Multiple Imputation

Summer Workshops
June 10, 2009
What is MI and Why do I have to use it?

• MI is a Monte Carlo technique.
  – Missing data are imputed with conditional random values
  – Each new dataset is analyses
  – Combining for the results

• Make your dataset as small as possible
What is MI and Why do I have to use it?

- Extreme missing data can decrease sample size, statistical power, and increase the possibility of bias.
- Data are expected to be missing at random.
  - The probability of missing data on any variable is not related to its particular value.
How do I do MI in SAS?

The MI technique in SAS assumes that the variables are multivariate normal. If the missing are small it will be ok. Also, you can use the transform command.
How do I do MI in SAS?

```sas
proc mi data=mi seed=24 out=outmi ;
    var wabused habused;
run;

proc reg data=outmi outest=outreg covout nocprint;
    model kids= wabused habused;
    by _Imputation_;
run;

proc mianalyze data=outreg;
   modeleffect Intercept wabused habused;
run;
```
This tells us if our data are monotone or arbitrary in missing pattern.
Proc Mianalyze

• This is needed to produce univariate and multivariate results for the variables.
• The Proc MI procedure will create a variable called _imputation
  – Use this as a by variable
This output tells what is going on with the variance when we have the new dataset.

This gives us the standard error and parameter estimate for each variable in our model.
What you can use with Proc Mi

- Proc Reg
- Proc Genmod
- Proc Logit
- Proc Mixed
- Proc GLM
SAS IVWare

- [http://www.isr.umich.edu/src/smp/ive/](http://www.isr.umich.edu/src/smp/ive/)
- Perform a variety of descriptive and model-based analyses accounting for such complex design features as clustering, stratification, and weighting.
- Perform multiple imputation analyses for both descriptive and model-based survey statistics.
SAS IVWare

- Currently the following SAS PROCS can be called: CALIS, CATMOD, GENMOD, LIFEREG, MIXED, NLIN, PHREG, and PROBIT
- Variables can be: continuous, binary, categorical, counts, or mixed
How do I do MI in STATA?

• First make sure you have the ice program
• Findit ice
• Findit mim
How do I do MI in STATA?
set more off

ice kids wabused habused, /*
*/saving (R:\CFDR\CFDR\HEIDI\workshop_imputed.dta, replace) m(5)
genmiss (m_)/*
*/ seed(123)

use R:\CFDR\CFDR\HEIDI\workshop_imputed.dta, clear

tab _mj

mim: regress kids wabused habused
Single-user Stata for Windows perpetual license:
  Serial number: 199048108
  Licensed to: CFDR computer
  CFDR computer

Notes:
1. (/m# option or -set memory-) 1.00 MB allocated to data

.use "R:\CFDR\CFDR\HEIDI\CM\ice.dta", clear
.do "R:\CFDR\CFDR\HEIDI\CM\workshop.txt"

.set more off

.ice kids wabused habused, /*
> */saving (R:\CFDR\CFDR\HEIDI\workshop_imputed.dta, replace) m(5) genmiss (m_)/*
> */ seed(123)

<table>
<thead>
<tr>
<th>#missing values</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>473</td>
<td>91.67</td>
<td>91.67</td>
</tr>
<tr>
<td>1</td>
<td>42</td>
<td>8.14</td>
<td>99.81</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.19</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>516</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>

Variable        | Command | Prediction equation
-----------------|---------|----------------------
kids wabused     | regress | [No missing data in estimation sample] kids habused
habused          | regress | kids wabused
Imputing ..........1...........2...........3...........4...........5
file R:\CFDR\CFDR\HEIDI\workshop_imputed.dta saved
.
use R:\CFDR\CFDR\HEIDI\workshop_imputed.dta, clear
.
* tab _mj

<table>
<thead>
<tr>
<th>imputation number</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>516</td>
<td>16.67%</td>
<td>16.67%</td>
</tr>
<tr>
<td>1</td>
<td>516</td>
<td>16.67%</td>
<td>33.33%</td>
</tr>
<tr>
<td>2</td>
<td>516</td>
<td>16.67%</td>
<td>50.00%</td>
</tr>
<tr>
<td>3</td>
<td>516</td>
<td>16.67%</td>
<td>66.67%</td>
</tr>
<tr>
<td>4</td>
<td>516</td>
<td>16.67%</td>
<td>83.33%</td>
</tr>
</tbody>
</table>
| 5                 | 516   | 16.67%  | 100.00%
| Total             | 3,096 | 100.00% |        |

* mim: regress kids wabused habused

Multiple-imputation estimates (regress) Imputations = 5
Linear regression Minimum obs = 516
Minimum dof = 178.2

|         | Coef.  | Std. Err. | t     | P>|t|  | [95% Conf. Int.] | MI.df |
|---------|--------|-----------|-------|------|------------------|------|
| kids    |        |           |       |      |                  |      |
| wabused | .023559| .01402    | 1.68  | 0.095| -.004107 .051225|      |
| habused | .039167| .016678   | 2.35  | 0.020| .006317 .072016 |    |
| _cons   | .4184  | .057147   | 7.32  | 0.000| .306107 .530692 | 475.2|

end of do-file
A more complex example - Add Health

ice happy rsat rschool hs twodayear grad notenrolled work parttime
fulltime married lwp cohab consequences risks behavior
depressed fitin notfuture rnocrowd maturity female hadsex
responsibilities bio income momed rrace black hisp otherrace
mlhs mhs msomec money, /*

*/saving (T:\Users\hlyons\min_impute.dta, replace) m(3) genmiss
(m_)/*

svyset [pweight=gswgt3_2], strata(region)psu(psuscid)
*/cmd(happy rsat consequence behavior risks fitin notfuture rnocrowd maturity responsibilities rschool momed: ologit, work : mlogit, married lwp cohab bio female hadsex : logit)/*

*/passive
(hs:rschool==1\twoyear:rschool==2\grad:rschool==4\notenrolled:rschool==5\parttime:work==2\fulltime:work==3\mlhs:momed==1\mhs:momed==2\msomec:momed==3)/*

*/substitute (rschool: hs twyear grad notenrolled, work: parttime fulltime, momed: mlhs mhs msomec)

*/ seed(123)

use T:\Users\hlyons\min_impute.dta, clear

tab _mj
mim:svy,subpop(marker):ologit happy twoyear grad notenrolled parttime fulltime married lwp cohab consequences behavior risks fitin notfuture rnocrowd bio income mlhs mhs msomec black hisp otherrace age hadsex female money rsat,or
What Svy commands Mim can do?

- Svy: regress
- Svy: mean
- Svy: proportion
- Svy: ratio
- Svy: logistic
- Svy: ologit
- Svy: mlogit
- Svy: probit
- Svy: oprobit
- Svy: poisson
SPSS

- Now, using SPSS Missing Values 17.0, you can impute missing values for categorical or continuous variables by multiple imputation.
Questions?

Next workshop:

*Introduction to Structural Equation Modeling*

Wednesday, June 17, 12:00-1:00
Room 314