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

Presented by Pat Berglund
Summer Institute Program
Survmeth 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

Stata 13.1-Examination of Missing Data

- 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)

Missing Data Pattern

- 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  2
  ------------+-------------
       93%    |  1  1
              |
        4     |  1  0
        3     |  0  1
  ------------+-------------
      100%    |

  Variables are  (1) mde  (2) educat
. misstable tree

  Nested pattern of missing values
      educat        mde
  --------------------
           4%         0%
                      4
          96          3
                     93
  --------------------
 (percent missing listed first)
```

Examine Missing Data, continued

```
. misstable summarize
                                                                  Obs<.
                                                  +--------------------------------
                |                                 | Unique
    Variable    |     Obs=.     Obs>.     Obs<.   | values        Min          Max
----------------+---------------------------------+--------------------------------
       educat   |       235                 5,457 |      4          1            4
          mde   |       165                 5,527 |      2          0            1
-------------------------------------------------------------------------------
```

- 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

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    sampleid |         0
         sex |      5692    .4184821    .4933534          0          1
      region |      5692    2.575018    1.021898          1          4
         age |      5692    43.37807     16.5794         17         98
         str |      5692    26.37878    11.15588          1         42
-------------+--------------------------------------------------------
        secu |      5692    1.505271    .5000161          1          2
   finalp2wt |      5692           1    .9582254   .1144058   10.10207
    racecat_ |      5692    3.423226    1.025544          1          4
      educat |      5457    2.651457    1.012665          1          4
         mde |      5527    .3155419    .4647734          0          1
-------------+--------------------------------------------------------
    str_secu |      5692     265.293    111.5497         11        422
```

- The variables sex, region, str, secu, racecat_, educat, str_secu and mde are categorial 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
Imputed: m=1 through m=5                        updated =        0

Initialization: monotone                     Iterations =       50
                                                burn-in =       10

              mde: logistic regression
           educat: ordered logistic regression

------------------------------------------------------------------
                  |               Observations per m
                  |-----------------------------------------------
         Variable |   Complete   Incomplete   Imputed |     Total
------------------+-----------------------------------+----------
              mde |       5527          165       165 |      5692
           educat |       5457          235       235 |      5692
------------------------------------------------------------------
(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

------------------------------------------------------------------------------
             |        Imputation variance                               Relative
             |    Within   Between     Total       RVI       FMI     efficiency
-------------+----------------------------------------------------------------
mde          |
           0 |   .000039   7.7e-07    .00004   .023773   .026471       .994734
           1 |   .000039   7.7e-07    .00004   .023773   .026471       .994734
-------------+----------------------------------------------------------------
educat       |
           1 |    .00007   5.0e-06   .000076   .084999   .086233       .983046
           2 |   .000139   4.0e-06   .000144    .03463   .037412       .992573
           3 |   .000064   4.7e-06    .00007    .08677   .087891       .982725
           4 |   .000113   3.6e-06   .000117   .038482   .041259       .991816
------------------------------------------------------------------------------
```

Use of mi estimate: vartable: svy: proportion command

```
Multiple-imputation estimates      Imputations     =          5
Survey: Proportion estimation      Number of obs   =       5692

Number of strata  =        42      Population size = 5692.0005
Number of PSUs    =        84
                                   Average RVI     =    0.0788
                                   Largest FMI     =    0.0879
                                   Complete DF     =        42
DF adjustment:   Small sample      DF:     min     =     34.88
                                           avg     =     37.40
Within VCE type:   Linearized              max     =     39.00

