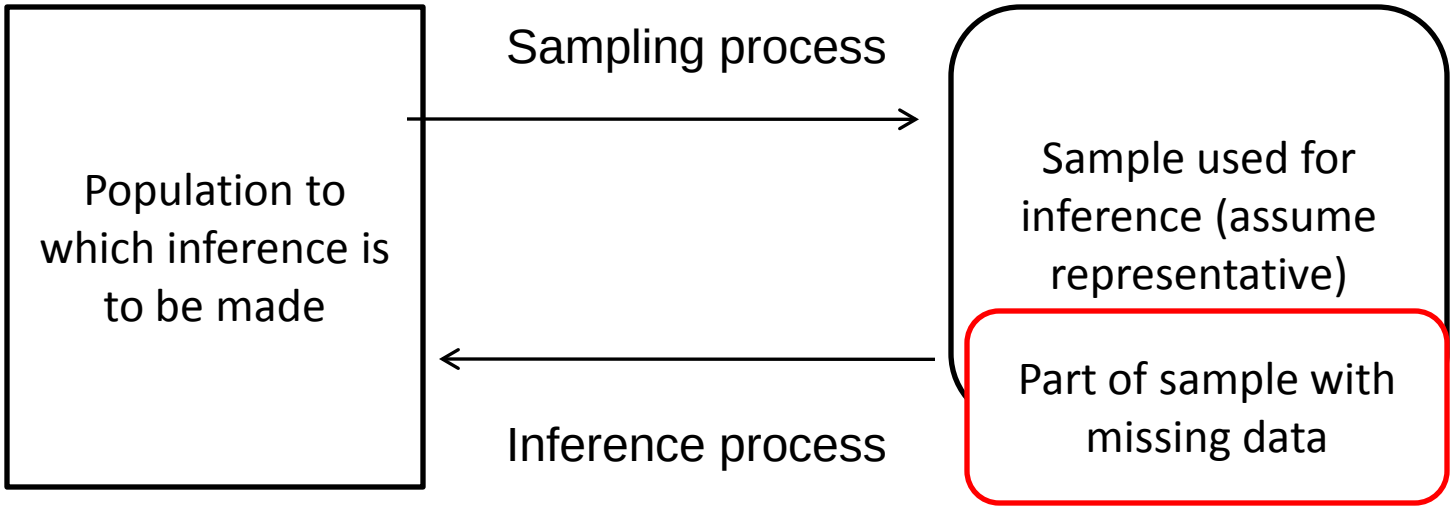
Missing Data: Part 2
Implementing Multiple Imputation in STATA and SPSS

Carol B. Thompson
Johns Hopkins Biostatistics Center
SON Brown Bag – 4/24/13

Overview

- Reminder – Steps in Multiple Imputation
- Implementation in STATA
- Implementation in SPSS

Statistical Analysis & Missingness



Is sample with missing data still representative enough to make appropriate inferences to population of interest?????

Missingness Mechanisms

- Process by which observations become missing
- Mechanism types
  - Missing Completely at Random (MCAR)
  - Missing at Random (MAR)
  - Missing Not at Random (MNAR)
- Using Multiple Imputation
  - Mostly for MAR
  - Likely not for MNAR

Multiple Imputation (MI)

- When used correctly, produces estimates:
  - Approximately unbiased, better as sample size increases
  - Asymptotically normal when data are MAR
    - Can construct CIs and p-values

MI cont'd

- Advantages
  - Can be used with virtually any kind of data,
  - any kind of model,
  - and with unmodified conventional software
- Disadvantages
  - Can be cumbersome to implement,
  - Is easy to do wrong,
  - Produces different estimates every time it is used (hopefully small differences)

MI Process

- Repeat the random imputation process more than once (5 times is generally enough)
- Each imputation process represents random sample from distribution of plausible values for missing values
- Important for imputation processes to be independent – large number of iterations between each saved data set
- Analyze data set from each imputation process as if no missing data

MI Process – Pooling Estimates

- Calculate mean of estimates
- Calculate mean of squared std errors
- Calculate variance of estimates
- Calculate square root of mean of variances plus variance of estimates
- Can be used with any parameter

MI Example (Howell- Part 2)

