Multiple Imputation

Summer Workshops

June 10, 2009

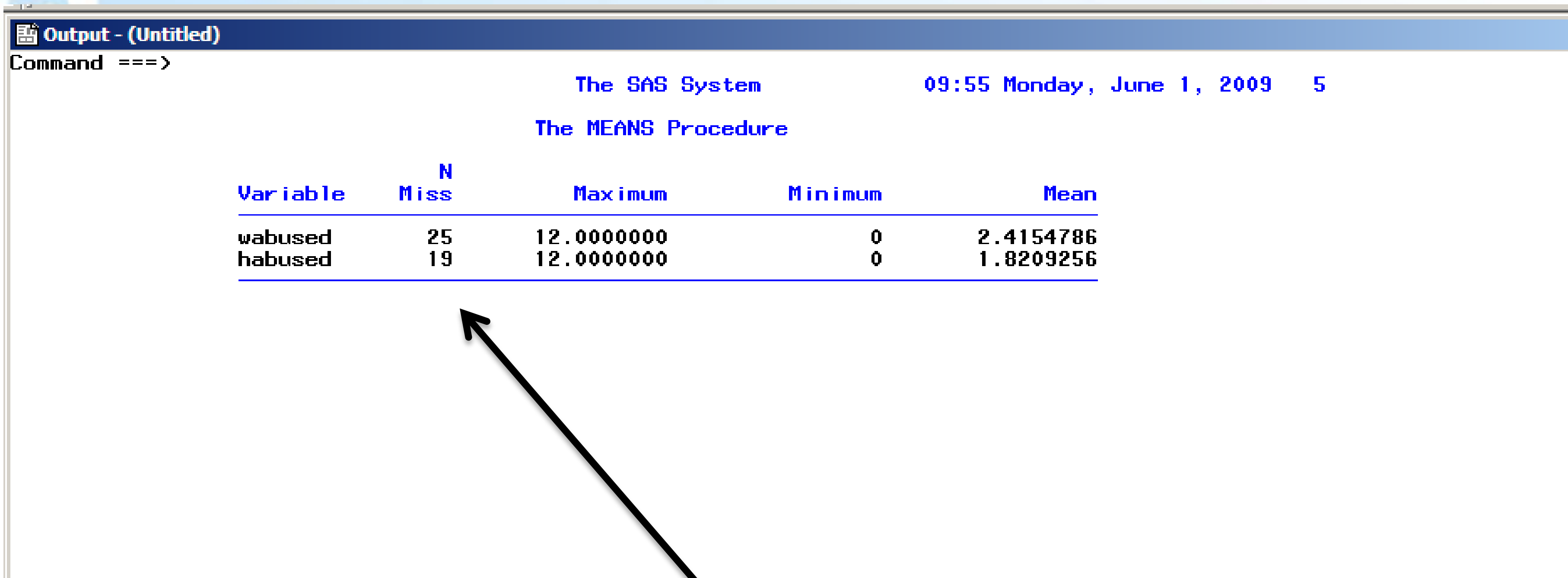
What is MI and Why do I have to use it?

- 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?



Output - (Untitled)

Command ===>

The SAS System 09:55 Monday, June 1, 2009 5

The MEANS Procedure

Variable	N Miss	Maximum	Minimum	Mean
wabused	25	12.0000000	0	2.4154786
habused	19	12.0000000	0	1.8209256

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?

Command ===>

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

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

proc mianalyze data=outreg;
modeleffect Intercept wabused habused;
run;
```

This tells use if our data are monotone or arbitrary in missing pattern

```
Command ===>
                    Parameter Estimates from Imputed Data Sets                    27
                                                       09:55 Monday, June 1, 2009

                                 The MI Procedure

                                Model Information

              Data Set                              WORK.MI
              Method                                MCMC
              Multiple Imputation Chain             Single Chain
              Initial Estimates for MCMC            EM Posterior Mode
              Start                                 Starting Value
              Prior                                 Jeffreys
              Number of Imputations                 5
              Number of Burn-in Iterations          200
              Number of Iterations                  100
              Seed for random number generator      24

                              Missing Data Patterns

                                                        ---------Group Means--------
Group    wabused    habused      Freq    Percent        wabused          habused

    1    X          X             473      91.67       2.410148         1.803383
    2    X          .              18       3.49       2.555556                .
    3    .          X              24       4.65              .         2.166667
    4    O          O               1       0.19              .                .