------------------------------------------------------------------
             | Proportion   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------
mde          |
           0 |   .8083499   .0062936      .7956198      .82108
           1 |   .1916501   .0062936        .17892    .2043802
-------------+----------------------------------------------------
educat       |
           1 |   .1676641   .0087298      .1499417    .1853865
           2 |   .3245322   .0119858       .300276    .3487885
           3 |   .2755689   .0083708       .258573    .2925647
           4 |   .2322348   .0108185      .2103365    .2541331
------------------------------------------------------------------
```

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                   Imputations        =          5
Survey: Logistic regression                     Number of obs      =       5692

Number of strata  =        42                   Population size    = 5692.0005
Number of PSUs    =        84
                                                Average RVI        =     0.0464
                                                Largest FMI        =     0.0933
                                                Complete DF        =         42
DF adjustment:   Small sample                   DF:     min        =      34.46
                                                        avg        =      38.43
                                                        max        =      39.79
Model F test:       Equal FMI                   F(   7,    39.7)   =      16.02
Within VCE type:   Linearized                   Prob > F           =     0.0000

------------------------------------------------------------------------------
         mde |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       1.sex |  -.5111926   .0638813    -8.00   0.000    -.6403719   -.3820132
             |
      region |
          2  |   .0065904   .1408085     0.05   0.963    -.2783006    .2914814
          3  |  -.1669361   .1330025    -1.26   0.217    -.4358414    .1019692
          4  |    .039644   .1319837     0.30   0.765    -.2271491    .3064372
             |
      educat |
          2  |    .094891   .1074464     0.88   0.383    -.1228383    .3126204
          3  |   .2324122   .1124072     2.07   0.046      .004085    .4607393
          4  |   .1770819   .1114483     1.59   0.120    -.0482573    .4024212
             |
       _cons |  -1.311143   .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

Examine Missing Data Pattern

```
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	educat_ 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	Group Means							
	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	3.361702	.

Missing Data Patterns

Group	Group Means			
	mde	mde_imp	educat_imp	str_secu
1	0.316515	0	0	265.268519
2	.	1.000000	0	271.103030
3	0.293617	0	1.000000	261.765957

Impute Missing Data Using FCS Method

- 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

Model Information	
Data Set	D.NCSR2_1
Method	FCS
Number of Imputations	5
Number of Burn-in Iterations	20
Seed for random number generator	876

FCS Model Specification	
Method	Imputed Variables
Regression	age finalp2wt
Logistic Regression	mde educat
Discriminant Function	sex region racecat_ str_secu

Logistic Models for FCS Method

Imputed Variable	Effect	sex	region	racecat_	str_secu	educat	Imputation 1	2	3	4	5
mde	Intercept	.	.	.	.	.	1.536822	1.407798	1.446521	1.648773	1.424868
mde	sex	0	.	.	.	.	-0.177406	-0.087597	-0.119781	-0.208306	-0.178475
mde	region	.	1.000000	.	.	.	1.575420	0.116013	-1.045273	-0.031539	-1.212921
mde	region	.	2.000000	.	.	.	-0.743625	0.155627	0.668689	0.145863	0.950015
mde	region	.	3.000000	.	.	.	-0.827360	-0.320575	0.072085	-0.478683	0.220971
mde	age	.	.	.	.	.	0.055691	-0.010080	0.013294	0.079427	-0.008374
mde	racecat_	.	.	1.000000	.	.	-0.045572	0.167123	-0.016189	-0.029828	0.135773
mde	racecat_	.	.	2.000000	.	.	0.301858	0.227859	0.393062	0.346515	0.363653
mde	racecat_	.	.	3.000000	.	.	-0.072845	-0.108918	-0.032006	-0.067074	-0.243884
mde	str_secu	.	.	.	11.000000	.	0.022327	1.067933	2.438004	-0.017650	1.774737
mde	str_secu	.	.	.	12.000000	.	-1.674096	-1.863723	0.522617	-0.900591	0.545178
mde	str_secu	.	.	.	21.000000	.	-1.846103	-0.563765	1.123181	0.796031	1.530252
mde	str_secu	.	.	.	22.000000	.	-1.321143	0.738015	0.856889	0.998848	2.326307
mde	str_secu	.	.	.	31.000000	.	-0.546819	-0.401253	-0.906626	-0.348298	-0.729161

Crosstabulations of Major Depressive Episode by Imputed Flag and Imputation (Partial Output)

Frequency	Table 1 of mde by mde_imp		
	Controlling for _Imputation_=1		
mde(Major Depressive Episode)	mde_imp		
	0	1	Total
0	3783	98	3881
1	1744	67	1811
Total	5527	165	5692

Frequency	Table 2 of mde by mde_imp		
	Controlling for _Imputation_=2		
mde(Major Depressive Episode)	mde_imp		
	0	1	Total
0	3783	100	3883
1	1744	65	1809
Total	5527	165	5692

```
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

