Imputation of Multiple-Response Items in SESTAT and Its Component Surveys

July 31, 2011

Joint Statistical Meetings
Nick Beyler • Bruno Vizcarra • Don Jang • Amang Sukasih
The Scientists and Engineers Statistical Data System (SESTAT) is an integrated data system sponsored by the National Science Foundation (NSF)

SESTAT provides users with characteristics based on a nationally-representative sample of scientists and engineers in the United States

Access is available at sestat.nsf.gov
SESTAT and Its Component Surveys

- SESTAT is comprised of data from three national surveys
  - National Survey of College Graduates (NSCG)
  - National Survey of Recent College Graduates (NSRCG)
  - Survey of Doctorate Recipients (SDR)

- Key eligibility criteria for all surveys
  - 75 or younger, non-institutionalized, and living in the U.S. on the survey reference date
  - Bachelor’s or higher degree (or employment) in science, engineering, or related health field
Nonresponse in the Component Surveys

- All three SESTAT component surveys are subject to unit, partial, and item nonresponse

- Unit nonresponse
  - Sampled unit does not return or complete the questionnaire
  - Accounted for using weighting adjustments

- Partial nonresponse
  - Sampled unit answers some but not all of the “critical” items in the questionnaire (items about working status, living in the U.S., etc.)
  - Treated as unit nonresponse
Nonresponse in the Component Surveys

- **Item nonresponse**
  - Sampled unit answers all critical items, but does not answer some non-critical items
  - Accounted for using hot deck imputation

- **Unweighted sequential hot deck imputation**
  - A missing response to an item for a sampled unit is replaced by the response from a similar sampled unit, called a donor
  - See Andridge and Little (2010)
During each survey round, the individual survey contractors perform hot deck imputation before the data are integrated into SESTAT.

To maintain consistency in the imputation procedures across surveys and survey years, the contractors are given SESTAT imputation specifications prior to imputation.

The specifications list classing and sorting variables to be used for sequential hot deck imputation.
Classing variables
- Used to divide the sample into imputation classes prior to imputation
- Usually predetermined as filter (or skipping) variables

Sorting variables
- Used to sort cases within imputation classes
- Determined through stepwise regressions using data from the previous survey round
Example
- Variable to be imputed: Salary
- Classing variable: Working Status
- Input variables for the regression model: Gender, Age, Race/Eth, Job Field, Year of Degree
- Significant variables from stepwise regression: Age, Gender, Year of Degree

Final specifications
- Classing variable: Working Status
- Sorting variables: Age, Gender, Year of Degree
Multiple-Response Items in SESTAT

Example

B21. (If Not related) Did these factors influence your decision to work in an area outside the field of your highest degree?

Mark Yes or No for each item.

1. Pay, promotion opportunities
2. Working conditions (e.g., hours, equipment, working environment)
3. Job location
4. Change in career or professional interests
5. Family-related reasons (e.g., children, spouse’s job moved)
6. Job in highest degree field not available
7. Some other reason – Specify

Yes  No

NRPAY  NRPAY
NRCON  NRCON
NRLOC  NRLOC
NRCHG  NRCHG
NRFAM  NRFAM
NROCNA  NROCNA
NROT  NROT
Multiple-Response Items in SESTAT

- Multiple responses are imputed individually using multiple donors

- This procedure partially ignores the potential correlation between responses during imputation

- Hot deck methods have been considered for multivariate missing data
  - See Andridge and Little (2010)

- We are interested in a method that can be easily implemented under the existing SESTAT framework
SRMI

- Sequential Regression Multivariate Imputation
  - Raghunathan et al. (2001)

- Joint model for multiple response variables is fit sequentially until convergence, while missing values are being imputed from posterior distributions

- IVEware implements the SRMI methodology with a SAS-callable application
  - Raghunathan et al. (2002)

- Users may specify a minimum marginal $R^2$ for stepwise model fitting
Joint density is $f(y_1, \ldots, y_k | X, \theta_1, \ldots, \theta_k)$, where $y_j$ is the $j$th response, $X$ is a set of covariates, and $\theta_j$ is a vector of parameters (e.g., regression coefficients) for $y_j$

Fitting the joint model
- Order $y$ values so that $y_1$ has the fewest number of missing values, $y_2$ has the second fewest, and so on
- Regress $y_1$ on $X$ and impute missing $y_1$ values
- Regress $y_2$ on $(X, y_1)$ and impute missing $y_2$ values
- Continue until $y_k$ is regressed on $(X, y_1, \ldots, y_{k-1})$ and the missing values for $y_k$ are imputed
- Repeat, but regress $y_1$ on $(X, y_2, \ldots, y_k)$ and re-impute missing $y_1$, regress $y_2$ on $(X, y_1, y_3, \ldots, y_k)$ and re-impute missing $y_2$, and so on, until convergence
SRMI for SESTAT Imputation

- Benefit
  - Accounts for correlation among multiple responses

- Limitation (in terms of our needs)
  - Not a hot deck imputation method

- Solution
  - Use SRMI to determine sorting variables for hot deck imputation
  - Implement hot deck imputation for multiple-response items under the usual SESTAT framework
Proposed Modified Hot Deck Approach

