we restricted our focus to five of them – clothing, insurance other than health, medical, miscellaneous, and utilities. These

Table 1: Description of Hypothetical Finite Population

Expenditure Category	Brief Description	Reporting Rates (%)	Mean ($)	Variance (SE) of the Mean[1]
Clothing	For persons age 2 and over	70.06	259.74	82.94 (9.1)
Insurance	Non-health (e.g., auto, life)	69.84	481.20	177.95 (13.3)
Medical	Medical expenses, including medical supplies	60.90	277.00	318.72 (17.9)
Miscellaneous	Various expenses (e.g., pet services, cash contributions)	65.86	262.86	541.72 (23.3)
Utilities	Utility expenses (e.g., electricity)	92.54	596.87	78.95 (8.9)

1: Theoretical variance (SE) of the sample mean of size 2,500 selected from the full population of 10,412

five expenditure categories tend to vary in interview two reporting rates, quarterly mean expenditures, and variances. We also have demographic information on sample units from both interviews one and two. This information includes family type, a variable describing the relationship of persons living within the household, housing tenure, and age, sex, and educational attainment of the respondent. Table 1 summarizes the hypothetical population from which we will be drawing our samples for each simulation scenario.

4.1 Simulation Setup

We have three simulation scenarios which are uniquely determined by the way the missing information is imputed. The same basic method is used for imputing the missing information in all three scenarios. The primary difference among them is whether and how prior information (i.e., from a previous administration of the survey) about the sample unit is used in imputing that sample unit's missing expenditure information. In order to create missing data, for each iteration of a specific simulation scenario, a sample unit was assigned to be asked one of the five expenditure categories. That expenditure category would be "observed directly" while the other four would be set to missing and subsequently imputed.

Before we outline the steps for each simulation scenario, it is essential to identify some common notation for each scenario. First, let S denote the full sample. Let S_k denote the sample members being asked about expenditure k for $k = 1, 2, \ldots 5$. As stated above, each sample unit is asked about only one expenditure category and values for the remaining categories are imputed. We use $\mathbf{x}_i'$ to denote a vector of covariates containing the demographic information identified above. This demographic information is available for the full population. We let p_{ik} represent the probability of reporting an expense on expenditure k at interview 2 and its estimated value as $\hat{p}_{ik}$. Finally, we have $\tilde{y}_{ik}$ which takes the value of the observed value, y_{ik}, if the sample unit was asked directly about expenditure category k and an imputed value otherwise.

The steps for imputing missing values under simulation scenario 1 are given below. It is important to bear in mind that this scenario does not make use of any panel information.

1. Fit $\text{logit}(p_{ik}) = \mathbf{x}_i'\gamma_k$ to all $i \in S_k$
2. Estimate p_{ik} for all $i \in \overline{S}_k$
3. Fit $y_{ik} = \mathbf{x}_i'\beta_k$ to all $i \in S_k$ with $y_{ik} > 0$
4. Draw $\theta_{ik} \sim \text{Uni}(0, 1)$
5. Set $\tilde{y}_{ik} = y_{ik}$ if $i \in S_k$
6. Impute for all $i \in \overline{S}_k$ as follows $\tilde{y}_{ik} = \begin{cases} 0 & \theta_{ik} > \hat{p}_{ik} \\ \mathbf{x}_i'\hat{\beta}_k & \theta_{ik} \leq \hat{p}_{ik} \end{cases}$

In words, we describe this simulation scenario. We fit a logistic regression model, based only on demographic information, to the sample members receiving expenditure k. We then estimate p_{ik} for all sample members not receiving that expenditure category. To obtain the non-zero part of the expenditure distribution, we fit a regression model to all sample members receiving expenditure k with non-zero expenditure reports. Based on the estimated parameters from this regression model and a random draw from a uniform zero-one random variable, we impute for all sample members not receiving expenditure k as follows. If their estimated probability is greater than the uniform random variable, then that sample unit is imputed as a zero. If their estimated probability is less than the uniform random variable, then their demographic variable covariate pattern is used to impute their expenditure amount. What these final steps accomplish is that they impute a non-zero dollar amount that should be consistent with the sample unit's estimated prevalence of actually incurring that expense.

Under scenario 2, the same general procedure, described above, is used, except this time we incorporate the panel information into the regression imputation model. Specifically, we include the sample unit's expenditure amount reported in the first interview, if non-zero, as a covariate in the regression imputation model. We denote this value as $y_{int1,ik}$. The rationale is that expenditures reported in one interview should be predictive of expenditures reported in a subsequent interview. The steps for imputing missing values under simulation scenario 2 are given as follows.

