Simulation Setup

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 matrix sampling methods into the estimation procedures of an adaptive matrix sampling design.

Mathematical Details

We are interested in drawing inferences about mean expenditures for various expenditure categories. The primary statistic of interest is \( \mu_i \). Suppose we draw a random sample, defined by \( S \), from the target population, then a design-based estimator would be:

\[
\hat{\mu}_i = \frac{1}{n} \sum_{j=1}^{n} w_{ij} \cdot \sum_{k=1}^{K} \hat{Y}_{ik}
\]

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 will be \( \hat{\mu}_i \), the imputed value.

A useful evaluative tool for judging the performance of this estimator is the variance of \( \hat{\mu}_i \) due to three sources of variation: (Initial sample selection, matrix subsampling, and imputation error) can be written as:

\[
\text{Var}(\hat{\mu}_i) = E\left[\left(\hat{\mu}_i - E[\hat{\mu}_i]\right)^2\right] = E\left[\left(\hat{Y}_{ik} - E[\hat{Y}_{ik}]\right)^2\right] + E\left[\left(E[\hat{Y}_{ik}] - E[E[\hat{Y}_{ik}]]\right)^2\right] + E\left[\left(E[E[\hat{Y}_{ik}]] - E[\hat{Y}_{ik}]\right)^2\right]
\]

where:

- \( E[\hat{Y}_{ik}] \) are the moments with respect to the original sample selection;
- \( E[E[\hat{Y}_{ik}]] \) are the moments with respect to the matrix subsampling, conditional on the initial sample, \( S \); and
- \( E[E[E[\hat{Y}_{ik}]]] \) 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 by shed light on the first question by considering two imputation procedures hypothesized to have differing predictive precisions.

Step 1: \( \Phi(i, p_{ik}) = \lambda \cdot \psi \) to all respondents (where \( p_{ik} \) is the probability of the \( i \)-th unit having an expense for \( k \)).

Step 2: Fit a regression model to all respondents (see Handout), draw \( d_i \sim \text{Unif}(0,1) \), and impute \( \hat{Y}_{ik} \) via Option 1 or Option 2 below.

Simulation Results

Table 1: Population Description (N = 10412)

<table>
<thead>
<tr>
<th>Expenditure Category</th>
<th>Brief Description</th>
<th>Reporting Rates (%)</th>
<th>Mean ($)</th>
<th>Variance (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clothing</td>
<td>For persons age 2 and over</td>
<td>70.05</td>
<td>259.74</td>
<td>92.94 (9.0)</td>
</tr>
<tr>
<td>Medical</td>
<td>Health services and supplies</td>
<td>78.95</td>
<td>596.87</td>
<td>78.95 (8.9)</td>
</tr>
<tr>
<td>Food</td>
<td>Food and beverages</td>
<td>77.69</td>
<td>323.02</td>
<td>49.06 (7.1)</td>
</tr>
<tr>
<td>Utilities</td>
<td>Utility expenses</td>
<td>75.75</td>
<td>283.95</td>
<td>43.55 (6.1)</td>
</tr>
</tbody>
</table>

Table 2: Scenario 1 Results

<table>
<thead>
<tr>
<th>Expenditure Category</th>
<th>Full Sample</th>
<th>Matrix Sample</th>
<th>Variance Component 1</th>
<th>Variance Component 2</th>
<th>Variance Ratio 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clothing</td>
<td>259.82</td>
<td>78.67 (9.5)</td>
<td>360.46</td>
<td>513.76 (22.7)</td>
<td>1.31</td>
</tr>
<tr>
<td>Medical</td>
<td>275.45</td>
<td>73.47 (16.4)</td>
<td>270.68</td>
<td>253.03 (32.3)</td>
<td>1.55</td>
</tr>
<tr>
<td>Food</td>
<td>262.25</td>
<td>104.25 (22.4)</td>
<td>262.09</td>
<td>409.07 (7.1)</td>
<td>1.58</td>
</tr>
<tr>
<td>Utilities</td>
<td>565.64</td>
<td>75.83 (8.7)</td>
<td>595.05</td>
<td>433.55 (20.3)</td>
<td>1.14</td>
</tr>
</tbody>
</table>

Simulation Setup

Data Source

- Data collected from April 2006 to March 2009 using the full CEQ
- Diary survey instrument; represents 8 calendar quarters
- Subsets 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 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)

<table>
<thead>
<tr>
<th>Expenditure Category</th>
<th>Reporting Rates (%)</th>
<th>Mean ($)</th>
<th>Variance (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clothing</td>
<td>70.05</td>
<td>259.74</td>
<td>92.94 (9.0)</td>
</tr>
<tr>
<td>Medical</td>
<td>78.95</td>
<td>596.87</td>
<td>78.95 (8.9)</td>
</tr>
<tr>
<td>Food</td>
<td>77.69</td>
<td>323.02</td>
<td>49.06 (7.1)</td>
</tr>
<tr>
<td>Utilities</td>
<td>75.75</td>
<td>283.95</td>
<td>43.55 (6.1)</td>
</tr>
</tbody>
</table>

Simulation Setup

Data Source

- Data collected from April 2006 to March 2009 using the full CEQ
- Diary survey instrument; represents 8 calendar quarters
- Subsets 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 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)

<table>
<thead>
<tr>
<th>Expenditure Category</th>
<th>Reporting Rates (%)</th>
<th>Mean ($)</th>
<th>Variance (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clothing</td>
<td>70.05</td>
<td>259.74</td>
<td>92.94 (9.0)</td>
</tr>
<tr>
<td>Medical</td>
<td>78.95</td>
<td>596.87</td>
<td>78.95 (8.9)</td>
</tr>
<tr>
<td>Food</td>
<td>77.69</td>
<td>323.02</td>
<td>49.06 (7.1)</td>
</tr>
<tr>
<td>Utilities</td>
<td>75.75</td>
<td>283.95</td>
<td>43.55 (6.1)</td>
</tr>
</tbody>
</table>
Imputation Methods for Adaptive Matrix Sampling
(Handout)

Jeffrey M. Gonzalez∗ and John L. Eltinge

2009 Joint Statistical Meetings
August 4, 2009

Imputation Procedure Details
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, $\logit(p_{ik}) = x_i'\gamma_k$, 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

†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

<table>
<thead>
<tr>
<th>Option 1</th>
<th>Option 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tilde{y}<em>{ik} = \begin{cases} 0 &amp; \theta</em>{ik} &gt; \hat{p}<em>{ik} \ x_i'\hat{\beta}<em>k &amp; \theta</em>{ik} \leq \hat{p}</em>{ik} \end{cases}$</td>
<td>$\tilde{y}<em>{ik} = \begin{cases} 0 &amp; \theta</em>{ik} &gt; \hat{p}<em>{ik} \ x_i'\hat{\beta}<em>k &amp; \theta</em>{ik} \leq \hat{p}</em>{ik}, \ y_{int1,ik} = 0 \ x_i'\hat{\beta}<em>k^* + y</em>{int1,ik}\hat{\beta}<em>Y_k &amp; \theta</em>{ik} \leq \hat{p}<em>{ik}, \ y</em>{int1,ik} &gt; 0 \end{cases}$</td>
</tr>
</tbody>
</table>

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