Multiple Imputation using Chained Equations: A Comparison of Stata, SAS, IVEware and R

Presented by Pat Berglund
Summer Institute Program
Surv meth 616 - July 21, 2014
Imputation using Chained Equations

- This presentation demonstrates use of Multiple Imputation of missing data using the “chained equations/sequential regression/FCS/ice/mice” method (all refer to the same approach)

- The data set (subset of the NCS-R data, n=5692) has an arbitrary missing data pattern with two categorical variables that require imputation

- The covariates are both continuous and categorical

- This process is demonstrated in four software packages for a comparison of software usage and results:
  - SAS (v9.4)
  - Stata (13.1)
  - IVEware (0.1 SAS Callable)
  - R (3.0.1) with additional packages (mice, foreign, mitools, etc.)
Chained Equations Approach

- This method is widely used in practice as it handles complex missing data problems relatively easily:
  - Some of the benefits of the chained equations approach are that each model can be specified as desired, i.e. you can declare exactly the type of model to be used and predictors included as covariates in SAS and Stata but not IVEware which performs this type of decision making for you
  - This method handles arbitrary missing data patterns with categorical and continuous variables easily, widely used in practice due to ease of implementation and reliable results
  - Another advantage is that the variable with the least amount of missing data is imputed first and then used in subsequent imputations, then the next variable with the 2nd least amount of missing data is imputed and used in subsequent imputations, etc..

- A disadvantage (for the statistically inclined) is lack of theoretical foundation yet results are robust and generally reliable, see Van Buuren (2007 and 2012) for more detail
Applications

- This example uses Part 2 NCS-R data with an arbitrary missing data pattern and a mix of continuous and categorical variables as donors.
- This imputation performed using SAS, Stata, IVEware (runs under SAS or as a stand-alone tool), and R 3.0.2.
- A comparison of code/results is included.
The data used, **ncsr2_v12.dta** is a subset of the NCS-R data set:
- Variables included are:
  - **sex** (categorical 0=FEMALE 1=MALE)
  - **region** (categorical 1=NE 2=MW 3=SOUTH 4=WEST)
  - **age** (continuous age in years)
  - **str** (continuous strata for complex sample design)
  - **secu** (categorical cluster/PSU for complex sample design)
  - **finalp2wt** (continuous final part 2 weight)
  - **racecat_** (categorical 1=WHITE 2=HISPANIC 3=BLACK 4=OTHER)
  - **educat** (categorical 1=0-11 YRS 2=12 YRS 3=13-15 YRS 4=16+ YRS)
  - **mde** (categorical 1=YES major depressive episode 0=NO MDE)
  - **Str_secu** (categorical combined str and secu variable)
Use of the *misstable patterns* command shows missing data on the Major Depressive Episode indicator (mde) and education level in categories (educat), full n=5692

```
use "P:\pberg\Statistics.com Missing Data Class 2012\ncsr2_v12.dta", clear
.misstable patterns

Missing-value patterns
   (1 means complete)

          |   Pattern | Percent |
----------|-----------|---------|
         1  |  1  1     | 93%     |
         4  |  1  0     | 4%      |
         3  |  0  1     | 3%      |
----------|-----------|---------|
         1  |           | 100%    |

Variables are (1) mde (2) educat
```

```
.misstable tree

 Nested pattern of missing values

educat   mde
----------
  4%     0%
  96%    3%
  93%    4

(percent missing listed first)
```
The `misstable summarize` command shows:

- 165 records with missing on the mde indicator (0,1 values)
- 235 with missing data on the educat variable (categorical education levels ranging from 1 to 4)
## Variables Used in the Imputation

```
. 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>sex</td>
<td>5692</td>
<td>.4184821</td>
<td>.4933534</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>region</td>
<td>5692</td>
<td>2.575018</td>
<td>1.021898</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>age</td>
<td>5692</td>
<td>43.37807</td>
<td>16.5794</td>
<td>17</td>
<td>98</td>
</tr>
<tr>
<td>str</td>
<td>5692</td>
<td>26.37878</td>
<td>11.15588</td>
<td>1</td>
<td>42</td>
</tr>
<tr>
<td>secu</td>
<td>5692</td>
<td>1.505271</td>
<td>.5000161</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>finalp2wt</td>
<td>5692</td>
<td>1</td>
<td>.9582254</td>
<td>.1144058</td>
<td>10.10207</td>
</tr>
<tr>
<td>racecat_</td>
<td>5692</td>
<td>3.423226</td>
<td>1.025544</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>educat</td>
<td>5457</td>
<td>2.651457</td>
<td>1.012665</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>mde</td>
<td>5527</td>
<td>.3155419</td>
<td>.4647734</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>str_secu</td>
<td>5692</td>
<td>265.293</td>
<td>111.5497</td>
<td>1</td>
<td>422</td>
</tr>
</tbody>
</table>
```

- The variables `sex`, `region`, `str`, `secu`, `racecat_`, `educat`, `str_secu` and `mde` are categorical with fully observed data on all but `educat` and `mde`.
- The variables `age` and `finalp2wt` are continuous with fully observed data and `sampleid` is a ID variable that is character (not used in imputation).
Imputation of Education and Major Depressive Episode

• Set up for imputation

• . mi set mlong
• . mi register imputed mde educat
• (400 m=0 obs. now marked as incomplete)
• . mi register regular sex region racecat_ age finalp2wt str secu str_secu

. mi impute chained (logit) mde (ologit) educat=i.sex i.region i.racecat_ age finalp2wt i.str_secu , add(5) rseed(2012)

Conditional models:
  mde: logit mde i.educat i.sex i.region i.racecat_ age finalp2wt i.str_secu
  educat: ologit educat i.mde i.sex i.region i.racecat_ age finalp2wt i.str_secu

Performing chained iterations ...