```
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

```
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 the default of treating variables as continuous.
- The MODELEFFECTS statement uses the same model specification from PROC SURVEYLOGISTIC (in the previous step), including the intercept

MIANALYZE Partial Output

- 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

The MIANALYZE Procedure

Model Information	
PARMS Data Set	WORK.OUTPARMS
Number of Imputations	5

Variance Information

Parameter	sex	educat	region	Variance			DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
				Between	Within	Total				
intercept				0.000137	0.020909	0.021074	65392	0.007883	0.007851	0.998432
sex	1			0.000025634	0.004869	0.004900	101490	0.006318	0.006298	0.998742
educat		2		0.000404	0.011551	0.012036	2462.1	0.042000	0.041085	0.991850
educat		3		0.011276	0.011015	0.024547	13.163	1.228441	0.606783	0.891777
educat		4		0.010145	0.012549	0.024723	16.496	0.970143	0.544492	0.901796
region			2	0.000252	0.017985	0.018287	14632	0.016812	0.016668	0.996677
region			3	0.000275	0.016343	0.016673	10200	0.020203	0.019995	0.996017
region			4	0.000329	0.016842	0.017236	7643.5	0.023412	0.023132	0.995395

Parameter Estimates

Parameter	sex	educat	region	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
intercept				-1.302637	0.145168	-1.58717	-1.01811	65392	-1.314556	-1.284028	0	-8.97	<.0001
sex	1			-0.504400	0.069998	-0.64160	-0.36720	101490	-0.510147	-0.496238	0	-7.21	<.0001
educat		2		0.095271	0.109708	-0.11986	0.31040	2462.1	0.071256	0.112619	0	0.87	0.3853
educat		3		0.309667	0.156675	-0.02838	0.64772	13.163	0.171845	0.396707	0	1.98	0.0694
educat		4		0.214843	0.157236	-0.11767	0.54736	16.496	0.126754	0.324696	0	1.37	0.1902
region			2	-0.022962	0.135231	-0.28803	0.24211	14632	-0.040051	-0.001512	0	-0.17	0.8652
region			3	-0.194508	0.129125	-0.44762	0.05860	10200	-0.210003	-0.167187	0	-1.51	0.1320
region			4	0.019060	0.131287	-0.23830	0.27642	7643.5	-0.007351	0.038419	0	0.15	0.8846

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)

IVEware Code to Impute Missing Data

```
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

IVEware Iterative Imputation Procedure, Tue Jul 15 13:39:24 2014 5

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

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
1	813	14.90	35	14.89	848	14.90
2	1641	30.07	75	31.91	1716	30.15
3	1638	30.02	74	31.49	1712	30.08
4	1365	25.01	51	21.70	1416	24.88
Total	5457	100.00	235	100.00	5692	100.00

Variable mde_r

	Observed		Imputed		Combined	
Code	Freq	Per	Freq	Per	Freq	Per
1	1744	31.55	48	29.09	1792	31.48
2	3783	68.45	117	70.91	3900	68.52
Total	5527	100.00	165	100.00	5692	100.00

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 ;
```

Logistic Regression Output

```
IVEware Jackknife Regression Procedure, Tue Jul 15 13:41:41 2014
1

Regression type:          Logistic
Dependent variable:       mde_r
Predictors:               sex  SEX
                          educat  Education in Categories
                          region  Region
Cat. var. ref. codes:     sex  1
                          region  4
                          educat  4
                          mde_r  2
Stratum variable:         str  Strata
Cluster variable:         secu  SECU
Weight variable:          finalp2wt  Final Part 2 weight
```

- 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

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

Variable	Odds Ratio	95% Confidence Interval Lower	Upper
Intercept			
sex	1.6570571	1.4604735	1.8801014
educat.1	0.8347416	0.6656214	1.0468315
educat.2	0.9089556	0.7617580	1.0845969
educat.3	1.0620662	0.9077993	1.2425485
region.1	0.9663815	0.7274006	1.2838775
region.2	0.9632122	0.7853363	1.1813764
region.3	0.8157540	0.6807464	0.9775367

Variable	Design Effect	SRS Estimate	% Diff SRS v Est
Intercept	1.21094	-0.9769619	-39.00545
sex	1.12449	0.5185654	2.67744
educat.1	1.39625	-0.1475597	-18.30971
educat.2	1.15261	-0.0470697	-50.69118
educat.3	1.04220	0.0265876	-55.84643
region.1	2.46956	-0.1698988	396.83027
region.2	1.56901	-0.0118203	-68.46355
region.3	1.30510	-0.1530469	-24.84526

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