                        EM (Posterior Mode) Estimates

                 _TYPE_    _NAME_          wabused         habused

                 MEAN                     2.419339        1.821697
                 COV       wabused       10.057401        1.573766
                 COV       habused        1.573766        6.873918


                  Multiple Imputation Variance Information

                      -----------------Variance-----------------
           Variable       Between        Within         Total       DF

           wabused       0.001011      0.019614      0.020827   342.29

                   Multiple Imputation Variance Information

                                Relative      Fraction
                                Increase       Missing       Relative
                Variable     in Variance   Information     Efficiency

                wabused         0.061830      0.059822       0.988177
```

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

Command ===>

Parameter Estimates from Imputed Data Sets 29
09:55 Monday, June 1, 2009

This output tells what is going on with the variance when we have the new dataset

The MIANALYZE Procedure

Model Information

Data Set	WORK.OUTREG
Number of Imputations	5

This gives us the standard error and parameter estimate for each variable in our model.

Multiple Imputation Variance Information

Parameter	Variance: Between	Variance: Within	Variance: Total	DF
Intercept	0.000159	0.003224	0.003415	1286.2
wabused	0.000010523	0.000175	0.000188	886
habused	0.000014963	0.000258	0.000275	941.65

Multiple Imputation Variance Information

Parameter	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
Intercept	0.059061	0.057232	0.988683
wabused	0.072031	0.069290	0.986331
habused	0.069719	0.067155	0.986747

Multiple Imputation Parameter Estimates

Parameter	Estimate	Std Error	95% Confidence Limits		DF
Intercept	0.425682	0.058437	0.31104	0.540324	1286.2
wabused	0.023758	0.013709	-0.00315	0.050664	886
habused	0.035350	0.016598	0.00278	0.067923	941.65

Multiple Imputation Parameter Estimates

Parameter	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
Intercept	0.406227	0.440495	0	7.28	<.0001
wabused	0.019762	0.027856	0	1.73	0.0834
habused	0.031986	0.039810	0	2.13	0.0334

What you can use with Proc Mi

- Proc Reg
- Proc Genmod
- Proc Logit
- Proc Mixed
- Proc GLM

SAS IVEware

- 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 IVEware

- 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?

```
. summarize

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    respid_w |       516    11329.81    826.6325      10008      12700
     wabused |       491    2.415479    3.183261          0         12
     habused |       497    1.820926    2.632406          0         12
        kids |       516    .5465116    .9429181          0          6

.
```

```
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
```

Results

```
                                   College Station, Texas 77845 USA
                                   800-STATA-PC         http://www.stata.com
                                   979-696-4600         stata@stata.com
                                   979-696-4601 (fax)

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)

   #missing |
     values |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        473       91.67       91.67
          1 |         42        8.14       99.81
          2 |          1        0.19      100.00
------------+-----------------------------------
      Total |        516      100.00

   Variable | Command | Prediction equation
------------+---------+-----------------------------------------------
       kids |         | [No missing data in estimation sample]
    wabused | regress | 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

 imputation |
     number |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        516       16.67       16.67
          1 |        516       16.67       33.33
          2 |        516       16.67       50.00
          3 |        516       16.67       66.67
          4 |        516       16.67       83.33
          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

------------------------------------------------------------------------------
        kids |     Coef.  Std. Err.     t    P>|t|    [95% Conf. Int.]   MI.df
-------------+----------------------------------------------------------------
     wabused |   .023559     .01402   1.68   0.095  -.004107  .051225   178.2
     habused |   .039167    .016678   2.35   0.020   .006317  .072016   246.5
       _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 twoyear 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 twoyear 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