Multivariate imputation
Imputations = 5
Chained equations
added = 5
Updated = 0
Imputed: m=1 through m=5
Initialization: monotone
Iterations = 50
Burn-in = 10

mde: logistic regression
educat: ordered logistic regression

<table>
<thead>
<tr>
<th>Variable</th>
<th>Complete</th>
<th>Incomplete</th>
<th>Imputed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>mde</td>
<td>5527</td>
<td>165</td>
<td>165</td>
<td>5692</td>
</tr>
<tr>
<td>educat</td>
<td>5457</td>
<td>235</td>
<td>235</td>
<td>5692</td>
</tr>
</tbody>
</table>

(complete + incomplete = total; imputed is the minimum across m of the number of filled-in observations.)
Use of *mi svyset* and *mi estimate, vartable* commands

```
   . mi svyset secu [pweight=finalp2wt], strata(str)

   pweight: finalp2wt
   VCE: linearized
   Single unit: missing
   Strata 1: str
   SU 1: secu
   FPC 1: <zero>

   . mi estimate, vartable: svy: proportion mde educat
```

Multiple-imputation estimates                   Imputations        =         5
Survey: Proportion estimation

Variance information

<table>
<thead>
<tr>
<th></th>
<th>Imputation variance</th>
<th>Relative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Within</td>
<td>Between</td>
</tr>
<tr>
<td>--------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>mde</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>.000039</td>
<td>7.7e-07</td>
</tr>
<tr>
<td>1</td>
<td>.000039</td>
<td>7.7e-07</td>
</tr>
<tr>
<td>educat</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>.00007</td>
<td>5.0e-06</td>
</tr>
<tr>
<td>2</td>
<td>.000139</td>
<td>4.0e-06</td>
</tr>
<tr>
<td>3</td>
<td>.000064</td>
<td>4.7e-06</td>
</tr>
<tr>
<td>4</td>
<td>.000113</td>
<td>3.6e-06</td>
</tr>
</tbody>
</table>

10
Use of mi estimate: vartable: svy: proportion command

Multiple-imputation estimates
Survey: Proportion estimation

Imputations = 5
Number of obs = 5692

Number of strata = 42
Number of PSUs = 84
Population size = 5692.0005

Average RVI = 0.0788
Largest FMI = 0.0879
Complete DF = 42

DF adjustment: Small sample
DF: min = 34.88
avg = 37.40
max = 39.00
Within VCE type: Linearized

<table>
<thead>
<tr>
<th>Proportion</th>
<th>Std. Err.</th>
<th>[95% Conf. Interval]</th>
</tr>
</thead>
<tbody>
<tr>
<td>mde</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>.8083499</td>
<td>.0062936</td>
</tr>
<tr>
<td>1</td>
<td>.1916501</td>
<td>.0062936</td>
</tr>
<tr>
<td>educat</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>.1676641</td>
<td>.0087298</td>
</tr>
<tr>
<td>2</td>
<td>.3245322</td>
<td>.0119858</td>
</tr>
<tr>
<td>3</td>
<td>.2755689</td>
<td>.0083708</td>
</tr>
<tr>
<td>4</td>
<td>.2322348</td>
<td>.0108185</td>
</tr>
</tbody>
</table>
Logistic Regression with Imputed Data Sets

. *mi estimate for logistic regression
. mi estimate: svy: logit mde i.sex i.region i.educat, or

Multiple-imputation estimates
Survey: Logistic regression
Number of strata = 42
Number of PSUs = 84

Imputations = 5
Population size = 5692.0005
Average RVI = 0.0464
Largest FMI = 0.0933
Complete DF = 42
DF: min = 34.46
avg = 38.43
max = 39.79

Model F test: Equal FMI
F( 7, 39.7) = 16.02
Prob > F = 0.0000

| mde | Coef. | Std. Err. | t  | P>|t|  | [95% Conf. Interval] |
|-----|-------|-----------|----|------|-------------------------|
|     |       |           |    |      |                         |
| 1.sex | -0.5111926 | 0.0638813 | -8.00 | 0.000 | -0.6403719 | -0.3820132 |
| region |     |           |    |      |                         |
| 2 | 0.0065904 | 0.1408085 | 0.05 | 0.963 | -0.2783006 | 0.2914814 |
| 3 | -0.1669361 | 0.1330025 | -1.26 | 0.217 | -0.4358414 | 0.1019692 |
| 4 | 0.039644 | 0.1319837 | 0.30 | 0.765 | -0.2271491 | 0.3064372 |
| educat |     |           |    |      |                         |
| 2 | 0.094891 | 0.1074464 | 0.88 | 0.383 | -0.1228383 | 0.3126204 |
| 3 | 0.2324122 | 0.1124072 | 2.07 | 0.046 | 0.004085 | 0.4607393 |
| 4 | 0.1770819 | 0.1114483 | 1.59 | 0.120 | -0.0482573 | 0.4024212 |
| _cons | -1.311143 | 0.1484784 | -8.83 | 0.000 | -1.611389 | -1.010897 |
SAS 9.3

- Application repeated using SAS 9.4
- SAS offers the FCS method for use with an arbitrary missing data pattern and continuous or categorical variables
- Performs “chained equations” as well, good comparison to Stata and other software
- Expectation is that imputed results will be similar to Stata output
options nocenter ls=135 ps=59 ;
proc mi nimpute=0 data=d.ncsr2_1 ;
run ;

Missing Data Patterns

Group  sex  region  age  str  secu  finalp2wt  racecat_  educat  mde  mde_imp  imp      str_secu      Freq   Percent
1    X    X       X    X    X     X          X         X       X    X        X        X             5292     92.97
2    X    X       X    X    X     X          X         X       .    X        X        X              165      2.90
3    X    X       X    X    X     X          .          X       X    X        X        X              235      4.13

Missing Data Patterns

---------------------------------------------Group Means---------------------------------------------
Group  sex  region  age  str  secu  finalp2wt  racecat_  educat
1  0.417989       2.579554      43.352419      26.376228       1.506236       0.992335       3.428193       2.649849
2  0.387879       2.454545      41.690909      26.957576       1.527273       1.084961       3.351515       2.703030
3  0.451064       2.557447      45.140426      26.029787       1.468085       1.112958              .

