Imputation Methods for Adaptive Matrix Sampling

Jeffrey M. Gonzalez and John L. Eltinge

U.S. Bureau of Labor Statistics, Office of Survey Methods Research

Abstract

Matrix sampling methods involve dividing a lengthy questionnaire into subsets of questions and administering each subset to subsamples of a full sample. In a panel survey, information about a sample unit can be learned during the first interview and this information can be used both to assign questions and to impute missing quantities at later interviews. Previous research has considered estimators based on available cases and simple adjustments to the design weights (Gonzalez and Eltinge, 2008). Here we extend this research by developing an imputation procedure for recovering the data not collected from a sample unit at subsequent interviews. We use data from the Consumer Expenditure Quarterly Interview Survey to explore potential efficiency gains incurred from incorporating these imputation methods into the estimation procedures of an adaptive matrix sampling design.

KEY WORDS: Adaptive design, Burden reduction, Multiple imputation, Panel survey, Sample survey, Variance estimation

Background and Motivation

The U.S. Consumer Expenditure Quarterly Interview Survey (CEQ) is an on-going rotating panel survey of a nationally representative sample of addresses that solicits information on the spending habits of American consumers. The data are also used in the calculation of the Consumer Price Index (CPI). Recently there has been increasing concern over (1) declining response rates; (2) high respondent burden; (3) potentially low data quality; and, (4) increasing data collection costs. To address these concerns we explore the use of matrix sampling methods under an adaptive design framework.

Definitions

- **Matrix sampling**: To divide a lengthy questionnaire into subsets of questions and, based on some probabilistic mechanism, administering each to subsamples of a full sample
- **Adaptive design** (responsive design): To use information learned about a sample unit during the data collection process with the purpose of tailoring features of the survey administration to that particular unit
- **Adaptive matrix sampling**: To use data collected in the first interview (e.g., expenditure and demographic characteristics) to (1) adjust matrix subsampling probabilities for subsequent interviews and (2) impute information not collected from a particular subsample

Here our **focus** is on the **imputation procedure** used when data are collected via an adaptive matrix sampling design.

Interesting Feature of Expenditure Data

Responses are often equal zero OR otherwise (approximately) follow a continuous distribution. Thus, the ability to impute "zero dollar" expenditure amounts is a desirable property of the chosen imputation procedure. We accomplish this with a two-stage imputation in which we (1) predict the presence or absence of an expenditure and then (2) conditional on that, impute the specific dollar amount.

Mathematical Details

We are interested in drawing inferences about mean expenditures for various categories. The primary statistic of interest is $\bar{Y}_k$. Suppose we draw a random sample, denoted by S, from the target population, then a design-based estimator would be:

$$\hat{\bar{Y}}_k = \left(\sum_{i\in S} w_i\right)^{-1}\left(\sum_{i\in S} w_i Y_{ik}\right)$$

Since not all information is collected, we can either: (1) make an appropriate adjustment to the sampling weights (Gonzalez and Eltinge, 2008), (2) impute the non-observed information. Under (2), a reasonable estimator would be (where $\tilde{Y}_{ik}$ is the imputed value):

$$\hat{\bar{Y}}_{Ik} = \left(\sum_{i\in S} w_i\right)^{-1}\left(\sum_{i\in S} w_i \tilde{Y}_{ik}\right)$$

A useful evaluative tool for judging the performance of this estimator is the variance. The variance of $\hat{\bar{Y}}_{Ik}$ due to three sources of variation (initial sample selection, matrix subsampling, and imputation error) can be written as:

$$\begin{aligned} V\left(\hat{\bar{Y}}_{Ik}\right) &= E_1\left[V_{23}\left(\hat{\bar{Y}}_{Ik}\,|\,\mathrm{S}\right)\right] + V_1\left[E_{23}\left(\hat{\bar{Y}}_{Ik}\,|\,\mathrm{S}\right)\right] \\ &= E_1\left\{V_2\left[E_3\left(\hat{\bar{Y}}_{Ik}\,|\,\mathrm{P},\mathrm{S}\right)\right] + E_2\left[V_3\left(\hat{\bar{Y}}_{Ik}\,|\,\mathrm{P},\mathrm{S}\right)\right]\right\} + V_1\left[E_{23}\left(\hat{\bar{Y}}_{Ik}\,|\,\mathrm{S}\right)\right] \\ &= E_1\left\{V_2\left[E_3\left(\hat{\bar{Y}}_{Ik}\,|\,\mathrm{P},\mathrm{S}\right)\right] + E_2\left[V_3\left(\hat{\bar{Y}}_{Ik}\,|\,\mathrm{P},\mathrm{S}\right)\right]\right\} + V_1\left[\hat{\bar{Y}}_k\right] \end{aligned}$$

where:

- $E_1(\cdot), V_1(\cdot)$ are the moments with respect to the original sample selection;
- $E_{23}(\cdot\,|\,\mathrm{S}), V_{23}(\cdot\,|\,\mathrm{S})$ are the moments with respect to the matrix subsampling, conditional on the initial sample, S; and,
- $E_3(\cdot\,|\,\mathrm{P},\mathrm{S}), V_3(\cdot\,|\,\mathrm{P},\mathrm{S})$ are the moments with respect to the imputation procedure, conditional on the matrix subsampling, P, and the initial sample.