Regression coefficients from five imputed data sets

Data set	Estimated parameter	b_0	b_1	b_2	b_3	b_4	b_5
1	Coefficient	-11.535	-2.780	1.029	-.031	-0.359	0.572
	Variance	43.204	3.323	0.013	0.013	0.013	0.012
2	Coefficient	-11.501	-4.149	1.040	-0.093	-0.583	0.876
	Variance	40.488	2.680	0.010	0.009	0.009	0.007
3	Coefficient	-10.141	-5.038	0.766	0.123	-0.252	0.625
	Variance.	42.055	3.301	0.010	0.010	0.010	0.009
4	Coefficient	-11.533	-6.920	0.870	0.084	-0.458	0.815
	Variance	28.751	1.796	0.081	0.007	0.007	0.007
5	Coefficient	-14.586	-1.115	0.718	0.050	-0.373	0.814
	Variance	32.856	2.362	0.009	0.009	0.009	0.008
	Mean b_i	-11.859	-4.000	0.885	0.027	-0.405	0.740
	Mean Var.($\bar{W}$)	37.471	2.692	0.025	0.010	0.010	0.009
	Var. of b_i (B)	2.682	4.859	0.022	0.008	0.015	0.018
	T	40.69	8.523	0.051	0.020	0.028	0.031
	$\sqrt{T}$	6.379	2.919	0.226	0.141	0.167	0.176
	t	-1.859	-1.370	3.916*	0.191	2.425*	4.204*

* $p < .05$ "Var." refers to the squared standard error of the coefficient.

Additional Rules of Thumb (Allison)

- Dependent Variable (DV) should always be included in imputation regression analysis
- Impute missing values on DV if:
  - There are auxiliary variables strongly correlated with DV.
- Don't impute DV if:
  - No missing predictor data or auxiliary variables
  - No auxiliary variables and missing predictor data

Preparation

- Explore missing data patterns
- Determine missingness mechanism and appropriateness for MI
- Assign missing "codes" in data set to missing designation
  - ., .a through .z in STATA
  - Missing Values command in SPSS
- Determine variables to be included in MI process – not just those included in model

MI in STATA – Data set

- Data set: use http://www.stata-press.com/data/r11/mheart5
- Fictional heart attack data; bmi and age missing
  - 12 cases with missing age
  - 28 cases with missing bmi
- Variables:
  - attack (binary, dependent variable)
  - smokes (binary)
  - age (continuous)
  - bmi (continuous)
  - female (binary)
  - hsgrad (binary)

MI in STATA – Set up/review

- Declare data to be “mi”
  - set mi mlong
  - mlong is most memory efficient
- Explore missing patterns
  - mi misstable sum – (other options)
- Register variables
  - mi register *type varlist*
  - imputed – required
  - passive - variable that is function of imputed variable(s)
  - regular – neither imputed nor passive
- Confirm mi data set up
  - mi describe

MI in STATA – Imputation Step

- Set seed for reproducibility or in mi impute command
  - set seed 29390
- Create imputed data sets
  - mi impute method …, options
  - Set up and options differ by method
  - mi impute mvn age bmi = attack smokes hsgrad female, rseed(29390) add(10)
  - Creates 10 imputation data sets with seed 29390 using multivariate normal regression
  - The more missing data, the more imputations needed.

MI in STATA – Imputation Step

- 11 data sets
  - Original data set (numbered as 0) with missing data
  - Imputed data sets (numbered as 1-10)
- Review imputed data sets
  - Show summary statistics for imputed variables
  - mi xeq 0 1 3 6 10, summarize age bmi

MI in STATA – Estimation Step

- Run estimation model
  - mi estimate, options: estimation command
  - always provides estimates as coefficients
  - mi estimate: logistic attack smokes age bmi hsgrad female
- Get estimate in terms of odds ratios
  - mi estimate, or

MI in STATA – Compare estimates

	complete data only			M=5			M=10			M=20		
	OR	se	p	OR	se	p	OR	se	p	OR	se	p
smokes	4.54	1.84	0	3.46	1.28	0.001	3.43	1.25	0.001	3.33	1.21	0.001
age	1.031	0.018	0.088	1.034	0.017	0.052	1.033	0.017	0.042	1.031	0.017	0.064
bmi	1.1	0.055	0.047	1.12	0.06	0.035	1.12	0.056	0.024	1.11	0.059	0.061
hsgrad	1.38	0.616	0.469	1.21	0.495	0.647	1.2	0.488	0.66	1.17	0.476	0.696
female	1.32	0.615	0.549	0.92	0.389	0.845	0.91	0.382	0.823	0.917	0.38	0.835

MI in SPSS

- Data Set – CancerHead_DCHowell_SPSS.xls
  - Child behavior problems when parent has cancer
  - All variables have missing data (value = -9)
- Variables:
  - SexP, SexChild (binary)
  - DeptP, DeptS (continuous)
  - AnxtP, AnxtS (continuous)
  - GSItP, GSItS (continuous)
  - Totbpt (continuous, dependent variable)

MI in SPSS – Set up/Review

- Assign missing values for all variables
  - MISSING VALUES SexP DeptP AnxtP GSItP DeptS AnxtS GSItS SexChild Totbpt (-9).
- Missing Value Analysis
  - Summary statistics listwise (non-missing cases) and all cases
  - Missing patterns by variables
  - Analyze → Missing Values Analysis
  - MVA …
- Analysis of Missing Value Patterns
  - Analyze → Multiple Imputation → Analyze Patterns
  - Multiple Imputation ….