- **Initial step**
  - Fit joint model using IVEware to determine sorting variables for each of the multiple responses

- **First pass**
  - Impute sequentially using non-missing items only (some sorting variables that are also response variables may have to be omitted in the first pass)

- **Second pass**
  - Impute sequentially using all items (response variables can now be used as sorting variables after the first pass)
Exploratory Analyses

- 2006 NSRCG
- Compare three imputation methods
  - Existing hot deck imputation approach
  - Standard SRMI imputation
  - Modified hot deck approach
Example 1 – Working Outside of Field

B21. (If Not related) Did these factors influence your decision to work in an area outside the field of your highest degree?

Mark Yes or No for each item.

1. Pay, promotion opportunities

2. Working conditions (e.g., hours, equipment, working environment)

3. Job location

4. Change in career or professional interests

5. Family-related reasons (e.g., children, spouse’s job moved)

6. Job in highest degree field not available

7. Some other reason – Specify

Yes  No

- NRPAY
- NRCON
- NRLOC
- NRCHG
- NRFAM
- NROCNA
- NROT
Example 1 – Working Outside of Field

Percent of imputed cases with an imputed response of “Yes” (n = 75)

<table>
<thead>
<tr>
<th>Response</th>
<th>Standard Hot Deck</th>
<th>Standard SRMI</th>
<th>Modified Hot Deck</th>
</tr>
</thead>
<tbody>
<tr>
<td>NRPAY</td>
<td>54.7</td>
<td>64.0</td>
<td>53.3</td>
</tr>
<tr>
<td>NRCON</td>
<td>53.3</td>
<td>53.3</td>
<td>61.3</td>
</tr>
<tr>
<td>NRLOC</td>
<td>49.3</td>
<td>46.7</td>
<td>46.7</td>
</tr>
<tr>
<td>NRCHG</td>
<td>28.0</td>
<td>33.3</td>
<td>30.7</td>
</tr>
<tr>
<td>NRFAM</td>
<td>12.0</td>
<td>20.0</td>
<td>8.0</td>
</tr>
<tr>
<td>NROCNA</td>
<td>46.7</td>
<td>38.7</td>
<td>44.0</td>
</tr>
<tr>
<td>NROT</td>
<td>13.3</td>
<td>9.3</td>
<td>9.3</td>
</tr>
</tbody>
</table>
Example 2 – Training

C2. (If Yes) For which of the following reasons did you take training during the past 12 months?

Mark Yes or No for each item.

- WTRSKL 1 To improve skills or knowledge in your current occupational field ............... 1 □ 2 □
- WTROPPS 2 To increase opportunities for promotion or advancement in your current occupational field ............... 1 □ 2 □
- WTRLIC 3 For licensure or certification in your current occupational field ............... 1 □ 2 □
- WTRCHOC 4 To facilitate a change to a different occupational field ........................................ 1 □ 2 □
- WTREM 5 Required or expected by employer ............. 1 □ 2 □
- WTRPERS 6 For leisure or personal interest ............... 1 □ 2 □
- WTROT 7 Other – Specify ........................................ 1 □ 2 □
Example 2 – Training

Percent of imputed cases with an imputed response of “Yes” (n = 266)

<table>
<thead>
<tr>
<th>Response</th>
<th>Standard Hot Deck</th>
<th>Standard SRMI</th>
<th>Modified Hot Deck</th>
</tr>
</thead>
<tbody>
<tr>
<td>WTRSKL</td>
<td>94.0</td>
<td>96.2</td>
<td>94.7</td>
</tr>
<tr>
<td>WTROPPS</td>
<td>55.6</td>
<td>55.6</td>
<td>56.0</td>
</tr>
<tr>
<td>WTRLIC</td>
<td>42.1</td>
<td>40.2</td>
<td>39.5</td>
</tr>
<tr>
<td>WTRCHOC</td>
<td>10.9</td>
<td>13.5</td>
<td>14.3</td>
</tr>
<tr>
<td>WTREM</td>
<td>70.7</td>
<td>65.0</td>
<td>70.3</td>
</tr>
<tr>
<td>WTRPERS</td>
<td>43.2</td>
<td>49.2</td>
<td>42.5</td>
</tr>
<tr>
<td>WTROT</td>
<td>0.8</td>
<td>0.8</td>
<td>2.3</td>
</tr>
</tbody>
</table>
Observations

- Exploratory analyses show
  - Differences in imputation across methods
  - Evidence of correlation among multiple responses

- Differences do not greatly impact estimation and inference
  - Low item nonresponse
Future Work

- Compare imputation methods for multiple-response items across all three SESTAT component surveys
- Conduct a rigorous evaluation of the imputation methods
  - Simulation study
- Implement new method for SESTAT imputation specifications
References


Thank you for your attention!

- For additional questions or comments please contact:
  - Nick Beyler
    - nbeyler@mathematica-mpr.com