Two Central Questions

1. What effect will the predictive precision of the imputation procedure have for the non-observed expenditure sections?
2. What is the additional variance reduction obtained by assigning sections to sample units with unequal probabilities?

The simulation study will begin to shed light on the first question by considering two imputation procedures hypothesized to have differing predictive precisions.

Step 1: Fit $\mathrm{logit}(p_{ik}) = X_i^T\gamma_k$ to all respondents (where p_{ik} is the probability of the i^{th} unit having an expense for k. Estimate p_{ik} for all sample units who were not asked about that expenditure using the relationship: $\hat{p}_{ik} = \left(1+\exp\left(X_i^T\hat{\gamma}_k\right)\right)^{-1}\exp\left(X_i^T\hat{\gamma}_k\right)$

Step 2: Fit a regression model to all respondents (see Handout), draw $\theta_{ik} \sim Uni(0,1)$, and impute Y_{ik} via Option 1 or Option 2 below:

Option 1:	Option 2:
$\tilde{Y}_{ik} = \begin{cases} Y_{ik} & \text{if } i \in S_k \\ 0 & \text{if } \theta_{ik} > \hat{p}_{ik} \\ X_i^T\hat{\beta}_k & \text{if } \theta_{ik} < \hat{p}_{ik} \end{cases}$	$\tilde{Y}_{ik} = \begin{cases} Y_{ik} & \text{if } i \in S_k \\ 0 & \text{if } \theta_{ik} > \hat{p}_{ik} \\ X_i^T\hat{\beta}_k & \text{if } \theta_{ik} \le \hat{p}_{ik} \text{ and } Y_{Int1,ik} = 0 \\ X_i^T\hat{\beta}_k^* + Y_{Int1,ik}\hat{\beta}_{Y_k} & \text{if } \theta_{ik} \le \hat{p}_{ik} \text{ and } Y_{Int1,ik} > 0 \end{cases}$

Simulation Setup

Data Source

- Data collected from April 2006 to March 2008 using the full CEQ CAPI survey instrument; represents 8 calendar quarters
- Subset to sample units responding to BOTH interviews 1 and 2
- Identified 5 expenditure categories with varying interview 2 reporting rates, quarterly mean expenditures, and variances (see Table 1)
- Demographic information on sample units: family type (describes the relationship among the persons living within the sample unit), housing tenure (own vs. rent), age, sex, and educational attainment of the respondent

Table 1: Population Description (N = 10412)

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., life, auto)	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 full sample mean

Two Simulation Scenarios

- Common features: SRSWOR (n = 2500) from population, randomly allocate each sample unit to 1 of the 5 expenditure categories, 1000 iterations (computing $\hat{\bar{Y}}_k$ and $\hat{\bar{Y}}_{Ik}$ each iteration)
- Scenario 1: Uses Option 1 to impute missing expenditure information
- Scenario 2: Uses Option 2 to impute missing expenditure information

Simulation Results

Table 2: Scenario 1 Results

Expenditure Category	Full Sample Mean	Full Sample Variance Component[1]	Matrix Sample Mean	Matrix Sample Variance Component[1]	Variance Ratio[2]
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
2: Ratio of the variance of the imputation-based mean relative to the variance of the full sample mean

Table 3: Scenario 2 Results

Expenditure Category	Full Sample Mean	Full Sample Variance Component[1]	Matrix Sample Mean	Matrix Sample Variance Component[1]	Variance Ratio[2]
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
2: Ratio of the variance of the imputation-based mean relative to the variance of the full sample mean

Discussion

Lessons Learned

- As a baseline for comparison, we would expect the standard errors (of the mean) to increase by a factor of √0.2 (due to the one-fifth matrix subsampling); however, in both Scenarios 1 and 2, the standard errors were inflated by factors larger than √0.2
- We formulated Option 2 under the assumption that interview 1 expenditure reports are "good" predictors of interview 2 reports, so Option 2 was thought to be a more precise imputation procedure
  - Since Option 2 did not yield the anticipated results, we think that modeling and the subsequent improvement in efficiency likely depends on the logistic and regression model specifications used to impute the non-observed information
  - It may also be the case that insufficient samples sizes are being used to estimate the regression model parameters
- Despite a decrease in efficiency, it appears that we can still obtain unbiased estimates of mean expenditures, but the potential lack of fit in the regression models results in an undesired inflation in variance