MI in SPSS – Imputation Step

- Set seed for imputation (separate from imputation command)
  - Set SEED 29390.
- Multiple Imputations
  - Analyze → Multiple Imputation → Impute Missing Values
  - MULTIPLE IMPUTATION SexP DeptP AnxtP GSItP DeptS AnxtS GSItS SexChild Totbpt /IMPUTE METHOD=AUTO NIMPUTATIONS=5 MAXPCTMISSING=NONE /MISSINGSUMMARIES NONE /IMPUTATIONSUMMARIES MODELS DESCRIPTIVES /OUTFILE IMPUTATIONS=SPSSImputations .

  - Set up method, # imputations, resulting summaries, and data set in SPSS session to contain imputations (here – SPSSImputations; can also save to an SPSS file)

MI in SPSS – Imputation Step

- SPSSImputations includes variable Imputation_
  - Window → SPSSImputations Data Set
  - 0 represent original data set
  - 1-5 represents imputed data sets
    - Imputed values are highlighted
- Output shows summary statistics for original data set, and imputed cases, and all data with imputed values by imputation

MI in SPSS – Estimation Step

- Select analysis from Analyze menu
  - Can only impute if icon shows
  - Specify imputed data sets to be used in analysis
- DATASET ACTIVATE SPSSImputations.
- REGRESSION /DESCRIPTIVES MEAN STDDEV CORR SIG N /SELECT=Imputation_ GE 1 …. /DEPENDENT Totbpt /METHOD=ENTER SexP DeptP AnxtP GSItP DeptS AnxtS GSItS SexChild.
- Shows summary statistics/analysis for original data, each imputation, and pooled estimates

References - SPSS

- http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/MissingDataSPSS.html - Howell, DC. "Multiple Imputation Using SPSS"
- http://www.gmw.rug.nl/~huisman/md/MD5_Imputation_2011.pdf - Huisman, M. "Missing Data – Session 5 – Imputation (SPSS)"
- http://www.appliedmissingdata.com/spss_multiple_imputation.pdf - Enders, CK. Excerpt from Applied Missing Data Analysis (mostly for Mplus, some for SPSS)
- http://www.unt.edu/rss/class/Jon/SPSS_SC/Module6/SPSS_M6_2.htm - University of North Texas" University IT Part of SPSS workshop.
- ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/20.0/en/client/Manuals/IBM_SPSS_Missing_Values.pdf - SPSS Missing Values Manual for V20