Missing Data Patterns

---------------------------------------------Group Means---------------------------------------------
Group  mde  mde_imp  educat_imp  str_secu
1  0.316515               0        0      135.268519
2             .        1.000000               0      273.141030
3  0.293617               0        1.000000      261.765957
Highlights of the SAS code include

- **5** imputed data sets
- **Class** statement to declare categorical variables
- Use of the **fcs logistic** statement requests logistic regression with chained equations method for imputation
  - The **(mde/details)** option produces model details per imputation
- The **var** statement lists the variables in order of those with fully observed data, then least amount of missing to most missing data

```
proc mi data=ncsr2_1 seed=876 nimpute=5 out=outfcs ;
class sex region racecat_ educat mde str_secu ;
fcs logistic (mde/details) logistic (educat) ;
var sex region age racecat_ str_secu finalp2wt mde educat ;
run;
```
## Imputation Details for Major Depressive Episode

### The MI Procedure

<table>
<thead>
<tr>
<th>Model Information</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data Set</strong></td>
</tr>
<tr>
<td><strong>Method</strong></td>
</tr>
<tr>
<td><strong>Number of Imputations</strong></td>
</tr>
<tr>
<td><strong>Number of Burn-in Iterations</strong></td>
</tr>
<tr>
<td><strong>Seed for random number generator</strong></td>
</tr>
</tbody>
</table>

### FCS Model Specification

<table>
<thead>
<tr>
<th>Method</th>
<th>Imputed Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>age, finalp2wt</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>mde, educat</td>
</tr>
<tr>
<td>Discriminant Function</td>
<td>sex, region, racecat, str_secu</td>
</tr>
</tbody>
</table>

### Logistic Models for FCS Method

<table>
<thead>
<tr>
<th>Imputed Variable</th>
<th>Effect</th>
<th>sex</th>
<th>region</th>
<th>racecat</th>
<th>str_secu</th>
<th>educat</th>
<th>Imputation</th>
</tr>
</thead>
<tbody>
<tr>
<td>mde</td>
<td>Intercept</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>1</td>
</tr>
<tr>
<td>mde</td>
<td>sex</td>
<td>0</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>-0.177406</td>
</tr>
<tr>
<td>mde</td>
<td>region</td>
<td>1.00000</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>1.575420</td>
</tr>
<tr>
<td>mde</td>
<td>region</td>
<td>2.00000</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>-0.475625</td>
</tr>
<tr>
<td>mde</td>
<td>region</td>
<td>3.00000</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>-0.827580</td>
</tr>
<tr>
<td>mde</td>
<td>age</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>0.055691</td>
</tr>
<tr>
<td>mde</td>
<td>racecat</td>
<td>1.00000</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>-0.045572</td>
</tr>
<tr>
<td>mde</td>
<td>racecat</td>
<td>2.00000</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>0.031858</td>
</tr>
<tr>
<td>mde</td>
<td>racecat</td>
<td>3.00000</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>-0.072845</td>
</tr>
<tr>
<td>mde</td>
<td>str_secu</td>
<td>.</td>
<td>11.00000</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>0.22327</td>
</tr>
<tr>
<td>mde</td>
<td>str_secu</td>
<td>.</td>
<td>12.00000</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>-1.674056</td>
</tr>
<tr>
<td>mde</td>
<td>str_secu</td>
<td>.</td>
<td>21.00000</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>-1.846103</td>
</tr>
<tr>
<td>mde</td>
<td>str_secu</td>
<td>.</td>
<td>22.00000</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>-1.321143</td>
</tr>
<tr>
<td>mde</td>
<td>str_secu</td>
<td>.</td>
<td>31.00000</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>-0.546819</td>
</tr>
</tbody>
</table>
Crosstabulations of Major Depressive Episode by Imputed Flag and Imputation (Partial Output)

### Table 1 of mde by mde_imp

<table>
<thead>
<tr>
<th>mde (Major Depressive Episode)</th>
<th>mde_imp</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>Total</td>
</tr>
<tr>
<td>0</td>
<td>3783</td>
<td>98</td>
<td>3881</td>
</tr>
<tr>
<td>1</td>
<td>1744</td>
<td>67</td>
<td>1811</td>
</tr>
<tr>
<td>Total</td>
<td>5527</td>
<td>165</td>
<td>5692</td>
</tr>
</tbody>
</table>

### Table 2 of mde by mde_imp

<table>
<thead>
<tr>
<th>mde (Major Depressive Episode)</th>
<th>mde_imp</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>Total</td>
</tr>
<tr>
<td>0</td>
<td>3783</td>
<td>100</td>
<td>3883</td>
</tr>
<tr>
<td>1</td>
<td>1744</td>
<td>65</td>
<td>1809</td>
</tr>
<tr>
<td>Total</td>
<td>5527</td>
<td>165</td>
<td>5692</td>
</tr>
</tbody>
</table>

proc freq data=outfcs ;
tables _imputation_ *mde*mde_imp / missing
    nopercent nocol nocum norow ;
run ;

- Tables show how the values of imputed MDE change over the 5 imputed data sets (just 2 of 5 shown here)
- This is expected due to the different logistic models run and different values assigned per imputation
Analysis of Imputed Data Sets with PROC SURVEYLOGISTIC

- PROC SURVEYLOGISTIC is used with the binary dependent variable (imputed) mde as well as imputed educat, and region and sex
- Output data set called “outparms” consists of parameter estimates with associated standard errors

```plaintext
proc surveylogistic data=outfcs ;
strata str ; cluster secu ; weight finalp2wt ; *complex sample/weight ;
class sex (ref='0') educat (ref='1') region (ref='1') / param=ref ;
model mde (event='1') = sex educat region ; *model probability of mde=1 ;
by _imputation_ ; *run for each imputed data set ;
ods output parameterestimates = outparms ; * save output data set for next step PROC MIANALYZE ;
run ;
```
PROC MIANALYZE SAS Code

```sas
proc mianalyze parms (classvar=classval)=outparms;
class sex educat region;
modeleffects intercept sex educat region;
run;
```

- Use of `(classvar=classval)` refers to the `classval` variable that is included in the `parms` output data set:
  - specifies the class variable’s values as presented in the `outparms` data (with one category omitted)
- The CLASS statement is needed in PROC MIANALYZE to identify categorical variables rather than treating variables as continuous.
- The MODELEFFECTS statement uses the same model specification from PROC SURVEYLOGISTIC (in the previous step), including the intercept.
Variance Information and Parameter Estimates incorporate the survey correction (PROC SURVEYLOGISTIC) and the imputation variability.

Results are similar to Stata output with small changes in the region variable.