```
> # obtain information about missing data
> inf <-mi.info(a)
> # print info about missing data
> inf
       names include order number.mis all.mis                  type collinear
1   sampleid     Yes    NA          0      No unordered-categorical        No
2        sex     Yes    NA          0      No                binary        No
3     region     Yes    NA          0      No unordered-categorical        No
4        age     Yes    NA          0      No   positive-continuous        No
5        str     Yes    NA          0      No   positive-continuous        No
6       secu     Yes    NA          0      No                binary        No
7  finalp2wt     Yes    NA          0      No   positive-continuous        No
8   racecat_     Yes    NA          0      No   ordered-categorical        No
9     educat     Yes     1        235      No unordered-categorical        No
10       mde     Yes     2        165      No                binary        No
11  str_secu     Yes    NA          0      No unordered-categorical        No
>
```

Mice: Impute Missing Data on EDUCAT and MDE

```
> # 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
Missing cells per column:
 sampleid       sex    region       age       str      secu finalp2wt  racecat_    educat       mde
        0         0         0         0         0         0         0         0       235       165
 str_secu
        0
Imputation methods:
 sampleid       sex    region       age       str      secu finalp2wt  racecat_    educat       mde
       ""        ""        ""        ""        ""        ""        ""        "" "polyreg"     "pmm"
 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

```
> # 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        -0.39672207 0.08145831 -0.55638193 -0.2370622      1 %
educat2      0.11943740 0.10932972 -0.09530298  0.3341778      9 %
educat3      0.25810927 0.11257402  0.03738999  0.4788286      3 %
educat4      0.18291288 0.12037457 -0.05308465  0.4189104      3 %
region2      0.08537139 0.13342511 -0.17615071  0.3468935      1 %
region3     -0.06108228 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

```
------------------------------------------------------------------------------
         mde |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       1.sex |  -.5111926   .0638813    -8.00   0.000    -.6403719   -.3820132
             |
      region |
          2  |   .0065904   .1408085     0.05   0.963    -.2783006    .2914814
          3  |  -.1669361   .1330025    -1.26   0.217    -.4358414    .1019692
          4  |    .039644   .1319837     0.30   0.765    -.2271491    .3064372
             |
      educat |
          2  |    .094891   .1074464     0.88   0.383    -.1228383    .3126204
          3  |   .2324122   .1124072     2.07   0.046      .004085    .4607393
          4  |   .1770819   .1114483     1.59   0.120    -.0482573    .4024212
             |
       _cons |  -1.311143   .1484784    -8.83   0.000    -1.611389   -1.010897
------------------------------------------------------------------------------
```

SAS

The MIANALYZE Procedure

Model Information	
PARMS Data Set	WORK.OUTPARMS
Number of Imputations	5

Variance Information

Parameter	sex	educat	region	Variance: Between	Variance: Within	Variance: Total	DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
intercept				0.000137	0.020909	0.021074	65392	0.007883	0.007851	0.998432
sex	1			0.000025634	0.004869	0.004900	101490	0.006318	0.006298	0.998742
educat		2		0.000404	0.011551	0.012036	2462.1	0.042000	0.041085	0.991850
educat		3		0.011276	0.011015	0.024547	13.163	1.228441	0.606783	0.891777
educat		4		0.010145	0.012549	0.024723	16.496	0.970143	0.544492	0.901796
region			2	0.000252	0.017985	0.018287	14632	0.016812	0.016668	0.996677
region			3	0.000275	0.016343	0.016673	10200	0.020203	0.019995	0.996017
region			4	0.000329	0.016842	0.017236	7643.5	0.023412	0.023132	0.995395

Parameter Estimates

Parameter	sex	educat	region	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
intercept				-1.302637	0.145168	-1.58717	-1.01811	65392	-1.314556	-1.284028	0	-8.97	<.0001
sex	1			-0.504400	0.069998	-0.64160	-0.36720	101490	-0.510147	-0.496238	0	-7.21	<.0001
educat		2		0.095271	0.109708	-0.11986	0.31040	2462.1	0.071256	0.112619	0	0.87	0.3853
educat		3		0.309667	0.156675	-0.02838	0.64772	13.163	0.171845	0.396707	0	1.98	0.0694
educat		4		0.214843	0.157236	-0.11767	0.54736	16.496	0.126754	0.324696	0	1.37	0.1902
region			2	-0.022962	0.135231	-0.28803	0.24211	14632	-0.040051	-0.001512	0	-0.17	0.8652
region			3	-0.194508	0.129125	-0.44762	0.05860	10200	-0.210003	-0.167187	0	-1.51	0.1320
region			4	0.019060	0.131287	-0.23830	0.27642	7643.5	-0.007351	0.038419	0	0.15	0.8846

Comparison of Results from Stata, SAS, IVEware and R, continued

IVEWARE:

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

Variable	Odds Ratio	95% Confidence Interval Lower	Upper
Intercept			
sex	1.6570571	1.4604735	1.8801014
educat.1	0.8347416	0.6656214	1.0468315
educat.2	0.9089556	0.7617580	1.0845969
educat.3	1.0620662	0.9077993	1.2425485
region.1	0.9663815	0.7274006	1.2838775
region.2	0.9632122	0.7853363	1.1813764
region.3	0.8157540	0.6807464	0.9775367