References - STATA

- http://www.ats.ucla.edu/stat/stata/seminars/missing_data/mi_in_stata_pt1.htm - UCLA Statistical Computing Seminars – part 1 (using **mi**)
- http://www.ats.ucla.edu/stat/stata/seminars/missing_data/mi_in_stata_pt2.htm - UCLA Statistical Computing Seminars – part 2 (using chained equations with **ice**)
- http://biostat.mc.vanderbilt.edu/wiki/pub/Main/QingxiaChen/MI_stata.pdf - Marchenko Y. 2009 UK Stata Users Group Meeting (v 11)
- http://www.ssc.wisc.edu/sscc/pubs/stata_mi_intro.htm - Beginning of series of MI topics
- STATA manual for Multiple-Imputation – available from Help menu → PDF documentaiton

```
**MI_SPSS_20130424.sps in C:\CBThompson\SON\Brown_Bag\Missing_Data_p2_20130424.

**Test MI from DC Howell - Multiple Imputation Using SPSS.

 **Copy .CancerHead-9.dat to Excel spreadsheet and relabel last few columns to be
consistent with documentation.
**Data set  contains bvariables related to child behavior problems among kids who
have a parent with cancer.

GET DATA
  /TYPE=XLS

/FILE='C:\CBThompson\SON\Brown_Bag\Missing_Data_p2_20130424\CancerHead_DCHowell_SPSS.
xls'
  /SHEET=name 'Sheet1'
  /CELLRANGE=full
  /READNAMES=on
  /ASSUMEDSTRWIDTH=32767.
EXECUTE.
DATASET NAME DataSet1 WINDOW=FRONT.

** add descriptors to variables.
VARIABLE LABELS SexP "Sex Parent" /
DeptP "Parent's Depression T score"/
AnxtP "Parent's Anxiety T score"/
GSItP "Parent's  Global Symptom Index T score"/
DeptS "Spouse's Depression T score"/
AnxtS "Spouse's Anxiety T score" /
GSItS "Spouse's  Global Symptom Index T score"/
SexChild "Sex Child"/
Totbpt " Total Behavior Problem T score for child".

**Assign missing values to variables.
MISSING VALUES SexP DeptP AnxtP GSItP DeptS AnxtS GSItS SexChild Totbpt (-9).

**Missing Values Analysis.
MVA VARIABLES=DeptP AnxtP GSItP DeptS AnxtS GSItS Totbpt SexP SexChild
  /MAXCAT=25
  /CATEGORICAL=SexP SexChild
  /MISMATCH PERCENT=5
  /TPATTERN PERCENT=1 DESCRIBE=DeptP AnxtP GSItP DeptS AnxtS GSItS Totbpt SexP
SexChild
  /LISTWISE.

*Analyze Patterns of Missing Values.
MULTIPLE IMPUTATION  SexP DeptP AnxtP GSItP DeptS AnxtS GSItS SexChild Totbpt
   /IMPUTE METHOD=NONE
   /MISSINGSUMMARIES  OVERALL VARIABLES (MAXVARS=25 MINPCTMISSING=10) PATTERNS.

**Set Seed.
Set SEED 29390.

**Impute Missing Data Values - 5 iterations.
DATASET DECLARE SPSSImputations.
MULTIPLE IMPUTATION SexP DeptP AnxtP GSItP DeptS AnxtS GSItS SexChild Totbpt
  /IMPUTE METHOD=AUTO NIMPUTATIONS=5 MAXPCTMISSING=NONE
  /MISSINGSUMMARIES NONE
  /IMPUTATIONSUMMARIES MODELS DESCRIPTIVES
```

```
  /OUTFILE IMPUTATIONS=SPSSImputations .

***Regression Analysis on each imputation and Pooled across imputation estimates.
DATASET ACTIVATE SPSSImputations.
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /SELECT=Imputation_ GE 1
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA CHANGE
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT Totbpt
  /METHOD=ENTER SexP DeptP AnxtP GSItP DeptS AnxtS GSItS SexChild.
```

```
capture log close
log using
"C:\CBThompson\SON\Brown_Bag\Missing_Data_p2_20130424\MI_STATA_20130424.log", replace
**MI_STATA_20130424.do in C:\CBThompson\SON\Brown_Bag\Missing_Data_p2_20130424

****Based on STATA version 12*******

clear
set more off

*Sec A - "A really simple example" from STATA MI intro
**Fictional heart attack data set
use http://www.stata-press.com/data/r11/mheart5, clear
**A.1 data set
describe
misstable summarize

**Sec A.2 - Basic analysis -- excludes missing data
logit attack smokes age bmi hsgrad female
logistic attack smokes age bmi hsgrad female

**Sec A.3 - Set up data set for MI
preserve
mi set mlong
mi register imputed age bmi
mi misstable summarize

**Sec A.4 - set seed for reproduciliblity or include in mi impute command
***set seed 29390

**Sec A.5 - run imputation model with 10 imputations and check resulting imputed data
**impute with nultivariate normal regression
mi impute mvn age bmi = attack smokes hsgrad female, add(10) rseed(29390)
mi describe
mi xeq 0 1 3 6 10: summarize age bmi

**Sec A.6 - run analysis model based on 10 sets of imputed values
mi estimate: logistic attack smokes age bmi hsgrad female
mi estimate, or
restore

**Sec A.7 - run imputation model with 5 imputations and then analysis model
**set seed 29390
preserve
mi set mlong
mi register imputed age bmi
mi impute mvn age bmi = attack smokes hsgrad female, add(5) rseed(29390)
mi estimate: logistic attack smokes age bmi hsgrad female
mi estimate, or
restore

**Sec A.8 - run imputation model with 20 imputations and then analysis model
**set seed 29390
preserve
mi set mlong
mi register imputed age bmi
mi impute mvn age bmi = attack smokes hsgrad female, add(20) rseed(29390)
mi estimate: logistic attack smokes age bmi hsgrad female
```

```
mi estimate, or
restore

**Sec B. continuous outcome data set
use http://www.stata-press.com/data/r11/mheart0, clear
generate lnbmi = ln(bmi)
mi set mlong
mi register imputed lnbmi
**impute with linear regression -- relies on normality of model
mi impute regress lnbmi age attack smokes age hsgrad female, add(20) rseed(2232)
**bmi will be function of original bmi - thus needs to be registered as passive
mi register passive bmi
quietly mi passive: replace bmi = exp(lnbmi)
mi estimate, dots: logit attack smokes age bmi hsgrad female
mi estimate, or


log close
```