Overall conclusions would be similar to Stata as well.
Application in IVEware

- IVEware runs under SAS in this example (also possible to run as a standalone version, see iveware.org for newest versions)
- This tool incorporates imputation (%impute macro) and complex sample design adjustments using the Jackknife Repeated Replication method for variance estimation (%regress and %describe macros)
- The IVEware macros run as SAS macros within program (really not necessary to understand SAS macros in-depth to run this program)
LIBNAME d 'p:\pberg\Statistics.com Missing Data Class 2012' ;
* Read in data set ;
data app4 ;
   set d.ncsr2_1 ;
* Recode the dependent variable to make highest category (no) the omitted ;
   if mde=0 then mde_r=2 ; else if mde=1 then mde_r=1 ;  else mde_r=. ;
run ;
proc freq ;
   tables mde_r ;
run ;

* use IVEware %impute to multiply impute missing data ;
%impute (name=app4, setup=new, dir=. ) ;
datain  app4  ;
dataout app4_imp ;
default continuous ;
categorical sex region racecat_ educat mde_r str_secu ;
transfer sampleid mde_imp educat_imp str secu mde ;
multiples 5 ;  * create m=5 imputed data sets , this is the default ;
seed 876 ;
run ;
Output from Imputation #5

IVWare Iterative Imputation Procedure, Tue Jul 15 13:39:24 2014

Imputation 5

Variable Observed Imputed Double counted
sex 5692 0 0
region 5692 0 0
age 5692 0 0
finalp2wt 5692 0 0
racecat_ 5692 0 0
educat 5457 235 0
str_secu 5692 0 0
mde_r 5527 165 0

Variable educat

<table>
<thead>
<tr>
<th>Code</th>
<th>Observed</th>
<th>Imputed</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Freq</td>
<td>Per</td>
<td>Freq</td>
</tr>
<tr>
<td>1</td>
<td>813</td>
<td>14.90</td>
<td>35</td>
</tr>
<tr>
<td>2</td>
<td>1641</td>
<td>30.07</td>
<td>75</td>
</tr>
<tr>
<td>3</td>
<td>1638</td>
<td>30.02</td>
<td>74</td>
</tr>
<tr>
<td>4</td>
<td>1365</td>
<td>25.01</td>
<td>51</td>
</tr>
<tr>
<td>Total</td>
<td>5457</td>
<td>100.00</td>
<td>235</td>
</tr>
</tbody>
</table>

Variable mde_r

<table>
<thead>
<tr>
<th>Code</th>
<th>Observed</th>
<th>Imputed</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Freq</td>
<td>Per</td>
<td>Freq</td>
</tr>
<tr>
<td>1</td>
<td>1744</td>
<td>31.55</td>
<td>48</td>
</tr>
<tr>
<td>2</td>
<td>3783</td>
<td>68.45</td>
<td>117</td>
</tr>
<tr>
<td>Total</td>
<td>5527</td>
<td>100.00</td>
<td>165</td>
</tr>
</tbody>
</table>
Logistic Regression Using Imputed Data Sets

* use %putdata to produce 5 separate data sets for correct MI estimation ;
  %putdata(name=app4,dir=., mult=1,dataout=d.imp1 );
  %putdata(name=app4,dir=., mult=2,dataout=d.imp2 );
  %putdata(name=app4,dir=., mult=3,dataout=d.imp3 );
  %putdata(name=app4,dir=., mult=4,dataout=d.imp4 );
  %putdata(name=app4,dir=., mult=5,dataout=d.imp5 );

  %regress (name=app4_2 , setup=new, dir=. ) ;
  datain  d.imp1 d.imp2 d.imp3 d.imp4 d.imp5 ;
  stratum str ;
  cluster secu ;
  weight  finalp2wt ;
  categorical sex educat region ;
  predictor sex educat region ;
  dependent mde_r ;
  link logistic ;
  run ;
By default, IVEware omits the highest category and so if you want to match Stata, you could use dummy variables rather than categorical variables.

Use of stratum, cluster and weight ensure the right complex sample design correction as well as accounting for imputation variability.

Note that we are predicting the probability that MDE=1 through use of the mde_r variable (1=has major depressive episode, 2=none).
Logistic Regression Output, continued

All imputations
Valid cases               5692
Sum weights        5692.000487
Degr freedom        179.406391
-2 LogLike         5487.725767

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>Std Error</th>
<th>Wald test</th>
<th>Prob &gt; Chi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.6017202</td>
<td>0.0937348</td>
<td>291.99244</td>
<td>0.00000</td>
</tr>
<tr>
<td>sex</td>
<td>0.5050432</td>
<td>0.0639964</td>
<td>62.27970</td>
<td>0.00000</td>
</tr>
<tr>
<td>educat.1</td>
<td>-0.1806331</td>
<td>0.1147336</td>
<td>2.47864</td>
<td>0.11540</td>
</tr>
<tr>
<td>educat.2</td>
<td>-0.0954590</td>
<td>0.0895300</td>
<td>1.13683</td>
<td>0.28632</td>
</tr>
<tr>
<td>educat.3</td>
<td>0.0602163</td>
<td>0.0795369</td>
<td>0.57318</td>
<td>0.44900</td>
</tr>
<tr>
<td>region.1</td>
<td>-0.0341966</td>
<td>0.1439643</td>
<td>0.05642</td>
<td>0.81224</td>
</tr>
<tr>
<td>region.2</td>
<td>-0.0374815</td>
<td>0.1034633</td>
<td>0.13124</td>
<td>0.71715</td>
</tr>
<tr>
<td>region.3</td>
<td>-0.2036425</td>
<td>0.0916866</td>
<td>4.93316</td>
<td>0.02635</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Odds Ratio</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.6570571</td>
<td>1.4604735 - 1.8801014</td>
</tr>
<tr>
<td>sex</td>
<td>0.834716</td>
<td>0.6656214 - 1.0468315</td>
</tr>
<tr>
<td>educat.1</td>
<td>0.9089556</td>
<td>0.7617580 - 1.0845969</td>
</tr>
<tr>
<td>educat.2</td>
<td>1.062062</td>
<td>0.9077993 - 1.2425485</td>
</tr>
<tr>
<td>educat.3</td>
<td>0.9663815</td>
<td>0.7274006 - 1.2838775</td>
</tr>
<tr>
<td>region.1</td>
<td>0.9632122</td>
<td>0.7853363 - 1.1813764</td>
</tr>
<tr>
<td>region.2</td>
<td>0.8157540</td>
<td>0.6807464 - 0.9775367</td>
</tr>
<tr>
<td>region.3</td>
<td>1.30510</td>
<td>0.1530469 - 24.84526</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Design Effect</th>
<th>SRS</th>
<th>% Diff SRS v Est</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.21094</td>
<td>-0.9769619</td>
<td>-39.00545</td>
</tr>
<tr>
<td>sex</td>
<td>1.12449</td>
<td>0.5185654</td>
<td>2.67744</td>
</tr>
<tr>
<td>educat.1</td>
<td>1.39625</td>
<td>-0.1475597</td>
<td>-18.30971</td>
</tr>
<tr>
<td>educat.2</td>
<td>1.15261</td>
<td>-0.0470697</td>
<td>-50.69118</td>
</tr>
<tr>
<td>educat.3</td>
<td>1.04220</td>
<td>0.0265876</td>
<td>-55.84643</td>
</tr>
<tr>
<td>region.1</td>
<td>2.46956</td>
<td>-0.1698988</td>
<td>396.83027</td>
</tr>
<tr>
<td>region.2</td>
<td>1.56901</td>
<td>-0.0118203</td>
<td>-68.46355</td>
</tr>
<tr>
<td>region.3</td>
<td>1.30510</td>
<td>-0.1530469</td>
<td>-24.84526</td>
</tr>
</tbody>
</table>
Application in R