Future Research

- Continued exploration of how we can modify the imputation model in order to achieve efficiency gains (FCSM, 2009)
  - We can potentially vary the set of covariates used to predict each expenditure category (e.g., housing tenure might predict well utility expenses but not miscellaneous expenses)
- This research primarily focused on the $V_2\left[E_3\left(\hat{\bar{Y}}_{Ik}\,|\,\mathrm{P},\mathrm{S}\right)\right]$ component of variation, the obvious next question is how to reduce the other component, $E_2\left[V_3\left(\hat{\bar{Y}}_{Ik}\,|\,\mathrm{P},\mathrm{S}\right)\right]$
  - We anticipate that this can be accomplished by a careful choice of matrix subsampling probabilities (e.g., potentially make use of a logistic model similar to Step 1)
- We will also evaluate the procedures using a multivariate framework

Acknowledgments

The authors would like to thank Jennifer Edgar, Scott Fricker, Karen Goldenberg, Bill Mockovak, Adam Safir, and Lucilla Tan for their useful discussion on the Consumer Expenditure Surveys, matrix sampling, and other contributions to this research.

The views expressed on this poster are those of the authors and do not reflect the policies of the U.S. Bureau of Labor Statistics.

Contact Information

Jeffrey M. Gonzalez
Mathematical Statistician

U.S. Bureau of Labor Statistics
Office of Survey Methods Research
2 Massachusetts Avenue NE, Suite 1950
Washington, DC 20212

T: 202.691.7415
F: 202.691.7426
E: Gonzalez.Jeffrey@bls.gov

Imputation Methods for Adaptive Matrix Sampling
(Handout)

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

2009 Joint Statistical Meetings
August 4, 2009

Imputation Procedure Details[1]

To impute y_{ik}, the expenditure amount for the i^{th} sample unit on item k, we will implement the following two-step procedure:

Step 1:
Fit the logistic regression model, $\text{logit}(p_{ik}) = x_i'\gamma_k$, to all sample units receiving expenditure section (or item) k. For this model, p_{ik} is the probability that the i^{th} sample unit reports an expense on item k and x_i' is a vector of covariates (family type, housing tenure, and age, educational attainment, and gender of the respondent).

Using the relationship $p_{ik} = (1 + \exp(x_i'\gamma_k))^{-1}(\exp(x_i'\gamma_k))$, estimate p_{ik} for all sample units not receiving expenditure section k. We will denote the estimated probability as $\hat{p}_{ik}$.

Step 2:

- ⋆ *Option 1:*
  Fit the linear regression model, $y_{ik} = x_i'\beta_k$, to all sample units receiving expenditure section k and reporting a positive expense (i.e., $y_{ik} > 0$). For this model, x_i' is the same vector of covariates as in Step 1.

- ⋆ *Option 2:*
  Estimate the regression parameters from the following two linear regression models:

  1. Fit $y_{ik} = x_i'\beta_k$ to all sample units receiving expenditure section k, reporting a positive expense (i.e., $y_{ik} > 0$) at the current interview, but a zero-dollar expense during the first interview (i.e., $y_{int1,ik} = 0$)
  2. Fit $y_{ik} = x_i'\beta_k^* + y_{int1,ik}\beta_{Y_k}$ to all sample units receiving expenditure section k, reporting a positive expense (i.e., $y_{ik} > 0$) at the current interview, and a positive expense during the first interview (i.e., $y_{int1,ik} > 0$)

Now draw, $\theta_{ik} \sim Uni(0, 1)$. For all sample units not receiving expenditure section k, impute the non-observed information using either Option 1 or 2 in the following manner:

*Gonzalez.Jeffrey@bls.gov

[1]The views expressed here are entirely those of the authors and do not necessarily reflect policies of the U.S. Bureau of Labor Statistics.

Table 1: Imputation Options

Option 1	Option 2
$\tilde{y}_{ik} = \begin{cases} 0 & \theta_{ik} > \hat{p}_{ik} \\ x_i'\hat{\beta}_k & \theta_{ik} \leq \hat{p}_{ik} \end{cases}$	$\tilde{y}_{ik} = \begin{cases} 0 & \theta_{ik} > \hat{p}_{ik} \\ x_i'\hat{\beta}_k & \theta_{ik} \leq \hat{p}_{ik},\ y_{int1,ik} = 0 \\ x_i'\hat{\beta}_k^* + y_{int1,ik}\hat{\beta}_{Y_k} & \theta_{ik} \leq \hat{p}_{ik},\ y_{int1,ik} > 0 \end{cases}$

FCSM 2009
Sensitivity of Inference under Imputation: An Empirical Study

Jeffrey M. Gonzalez and John L. Eltinge

Abstract

Item nonresponse, a common problem in many surveys, occurs 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 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 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 three features on this potential sensitivity. They are (1) panel survey design; (2) range of expenditure dollar amounts; and (3) prevalence of certain expenditures (i.e., rare vs. frequently incurred expenses). The imputation models should account for these special features of the CE. Thus, we develop several imputation models for imputing a variety of expenditures. These expenditures vary in dollar amount, proportion of item nonresponse, 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 offer a commentary on imputation model parsimoniousness and implementation feasibility.

Key Words: Zero-inflated distribution; Panel survey; Missing data; Regression imputation; Two-stage imputation

Reference Cited on Poster

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*, 2081-2088.