1. Fit $\text{logit}(p_{ik}) = \mathbf{x}_i'\gamma_k$ to all $i \in S_k$
2. Estimate p_{ik} for all $i \in \overline{S}_k$
3. Fit $y_{ik} = \mathbf{x}_i'\beta_k$ to all $i \in S_k$ with $y_{ik} > 0$ but $y_{int1,ik} = 0$
4. Fit $y_{ik} = \mathbf{x}_i'\beta_k^* + y_{int1,ik}\beta_{Y_k}$ to all $i \in S_k$ with $y_{ik} > 0$, and $y_{int1,ik} > 0$
5. Draw $\theta_{ik} \sim \text{Uni}(0,1)$
6. Set $\tilde{y}_{ik} = y_{ik}$ if $i \in S_k$
7. Impute for all $i \in \overline{S}_k$ as follows $\tilde{y}_{ik} = \begin{cases} 0 & \theta_{ik} > \hat{p}_{ik} \\ \mathbf{x}_i'\hat{\beta}_k & \theta_{ik} \leq \hat{p}_{ik},\ y_{int1,ik} = 0 \\ \mathbf{x}_i'\hat{\beta}_k^* + y_{int1,ik}\hat{\beta}_{Y_k} & \theta_{ik} \leq \hat{p}_{ik}, y_{int1,ik} > 0 \end{cases}$

Finally, for scenario 3, similar to scenario 2, we incorporate the panel information into the regression imputation model. In this scenario, however, we also incorporate it into the logistic regression model predicting whether or not the sample unit would have incurred an expense of type k at interview 2. The steps for imputing missing values under scenario 3 are given below.

1. Fit $\text{logit}(p_{ik}) = \mathbf{x}_i'\gamma_k^* + y_{int1,ik}\gamma_{Y_k}$ to all $i \in S_k$
2. Estimate p_{ik} for all $i \in \overline{S}_k$
3. Fit $y_{ik} = \mathbf{x}_i'\beta_k$ to all $i \in S_k$ with $y_{ik} > 0$ but $y_{int1,ik} = 0$
4. Fit $y_{ik} = \mathbf{x}_i'\beta_k^* + y_{int1,ik}\beta_{Y_k}$ to all $i \in S_k$ with $y_{ik} > 0$, and $y_{int1,ik} > 0$
5. Draw $\theta_{ik} \sim \text{Uni}(0,1)$
6. Set $\tilde{y}_{ik} = y_{ik}$ if $i \in S_k$
7. Impute for all $i \in \overline{S}_k$ as follows $\tilde{y}_{ik} = \begin{cases} 0 & \theta_{ik} > \hat{p}_{ik} \\ \mathbf{x}_i'\hat{\beta}_k & \theta_{ik} \leq \hat{p}_{ik},\ y_{int1,ik} = 0 \\ \mathbf{x}_i'\hat{\beta}_k^* + y_{int1,ik}\hat{\beta}_{Y_k} & \theta_{ik} \leq \hat{p}_{ik}, y_{int1,ik} > 0 \end{cases}$

For each iteration of each simulation scenario, we calculated estimates of the population means of the five expenditure categories using a full-sample mean (as if there was no item missing data) and an imputation-based unweighted mean. The formula for the imputation-based mean is given by the formula below.

$$\hat{\tilde{y}}_{Ik} = \frac{1}{n}\sum_{i \in S} \tilde{y}_{ik}$$

This is similar in form to a standard design-based estimator of the population mean (assuming SRSWOR), with the only difference being the $\tilde{y}_{ik}$ term. An overall mean of the $(M =)1,000$ imputation-based means for each expenditure category, denoted by $\bar{\theta}_k$, where m is the index for the simulation iteration.

$$\overline{\theta}_k = \frac{1}{M}\sum_{m=1}^{M} \hat{\tilde{y}}_{mIk}$$

Finally, we computed a simulation variance, denoted by V_k.

$$V_k = \frac{1}{M-1}\sum_{m=1}^{M} \left(\hat{\tilde{y}}_{mIk} - \overline{\theta}_k\right)^2$$

4.2 Simulation Results

In this section, we present the results from each simulation scenario. Tables 2-4 display the simulation means based on the full sample, associated measures of variability, as well as the simulation means based on imputed data and their associated measures of variability for each of the three scenarios, respectively. We also compute a variance ratio for each expenditure category, displayed in the last column of each of Tables 2-4. This is the ratio of the variance component for the imputation-based mean, using an original, full sample of 2,500 and a 20% subsampling rate, to the variance component of the full sample mean using a sample size of 500. Recall that scenario 1 does not incorporate any prior information about the sample

Table 2: Scenario 1 Results

Expenditure Category	Full Sample		Imputation-Based		Variance Ratio
	Mean	Variance Component[1]	Mean	Variance Component[1]	
Clothing	259.82	78.47 (8.9)	260.48	513.79 (22.7)	1.31
Insurance	480.90	184.66 (13.6)	480.71	1031 (32.1)	1.12
Medical	277.45	337.49 (18.4)	279.68	2533 (50.3)	1.50
Miscellaneous	262.25	500.35 (22.4)	268.06	4509 (67.1)	1.80
Utilities	596.64	75.83 (8.7)	595.05	433.55 (20.8)	1.14

