Supplemental Analysis Example
Use of Stata 11.1 MI Tools for Imputation and Analysis of Complex Sample Data
April 13, 2011

This example uses a subset of NCS-R data to impute missing data on the variable, educ (continuous years of education) and then analyze the imputed data sets taking the complex sample design into account. The goal of this example is to demonstrate new imputation and analysis of imputed data sets tools included in Stata v11.1.

Commonly used features such as declaring the data as "mi", defining variables as either imputed or regular, use of the regression method for imputation of a continuous variable with a monotone missing data pattern, mi svyset to set mi data for survey analysis, and mi estimate: with svy commands are demonstrated. The imputed values for the variable, educ, may require rounding to obtain an integer value. For more information, see the Stata 11 documentation for the mi suite of commands.

----------------------------
use "F:\brahms\applied_analysis_book\Supplemental Code Examples for ASDA Website\stata 11 mi example\ncsr_exercise3.dta", clear
* Examine data for variables with missing data
.misstable summarize

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs&lt;.</th>
<th>Obs&gt;.</th>
<th>Obs&lt;.</th>
<th>Unique values</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>educ</td>
<td>872</td>
<td>4,820</td>
<td>0</td>
<td>18</td>
<td>0</td>
<td>17</td>
</tr>
</tbody>
</table>

* Use summarize to examine all numeric variables in data set
.summarize

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>sampleid</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>str</td>
<td>5692</td>
<td>26.3788</td>
<td>11.15588</td>
<td>1</td>
<td>42</td>
</tr>
<tr>
<td>secu</td>
<td>5692</td>
<td>1.50527</td>
<td>.5000161</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>finalp2wt</td>
<td>5692</td>
<td>.581518</td>
<td>.4933534</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>racecat_</td>
<td>5692</td>
<td>3.42323</td>
<td>1.025544</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>agecat</td>
<td>5692</td>
<td>2.36858</td>
<td>1.028047</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>hhinc</td>
<td>5692</td>
<td>62001.4</td>
<td>57590.46</td>
<td>0</td>
<td>337500</td>
</tr>
<tr>
<td>educ</td>
<td>4820</td>
<td>13.2666</td>
<td>2.417788</td>
<td>0</td>
<td>17</td>
</tr>
</tbody>
</table>

. * prior to imputation and estimation, need to set as "mi" data, first see if already set?
. mis query
(data not mi set)
. mi set mlong
. register variable educ as imputed
. register imputed educ
(872 m=0 obs. now marked as incomplete)
. * declare all other variables as regular, note that sampleid is omitted since it is not used in imputation
. mi register regular str secu finalp2wt sexf racecat_ agecat hhinc
. * impute missing data on educ using "regular" variables
. mi impute regress educ str secu finalp2wt sexf racecat_ agecat hhinc, add(5) rseed(655)

Univariate imputation                  Imputations = 5
Linear regression                     added = 5
Imputed: m=1 through m=5              updated = 0

<table>
<thead>
<tr>
<th>Variable</th>
<th>Complete</th>
<th>Incomplete</th>
<th>Imputed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>educ</td>
<td>4820</td>
<td>872</td>
<td>872</td>
<td>5692</td>
</tr>
</tbody>
</table>

(complete + incomplete = total; imputed is the minimum across m of the number of filled-in observations.)
. * set survey data
. mi svyset secu [pweight=finalp2wt], strata(str)
  pweight: finalp2wt
  VCE: linearized
  Single unit: missing
  Strata 1: str
  SU 1: secu
  FPC 1: <zero>

. * use of mi estimate to analyze imputed data sets using svy: regress
. mi estimate: svy: regress hhinc sexf educ i.racecat_

Multiple-imputation estimates
Survey: Linear regression
Imputations = 5
Number of obs = 5692
Population size = 5692.0005
Number of strata = 42
Number of PSUs = 84
Average RVI = 0.0184
Complete DF = 42
DF adjustment: Small sample
DF: min = 38.30
avg = 39.34
max = 39.98
Model F test: Equal FMI
F( 5, 40.0) = 114.79
Within VCE type: Linearized
Prob > F = 0.0000

|        | Coef.  | Std. Err. | t     | P>|t|   | [95% Conf. Interval] |
|--------|--------|-----------|-------|-------|----------------------|
| hhinc  | -15534.96 | 1733.276  | -8.96 | 0.000 | -19039.09 -12030.83 |
| sexf   | 7104.907 | 486.2448  | 14.61 | 0.000 | 6121.55 8088.263   |
| educ   | 1091.716 | 6420.198  | 0.17  | 0.866 | -11884.15 14067.58 |
| racecat_2 | -12870.44 | 4987.082  | -2.58 | 0.014 | -22955.37 -2785.515|
| racecat_3 | 1091.716 | 6420.198  | 0.17  | 0.866 | -11884.15 14067.58 |
| racecat_4 | 2276.086 | 4155.95   | 0.55  | 0.587 | -6125.977 10678.15 |
| _cons  | -22652.71 | 7270.611  | -3.12 | 0.003 | -37367.49 -7937.934|

. * use of mi estimate with svy mean to obtain mean on education (imputed) with svy correction
. mi estimate: svy: mean educ

Multiple-imputation estimates
Survey: Mean estimation
Imputations = 5
Number of obs = 5692
Population size = 5692.0005
Number of strata = 42
Number of PSUs = 84
Average RVI = 0.0932
Complete DF = 42
DF adjustment: Small sample
DF: min = 34.41
avg = 34.41
max = 34.41
Within VCE type: Linearized

Mean | Coef.  | Std. Err. | t     | P>|t|   | [95% Conf. Interval] |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>educ</td>
<td>13.05067</td>
<td>0.0719784</td>
<td>181.31</td>
<td>0.000</td>
<td>12.90445 13.19688</td>
</tr>
</tbody>
</table>