- In R we use **mice** and **mitools** with the survey package for design adjusted variance estimates
- Mice uses chained equations for the imputation
- Mitools is a survey related package for analysis of imputed survey data
- Survey is an R package for analysis of survey data
- This example uses mice defaults for the most part but the imputation can be customized if desired
R Command Syntax

# R code for SI 616 2014 Presentation (July 21, 2014)
# Berglund
# data management
# load packages using library command
library(foreign)
library(mi)
library(mice)
# read Stata format data set into R
a <- read.dta("P:/pberg/statistics.com Missing Data Class 2012/ncsr2_v12.dta")
summary(a)

# create factor variables
a$sex <- factor(a$sex)
a$educat <- factor(a$educat)
a$region <- factor(a$region)
a$str_secu <- factor(a$str_secu)

# obtain information about missing data
inf <- mi.info(a)
# print info about missing data
inf

# use mice to impute and pool
library(mice)
imp <- mice(a,n.imp=5,seed=1934)
summary(imp)

# convert mids to data useable for work in mitools
library(mitools)
mydata <- imputationList(lapply(1:5, complete, x=imp))
summary(mydata)

# set survey design
library(survey)
des <- svydesign(id=~secu, strat=~str, weight=~finalp2wt, data=(mydata), nest=TRUE)
summary(des)

# run design based model with svyglm using 5 imputed data sets contained in des (from mydata)
fit2 <- with (des, svyglm (mde ~ sex + educat + region, family=quasibinomial))
summary(MIcombine(fit2))
### Data Setup and Information about Missing Data

```r
> # obtain information about missing data
> inf <- mi.info(a)
> # print info about missing data
> inf

<table>
<thead>
<tr>
<th>names</th>
<th>include</th>
<th>order</th>
<th>number.mis</th>
<th>all.mis</th>
<th>type</th>
<th>collinear</th>
</tr>
</thead>
<tbody>
<tr>
<td>sampleid</td>
<td>Yes</td>
<td>NA</td>
<td>0</td>
<td>No</td>
<td>unordered-categorical</td>
<td>No</td>
</tr>
<tr>
<td>sex</td>
<td>Yes</td>
<td>NA</td>
<td>0</td>
<td>No</td>
<td>binary</td>
<td>No</td>
</tr>
<tr>
<td>region</td>
<td>Yes</td>
<td>NA</td>
<td>0</td>
<td>No</td>
<td>unordered-categorical</td>
<td>No</td>
</tr>
<tr>
<td>age</td>
<td>Yes</td>
<td>NA</td>
<td>0</td>
<td>No</td>
<td>positive-continuous</td>
<td>No</td>
</tr>
<tr>
<td>str</td>
<td>Yes</td>
<td>NA</td>
<td>0</td>
<td>No</td>
<td>positive-continuous</td>
<td>No</td>
</tr>
<tr>
<td>secu</td>
<td>Yes</td>
<td>NA</td>
<td>0</td>
<td>No</td>
<td>binary</td>
<td>No</td>
</tr>
<tr>
<td>finalp2wt</td>
<td>Yes</td>
<td>NA</td>
<td>0</td>
<td>No</td>
<td>positive-continuous</td>
<td>No</td>
</tr>
<tr>
<td>racecat_</td>
<td>Yes</td>
<td>NA</td>
<td>0</td>
<td>No</td>
<td>ordered-categorical</td>
<td>No</td>
</tr>
<tr>
<td>educat</td>
<td>Yes</td>
<td>1</td>
<td>235</td>
<td>No</td>
<td>unordered-categorical</td>
<td>No</td>
</tr>
<tr>
<td>mde</td>
<td>Yes</td>
<td>2</td>
<td>165</td>
<td>No</td>
<td>binary</td>
<td>No</td>
</tr>
<tr>
<td>str_secu</td>
<td>Yes</td>
<td>NA</td>
<td>0</td>
<td>No</td>
<td>unordered-categorical</td>
<td>No</td>
</tr>
</tbody>
</table>
```

>
Mice: Impute Missing Data on EDUCAT and MDE

```r
> # use mice to impute and pool
> library(mice)
> imp <- mice(a,n.imp=5,seed=1934)
> summary(imp)
```

Multiply imputed data set
Call: mice(data = a, seed = 1934, n.imp = 5)
Number of multiple imputations: 5

<table>
<thead>
<tr>
<th>Missing cells per column:</th>
</tr>
</thead>
<tbody>
<tr>
<td>sampleid       sex    region       age       str      secu finalp2wt  racecat_    educat       mde</td>
</tr>
<tr>
<td>0         0         0         0         0         0         0         0       235       165</td>
</tr>
<tr>
<td>str_secu</td>
</tr>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

Imputation methods:

```
>     sampleid       sex    region       age       str      secu finalp2wt  racecat_    educat       mde
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&quot;&quot;</td>
<td>&quot;&quot;</td>
<td>&quot;&quot;</td>
<td>&quot;&quot;</td>
<td>&quot;&quot;</td>
<td>&quot;&quot;</td>
<td>&quot;&quot;</td>
<td>&quot;&quot;</td>
<td>&quot;&quot;</td>
<td>&quot;&quot;</td>
</tr>
</tbody>
</table>

| str_secu     |
| "" |

VisitSequence:

```
educat  mde
9     10
```

PredictorMatrix:

```
sampleid sex region age str secu finalp2wt racecat_ educat mde str_secu
sampleid    0   0      0   0   0    0         0        0      0   0        0
sex         0   0      0   0   0    0         0        0      0   0        0
region      0   0      0   0   0    0         0        0      0   0        0
age         0   0      0   0   0    0         0        0      0   0        0
str          0   0      0   0   0    0         0        0      0   0        0
secu         0   0      0   0   0    0         0        0      0   0        0
finalp2wt    0   0      0   0   0    0         0        0      0   0        0
racecat_     0   0      0   0   0    0         0        0      0   0        0
educat       0   1      1   1   1    1         1        1      0   1        0
mde          0   1      1   1   1    1         1        1      1   0        0
str_secu     0   0      0   0   0    0         0        0      0   0        0
```

Random generator seed value: 1934
Set Survey Design and Use Micombine for Survey Logistic Regression with Imputed Data Sets

```r
> # convert mids to data useable for work in mitools
> library(mitools)
> mydata <- imputationList(lapply(1:5, complete, x=imp))
> summary(mydata)
    Length Class Mode
imputations 5   -none-  list
call         2   -none-   call
>
> # set survey design
> library(survey)
> des <- svydesign(id=~secu, strat=~str, weight=~finalp2wt, data=(mydata), nest=TRUE)
> summary(des)
    Length Class Mode
designs 5    -none-  list
call         6   -none-   call
>
> # run design based model with svyglm using 5 imputed data sets contained in des (from mydata)
> fit2 <- with (des, svyglm (mde ~ sex + educat + region, family=quasibinomial))
> summary(MIcombine(fit2))
Multiple imputation results:
  with(des, svyglm(mde ~ sex + educat + region, family = quasibinomial))
  MIcombine.default(fit2)