1: The corresponding standard deviations are listed in parentheses

unit into the imputation procedure. In fact, the only information used to impute the item missing expenditure information is the demographic information available for the sample unit. A comparison of the full sample means and the imputation-based means for each of the expenditure categories suggests that (nearly) unbiased estimates of mean quarterly expenditures can still be obtained using the imputation procedure. In terms of empirical bias, perhaps the only expenditure category that may warrant concern is miscellaneous expenditures. The relative empirical bias for this expenditure category is 2.21%. It is interesting to note that this expenditure category, among those considered, contains the most diverse set of expenses and is the most variable. These factors may contribute to the slightly biased estimates. In terms of the variance ratios for the five expenditure categories, relative to the full sample variance components we do observe inflations in variances ranging from 12% to 80%. It is not surprising that miscellaneous expendi-

tures results in the largest inflation in variance while insurance and utility expenditures result in the smallest.

In scenario 2, for three of the five expenditure categories – clothing, medical, and utilities – we still obtain (nearly) unbiased estimates of mean quarterly expenditures using the imputation-based mean. For insurance expenditures the empirical relative bias is 3.13% and for miscellaneous expenditures it is 2.14%. We hypothesized that characteristics of expenditure reports in interview one were "good" predictors of interview two reports; thus, in scenario 2 we incorporated the expenditure reported in interview one into the regression imputation model in order to obtain a more precise imputation. Relative to the full sample estimates, the variance inflations for the five expenditure categories range from 13% to 153% with the smallest occurring for utility expenditures and the largest for miscellaneous expenditures. With the exception of medical and utility expenditures, the variance inflations incurred under this scenario were actually larger than those incurred under scenario 1.

Table 3: Scenario 2 Results

Expenditure Category	Full Sample		Imputation-Based		Variance Ratio
	Mean	Variance Component[1]	Mean	Variance Component[1]	
Clothing	259.75	79.28 (8.9)	259.00	534.97 (23.1)	1.35
Insurance	479.87	171.83 (13.1)	494.88	1469 (38.3)	1.71
Medical	277.17	322.06 (17.9)	278.99	2302 (48.0)	1.43
Miscellaneous	263.14	570.38 (23.9)	268.76	7217 (85.0)	2.53
Utilities	596.42	83.23 (9.1)	595.11	469.85 (21.7)	1.13

1: The corresponding standard deviations are listed in parentheses

For scenario 3, we used interview one expenditure information in both the logistic regression model and the regression imputation model. For clothing and utility expenditures, we obtain (nearly) unbiased estimates of mean quarterly expenditures using the imputation-based mean estimator. For insurance, medical,

Table 4: Scenario 3 Results

Expenditure Category	Full Sample		Imputation-Based		Variance Ratio
	Mean	Variance Component[1]	Mean	Variance Component[1]	
Clothing	259.60	79.63 (8.9)	259.84	492.21 (22.1)	1.23
Insurance	481.58	160.31 (12.7)	488.90	1155 (34.0)	1.44
Medical	277.19	354.37 (18.8)	282.36	2439 (49.4)	1.38
Miscellaneous	262.08	550.93 (23.5)	280.11	7173 (84.7)	2.60
Utilities	596.51	79.07 (8.9)	595.06	359.95 (19.0)	0.91

1: The corresponding standard deviations are listed in parentheses

and miscellaneous expenditures the empirical relative biases are 1.52%, 1.87%, and 6.88%, respectively. With the exception of utility expenditures the variance inflation amounts range from 23% to 160% where clothing is the lowest and, again, miscellaneous is the highest. For utility expenditures, we actually observe a reduction in variance using the imputation-based mean relative to the full sample mean. This reduction is about 9%.

5 Discussion

In order to examine the extent to which imputation model assumptions affect the estimation of aggregate values, such as population means, we devised three simulation scenarios in which we varied the information we incorporated into the imputation models. We varied this information because we hypothesized that we

could obtain more precise predictions, on average, by using expenditure information about the sample unit from a previous administration of the survey. Our primary interest was imputing item missing expenditure data for five major expenditure categories – clothing, insurance other than health, medical, miscellaneous, and utilities. These five categories varied in their dollar amounts as well as the proportion of sample members reporting a true "zero-dollar" expense. We calculated and compared estimates of mean expenditures for these five categories using an imputation-based mean estimator according to each of the three simulation scenarios.