```
> # 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        -0.39672207 0.08145831 -0.55638193 -0.2370622      1 %
educat2      0.11943740 0.10932972 -0.09530298  0.3341778      9 %
educat3      0.25810927 0.11257402  0.03738999  0.4788286      3 %
educat4      0.18291288 0.12037457 -0.05308465  0.4189104      3 %
region2      0.08537139 0.13342511 -0.17615071  0.3468935      1 %
region3     -0.06108228 0.13851353 -0.33256725  0.2104027      1 %
region4      0.10233875 0.13072914 -0.15390103  0.3585785      1 %
```

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
```

SAS Code

```
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 ;
```

IVEware Code

```
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

Allison, Paul D., “Missing Data”, Sage Publications 2001.

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

Heitjan, Daniel F. 1997. “Annotation: What can be done about missing data? Approaches to imputation.” *American Journal of Public Health* 87: 548–550.

Horton, N.J. and Lipsitz, S.R. 2001. “Multiple Imputation in Practice: Comparison of Software Packages for Regression Models with Missing Variables.” *Journal of the American Statistical Association* 55: 244–254.

Raghunathan, T.E., Lepkowski, J., Van Hoewyk, J., and Solenberger, P. 2001. “A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models.” Survey Methodology: 27, pages 85-95.

Reiter, J.P., Raghunathan, T.E., and Kinney, V. 2006. “The importance of modeling the sampling design in multiple imputation for missing data.” *Survey Methodology* 32.2: 143–150.

References

Rubin, D.B. 1976. “Inference and Missing Data.” *Biometrika* 63: 581–592.

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons, Inc.

Rubin, D.B. 1996. “Multiple Imputation After 18+ Years.” *Journal of the American Statistical Association* 91: 473–489.

Schafer, J.L. 1997. *Analysis of Incomplete Multivariate Data.* New York: Chapman and Hall.

Schafer, J.L. 1999. “Multiple Imputation: A Primer.” *Statistical Methods in Medical Research* 8: 3–15.

van Buuren, S., Boshuizen, H.C., and Knook, D.L. 1999. “Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis.” *Statistics in Medicine* 18: 681–694.

van Buuren, S. (2007), "Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification," *Statistical Methods in Medical Research*, 16, 219–242.

Van Buuren, S. (2012), *Flexible Imputation of Missing Data*. Chapman & Hall/CRC, Boca Raton, FL. ISBN 9781439868249.

Yuan, Y.C. “Multiple Imputation for Missing Data: Concepts and New Development.” *Proceedings of the SAS Users Group International Conference*. Paper 267–25.