results   se    (lower     upper) missInfo
(Intercept) -1.48749336 0.14096000 -1.76381159 -1.2111751 2 %
sex2        -0.3972207 0.08145831 -0.55638193 -0.2370622 1 %
educat2    0.11943740 0.10932972 0.00530298  0.2341778 9 %
educat3    0.25810927 0.11257402 0.03738999  0.4788286 3 %
educat4    0.18291288 0.12037457 -0.09530298  0.4548465 3 %
region2    0.08537139 0.13342511 -0.17615071  0.3468935 1 %
region3   -0.06108288 0.13851353 -0.33256725  0.2104027 1 %
region4    0.10233875 0.13072914 -0.15390103  0.3585785 1 %
```
Comparison of Results from Stata, SAS, IVEware and R

STATA

|           | Coef. | Std. Err. |       t | P>|t| | 95% Conf. Interval |
|-----------|-------|-----------|--------|-----|---------------------|
| 1.sex     | -0.5111926 | 0.0638813 | -8.00  | 0.000 | -0.6403719, -0.3820132 |
| region    | 2     | 0.0065904 | 0.1508085 | 0.05  | 0.963, -0.2783006, 0.2914814 |
|           | 3     | -0.1669361 | 0.133025 | -1.26 | 0.217, -0.4358414, 0.1019632 |
|           | 4     | 0.039644 | 0.1319387 | 0.30  | 0.765, -0.2271491, 0.3064372 |
| educat    | 2     | 0.094891 | 0.1074464 | 0.88  | 0.383, -0.1228383, 0.3126204 |
|           | 3     | 0.2324122 | 0.1124072 | 2.07  | 0.046, 0.004085, 0.4607393 |
|           | 4     | 0.1770819 | 0.114483 | 1.59  | 0.120, -0.0482573, 0.4024212 |
| _cons     |       | -1.311143 | 0.1484784 | -8.83 | 0.000, -1.611389, -1.010897 |

SAS

The MEANALYZE Procedure

Model Information
PARAMS Data Set WORK-OUTPARAMS
Number of Imputations 5

Variance Information

<table>
<thead>
<tr>
<th>Parameter</th>
<th>sex</th>
<th>educat</th>
<th>region</th>
<th>Variance</th>
<th>Between</th>
<th>Within</th>
<th>Total</th>
<th>DF</th>
<th>Relative Increase in Variance</th>
<th>Fraction Missing Information</th>
<th>Relative Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td></td>
<td></td>
<td></td>
<td>0.000137</td>
<td>0.020969</td>
<td>0.021074</td>
<td>0.02</td>
<td>1.02</td>
<td>0.007083</td>
<td>0.007051</td>
<td>0.989432</td>
</tr>
<tr>
<td>sex</td>
<td>1</td>
<td></td>
<td></td>
<td>0.00025634</td>
<td>0.064869</td>
<td>0.053960</td>
<td>10.490</td>
<td>0.056318</td>
<td>0.006298</td>
<td>0.98742</td>
<td></td>
</tr>
<tr>
<td>educat</td>
<td>2</td>
<td></td>
<td></td>
<td>0.000404</td>
<td>0.11661</td>
<td>0.015636</td>
<td>1.4621</td>
<td>0.042660</td>
<td>0.041085</td>
<td>0.96168</td>
<td></td>
</tr>
<tr>
<td>educat</td>
<td>3</td>
<td></td>
<td></td>
<td>0.011276</td>
<td>0.119615</td>
<td>0.024647</td>
<td>13.163</td>
<td>1.228441</td>
<td>0.007083</td>
<td>0.989777</td>
<td></td>
</tr>
<tr>
<td>educat</td>
<td>4</td>
<td></td>
<td></td>
<td>0.1614145</td>
<td>0.125429</td>
<td>0.84723</td>
<td>16.496</td>
<td>0.870433</td>
<td>0.544492</td>
<td>0.961795</td>
<td></td>
</tr>
<tr>
<td>region</td>
<td>2</td>
<td></td>
<td></td>
<td>0.000275</td>
<td>0.168342</td>
<td>0.172366</td>
<td>7.6435</td>
<td>0.023412</td>
<td>0.023132</td>
<td>0.966396</td>
<td></td>
</tr>
<tr>
<td>region</td>
<td>3</td>
<td></td>
<td></td>
<td>0.000329</td>
<td>0.16842</td>
<td>0.172366</td>
<td>7.6435</td>
<td>0.023412</td>
<td>0.023132</td>
<td>0.966396</td>
<td></td>
</tr>
<tr>
<td>region</td>
<td>4</td>
<td></td>
<td></td>
<td>0.000329</td>
<td>0.16842</td>
<td>0.172366</td>
<td>7.6435</td>
<td>0.023412</td>
<td>0.023132</td>
<td>0.966396</td>
<td></td>
</tr>
</tbody>
</table>

Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>sex</th>
<th>educat</th>
<th>region</th>
<th>Estimate</th>
<th>Std Error</th>
<th>95% Confidence Limits</th>
<th>DF</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Theta0</th>
<th>t for H0: Parameter=Theta0</th>
<th>Pr &gt;</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td></td>
<td></td>
<td></td>
<td>-1.3922977</td>
<td>0.350398</td>
<td>-1.08772, -1.70911</td>
<td>0.3520</td>
<td>-1.314556</td>
<td>-1.284026</td>
<td>0</td>
<td>-9.97 &lt; 0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sex</td>
<td>1</td>
<td></td>
<td></td>
<td>0.054498</td>
<td>0.109598</td>
<td>0.04165, 0.05733</td>
<td>0.3520</td>
<td>-0.519147</td>
<td>-0.49623</td>
<td>0</td>
<td>-7.21 &lt; 0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>educat</td>
<td>2</td>
<td></td>
<td></td>
<td>0.059717</td>
<td>0.150768</td>
<td>0.03065, 0.10347</td>
<td>0.3520</td>
<td>0.071355</td>
<td>0.12016</td>
<td>0</td>
<td>9.87 &lt; 0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>educat</td>
<td>3</td>
<td></td>
<td></td>
<td>0.097676</td>
<td>0.166927</td>
<td>-0.23921, 0.43276</td>
<td>0.3520</td>
<td>0.17813</td>
<td>0.36760</td>
<td>0</td>
<td>1.85 0.0654</td>
<td></td>
<td></td>
</tr>
<tr>
<td>educat</td>
<td>4</td>
<td></td>
<td></td>
<td>0.137402</td>
<td>0.18304</td>
<td>-0.35016, 0.62097</td>
<td>0.3520</td>
<td>0.24695</td>
<td>0.27695</td>
<td>0</td>
<td>1.37 &lt; 0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>region</td>
<td>2</td>
<td></td>
<td></td>
<td>0.002392</td>
<td>0.125321</td>
<td>-0.29803, 0.27903</td>
<td>0.3520</td>
<td>-0.04091</td>
<td>-0.00112</td>
<td>0</td>
<td>-9.17 &lt; 0.0052</td>
<td></td>
<td></td>
</tr>
<tr>
<td>region</td>
<td>3</td>
<td></td>
<td></td>
<td>-0.18902</td>
<td>0.129125</td>
<td>-0.44702, 0.06800</td>
<td>0.3520</td>
<td>-0.21901</td>
<td>-0.16718</td>
<td>0</td>
<td>-1.51 &lt; 0.1320</td>
<td></td>
<td></td>
</tr>
<tr>
<td>region</td>
<td>4</td>
<td></td>
<td></td>
<td>0.019050</td>
<td>0.131287</td>
<td>-0.23850, 0.25262</td>
<td>0.3520</td>
<td>0.00376</td>
<td>0.05811</td>
<td>0</td>
<td>0.15 &lt; 0.8864</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Comparison of Results from Stata, SAS, IVEware and R, continued

IVEWARE:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>Std Error</th>
<th>Wald test</th>
<th>Prob &gt; Chi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.6017202</td>
<td>0.0937348</td>
<td>291.99244</td>
<td>0.00000</td>
</tr>
<tr>
<td>sex</td>
<td>0.5050432</td>
<td>0.0639964</td>
<td>62.27970</td>
<td>0.00000</td>
</tr>
<tr>
<td>educat.1</td>
<td>-0.1806331</td>
<td>0.1147336</td>
<td>2.47864</td>
<td>0.11540</td>
</tr>
<tr>
<td>educat.2</td>
<td>-0.0954590</td>
<td>0.0895300</td>
<td>1.13683</td>
<td>0.28632</td>
</tr>
<tr>
<td>educat.3</td>
<td>0.0602163</td>
<td>0.0795369</td>
<td>0.57318</td>
<td>0.44900</td>
</tr>
<tr>
<td>region.1</td>
<td>-0.031966</td>
<td>0.1439643</td>
<td>0.05642</td>
<td>0.81224</td>
</tr>
<tr>
<td>region.2</td>
<td>-0.0374815</td>
<td>0.1034633</td>
<td>0.13124</td>
<td>0.71715</td>
</tr>
<tr>
<td>region.3</td>
<td>-0.2036425</td>