With the exception of miscellaneous expenditures, at least one of the scenarios resulted in unbiased estimates of the mean expenditure, using the imputation-based estimator. This is an encouraging finding, but it suggests that part of the success of the imputation procedure depends on how the model is specified. In the context of the CE, finding a "successful" imputation procedure for each expenditure category could potentially be a daunting task since there are about 14 major expenditure categories and several minor subcategories. We should note that our research only investigated 5 of these 14 expenditure categories. Depending on what level of detail the imputation is performed; imputing missing expenditure information could result in a rather intense modeling exercise.

As previously mentioned, we formulated simulation scenarios 2 and 3 under the assumption that interview one reports are "good" predictors of interview 2 reports, so these scenarios were thought to contain more precise imputation procedures. A comparison across the three simulation scenarios suggests different conclusions about the added precision by using interview one information in the imputation procedures. More specifically, the added precision of this information may depend on which expenditure category it is being used to impute. For instance, for utility expenditures, we actually observe a reduction in variance inflation as this information is incorporated in more steps of the imputation procedure and a similar trend is observed for medical expenditures. The results for miscellaneous expenditures, however, are less conclusive because we actually appear somewhat worse-off when prior information is used in the imputation procedure.

We had hoped that using information from a previous administration of the survey would result in a variance reduction for all five expenditure categories. It is possible that by incorporating some of this information, we actually over-fit some of the imputation models. The criteria used to fit the regression imputation models, especially in scenarios 2 and 3, may contribute to insufficient sample sizes being used to estimate the model parameters. This can potentially be remedied by drawing a larger sample from the population and/or using additional historical CE data to estimate the imputation model parameters.

To reiterate, our findings suggest that there is likely not one imputation procedure that works for all expenditure categories, but rather the imputation procedure should be expenditure specific. We plan to continue research on how we can modify the imputation procedures, so that when imputing each type of expenditure some efficiency gains are achieved. One way to do this, as we have done in this research, is to vary the set of covariates used in the regression imputation model. The caveat being that we need to develop these imputation procedures within the constraints of the CE program. More specifically, any set of imputation procedures developed should be able to be handled by the processing systems of the CE program else the current processing systems must be modified. As suggested above, an additional consideration given in imputation model development must be the level of detail needed for the imputation.

6 Acknowledgments

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of the U.S. Bureau of Labor Statistics. The authors would like to thank Jennifer Edgar, Scott Fricker, Karen Goldenberg, Phil Kott, Trivellore Raghunathan, Adam Safir, and Lucilla Tan for helpful discussions of the application of matrix sampling, split questionnaire designs, and imputation methods to the U.S. Consumer Expenditure Surveys.

References

[1] Bureau of Labor Statistics, U.S. Department of Labor, *Handbook of Methods*, Chapter 16, April 2007 edition, Consumer Expenditures and Income. http://www.bls.gov/opub/hom/pdf/homch16.pdf (visited September 10, 2009).

[2] De Leeuw, E. and de Heer, W. (2002). Trends in Household Survey Nonresponse: A Longitudinal and International Perspective. Chapter 3 in R.M. Groves et al. (eds.) Survey Nonresponse, New York: Wiley, 41–54.

[3] Fay, R. E. (1996). Alternative Paradigms for the Analysis of Imputed Survey Data. *Journal of the American Statistical Association*, 91, 490–498.

[4] Gonzalez, J. M. and Eltinge, J. L. (2007a). Multiple Matrix Sampling: A Review. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 3069–75.

[5] Gonzalez, J. M. and Eltinge, J. L. (2007b). Properties of Alternative Sample Design and Estimation Methods for the Consumer Expenditure Surveys. *Paper presented at the 2007 Research Conference of the Federal Committee on Statistical Methodology*, Arlington, VA, November 2007.

[6] Gonzalez, J. M. and Eltinge, J. L. (2008). Adaptive Matrix Sampling for the Consumer Expenditure Quarterly Interview Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

[7] Greenlees, J.S., W.S. Reece, and K.D. Zieschang (1982). Imputation of Missing Values when the Probability of Response Depends on the Variable Being Imputed. *Journal of the American Statistical Association*, 77, 251-261.

[8] Herzog, A. R. and Bachman, J. G. (1981). Effects of Questionnaire Length on Response Quality. *The Public Opinion Quarterly*, 45, 549–59.

[9] Johnson, W. R., Sieveking, N. A., and Clanton, E. S. (1974). Effects of Alternative Positioning of Open-Ended Questions in Multiple-Choice Questionnaires. *Journal of Applied Psychology*, **59**, 776–8.

[10] Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data.* Second Edition. Wiley.

[11] Office of Management and Budget (2006). *Standards and guidelines for statistical surveys.* Washington, D.C. http://www.whitehouse.gov/omb/inforeg/statpolicy/standards.pdf (visited September 10, 2009).

[12] Raghunathan, T. E. and Grizzle, J. E. (1995). A Split Questionnaire Survey Design. *Journal of the American Statistical Association*, 90, 54–63.

[13] Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, 63, 581–92.