<td>0.0916866</td>
<td>4.93316</td>
<td>0.02635</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Odds Ratio</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.6570571</td>
<td>1.4604735 - 1.8801014</td>
</tr>
<tr>
<td>sex</td>
<td>1.5030452</td>
<td>1.4065214 - 1.6098315</td>
</tr>
<tr>
<td>educat.1</td>
<td>0.8345746</td>
<td>0.6616214 - 1.0468315</td>
</tr>
<tr>
<td>educat.2</td>
<td>1.0954590</td>
<td>0.8617580 - 1.345969</td>
</tr>
<tr>
<td>educat.3</td>
<td>1.0602163</td>
<td>0.9077993 - 1.2425485</td>
</tr>
<tr>
<td>region.1</td>
<td>0.9631222</td>
<td>0.7274006 - 1.2837775</td>
</tr>
<tr>
<td>region.2</td>
<td>0.9631222</td>
<td>0.7853363 - 1.1813764</td>
</tr>
<tr>
<td>region.3</td>
<td>0.8157540</td>
<td>0.6807464 - 0.9775367</td>
</tr>
</tbody>
</table>

> # run design based model with svyglm using 5 imputed data sets contained in des (from mydata)
> fit2 <- with (des, svyglm (mde ~ sex + educat + region, family=quasibinomial))
> summary(MICombine(fit2))
Multiple imputation results:

with(des, svyglm(mde ~ sex + educat + region, family = quasibinomial))

<table>
<thead>
<tr>
<th>results</th>
<th>se</th>
<th>(lower</th>
<th>upper</th>
<th>missInfo</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-1.48749336</td>
<td>0.14096000</td>
<td>-1.76381159</td>
<td>-1.2111751</td>
</tr>
<tr>
<td>sex</td>
<td>-0.39672207</td>
<td>0.08145831</td>
<td>-0.55638193</td>
<td>-0.2370622</td>
</tr>
<tr>
<td>educat2</td>
<td>0.11943740</td>
<td>0.1396432</td>
<td>0.05642</td>
<td>0.81224</td>
</tr>
<tr>
<td>educat3</td>
<td>0.1034633</td>
<td>0.0602163</td>
<td>0.13124</td>
<td>0.71715</td>
</tr>
<tr>
<td>region2</td>
<td>0.10233875</td>
<td>0.13072914</td>
<td>-0.15390103</td>
<td>0.3585785</td>
</tr>
</tbody>
</table>
Summary

- All four major software tools provide convenient methods for imputation and analysis of survey data sets

- The analyst can select from a variety of imputation model types or let the program select for you (IVEware)

- MI by chained equations offers a very flexible and easy method of imputation of complex missing data patterns
Stata Code

use "P:\pberg\Statistics.com Missing Data Class 2012\ncsr2_v12.dta", clear
misstable patterns
misstable tree
misstable summarize

mi set mlong
mi register imputed mde educat
mi register regular sex region racecat_age finalp2wt str secu str_secu

mi impute chained (logit) mde (ologit) educat=i.sex i.region i.racecat_age finalp2wt i.str_secu , add(5) rseed(2012)

* set survey variables for mi analysis with complex sample survey data
mi svyset secu [pweight=finalp2wt], strata(str)
* examine mean of price by each imputation
mi estimate, vartable: svy: proportion mde educat

*mi estimate for logistic regression
mi estimate: svy: logit mde i.sex i.region i.educat, or
LIBNAME d 'p:\pberg\Statistics.com Missing Data Class 2012' ;
options nofmterr nocenter nodate nonumber ;
data ncsr2_1 ;
set d.ncsr2_1 ;
if mde eq . then mde_imp=1 ; else mde_imp=0 ;
if educat eq . then educat_imp=1 ; else educat_imp=0 ;
run ;
proc freq ;
tables mde_imp educat_imp ;
run ;
proc mi nimpute=0 data=d.ncsr2_1 ;
run ;
proc mi data=ncsr2_1 seed=876 nimpute=5 out=outfcs ;
class sex region racecat_ educat mde str_secu ;
fcs logistic (mde/ details) logistic (educat) ;
var sex region age racecat_ str_secu finalp2wt mde educat ;
run;
* check imputed values of educat and mde ;
proc freq data=outfcs ;
tables _imputation_*mde*mde_imp / missing nopercent nocol nocum norow ;
run ;
*step 3 analyze combined datasets using logistic regression*** ;
proc surveylogistic data=outfcs ;
strata str ; cluster secu ; weight finalp2wt ;
class sex (ref='0') educat (ref='1') region (ref='1') / param=ref ;
model mde (event='1') = sex educat region ;
by _imputation_ ;
ods output parameterestimates = outparms ;
run ;
proc print data=outparms ;
run ;
*use mianalyze on combined imputed datasets* ;
options orientation=landscape ls=165 ps=45 ;
proc mianalyze parms (classvar=classval)=outparms ;
class sex educat region ;
modeleffects intercept sex educat region ;
run ;
options set=srclib "c:\srclib" sasautos=('!srclib' sasautos) mautosource;

LIBNAME d 'p:\pberg\Statistics.com Missing Data Class 2012';
* Read in data set;
data app4;
  set d.ncsr2_1;
  if mde=0 then mde_r=2; else if mde=1 then mde_r=1; else mde_r=.;
run;
proc freq;
tables mde_r;
run;

* use IVEware %impute to multiply impute missing data;
%impute (name=app4, setup=new, dir=.);
datain app4;
dataout app4_imp;
default continuous;
categorical sex region racecat educat mde_r str_secu;
transfer sampleid mde_imp educat_imp str secu mde ;
multiples 5; * create m=5 imputed data sets, this is the default;
seed 876;
run;

* use %putdata to produce 5 separate data sets for correct MI estimation;
%putdata(name=app4,dir=., mult=1, dataout=d.imp1);
%putdata(name=app4,dir=., mult=2, dataout=d.imp2);
%putdata(name=app4,dir=., mult=3, dataout=d.imp3);
%putdata(name=app4,dir=., mult=4, dataout=d.imp4);
%putdata(name=app4,dir=., mult=5, dataout=d.imp5);

%regress (name=app4_2, setup=new, dir=.)
datain  d.imp1 d.imp2 d.imp3 d.imp4 d.imp5;
stratum str;
cluster secu;
weight  finalp2wt;
categorical sex educat region;
predictor sex educat region;
dependent mde_r;
link logistic;
run;
# R Code

# R code for SI 616 2014 Presentation (July 21, 2014)
# Berglund

# data management
# load packages using library command
library(foreign)
library(mi)
library(mice)
# read Stata format data set into R
a <- read.dta("P:/pberg/statistics.com Missing Data Class 2012/ncsr2_v12.dta")
summary(a)

a$sex <- factor(a$sex)
a$educat <- factor(a$educat)
a$region <- factor(a$region)
a$str_secu <- factor(a$str_secu)
t <- table(a$mde, a$sex)

# obtain information about missing data
inf <- mi.info(a)
# print info about missing data
inf

# use mice to impute and pool
library(mice)
imp <- mice(a,n.imp=5,seed=1934)
summary(imp)

# convert mids to data useable for work in mitools
library(mitools)
mydata <- imputationList(lapply(1:5, complete, x=imp))
summary(mydata)

# set survey design
library(survey)
des <- svydesign(id=~secu, strat=~str, weight=~finalp2wt, data=(mydata), nest=TRUE)
summary(des)

# run design based model with svyglm using 5 imputed data sets contained in des (from mydata)
fit2 <- with (des, svyglm (mde ~ sex + educat + region, family=quasibinomial))
summary(MIcombine(fit2))
References


Heeringa, S., “Imputation Module Notes from Analysis of Complex Sample Data,” Institute for Social Research, University of Michigan Summer Institute Training Program.


References


