More on Multiple Imputation of Complex Sample Design Data Using SAS 9.3

Patricia A. Berglund, Institute for Social Research - University of Michigan
Wisconsin and Illinois SAS User's Group
November 12, 2012

Overview of Presentation

- Builds on previous Global Forum paper (2010) "An Introduction to Multiple Imputation of Complex Sample Survey Data Using SAS 9.2" including a brief review of missing data patterns and additional/new features of the MI process in SAS 9.3:
    - FCS imputation method (experimental in SAS 9.3)
    - MCMC diagnostic tools
    - Imputation of missing data in longitudinal data set

- Applications using data from a complex sample survey data set with demonstration of 3 steps of multiple imputation
    1. Imputation of missing data using PROC MI
    2. Analysis of imputed data sets using SAS SURVEY procedures, differs from "standard" SAS procedures which use SRS assumption
    3. Analysis of pooled results from Steps 1 and 2 using PROC MIANALYZE

Analysis of Data Sets with Item Missing Data

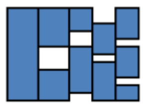
- How to analyze?
  - Do nothing, use either complete case or available cases, can be significant loss of data to analyze
  - Simple imputation using mean/median substitution, Hotdeck (similar record used to impute missing data), these approaches are easy to implement but lack precision for variance estimates
  - Multiple imputation is generally preferred to simple imputation because it uses statistically appropriate methods and accounts for variability introduced by the imputation process, better precision of variance estimates

- PROC MI for multiple imputation in SAS, assumes data is missing at random (MAR)
  - Means that missingness can be predicted from observed covariates
  - Basic statistical assumption of PROC MI

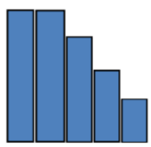
Missing Data Patterns

- The pattern of missing data has an impact on how the imputation process is applied, two types of missing data patterns:

  - Arbitrary



  - Monotone



5

Overview of Imputation Methods Table 56.5 (PROC MI)

Pattern of Missingness	Type of Imputed Variable	Type of Covariates	Available Methods
Monotone	Continuous	Arbitrary	Monotone regression
			Monotone predicted mean matching
			Monotone propensity score
Monotone	Classification (ordinal)	Arbitrary	Monotone logistic regression
Monotone	Classification (nominal)	Arbitrary	Monotone discriminant function
Arbitrary	Continuous	Continuous	MCMC full-data imputation
			MCMC monotone-data imputation
Arbitrary	Continuous	Arbitrary	FCS regression
			FCS predicted mean matching
Arbitrary	Classification (ordinal)	Arbitrary	FCS logistic regression
Arbitrary	Classification (nominal)	Arbitrary	FCS discriminant function

Detail on Imputation Methods

- MCMC
  - Markov Chain Monte Carlo method, assume MVN (MultiVariateNormal)
  - Recommended for imputation of continuous variables with continuous covariates and with arbitrary missing data pattern, robust to violations though
  - How to assess convergence of MCMC?
    - Trace, WLF (worst linear function), and autocorrelation plots
- Monotone methods
  - Use of appropriate method depending on type of variable to be imputed, for example binary, ordinal, count, continuous imputed variables
  - Monotone pattern is convenient since a series of independent models can be estimated for imputation, builds on previous model(s)
- FCS method
  - Convenient for typically "messy" missing data problems with a variety of variables to be imputed and arbitrary missing data pattern

Example 1 - FCS Imputation Method

- Experimental in SAS 9.3, the FCS (Fully Conditional Specification) method allows the user to impute missing data with arbitrary missing data patterns
- FCS belongs to a class of imputation methods that use flexible “chained models” to impute missing data, different approach than used in imputation of montone missing data, see SAS/STAT PROC MI documentation for details or Van Buuren (2012) “Flexible Imputation of Missing Data” Chapman Hall
- This example demonstrates use of the FCS method for imputation of missing data on both continuous and categorical variables with an arbitrary missing data pattern
- Data is from the NCS-R data set, a nationally representative survey focused on mental health and related issues and based on a complex sample design survey

Means Analysis of NCS-R Data Set

The MEANS Procedure

Variable	Label	N	N Miss	Mean	Minimum	Maximum
DSM_GAD	DSM-IV Generalized Anxiety Disorder(Lifetime)	5692	0	4.4757554	1.0000000	5.0000000
DSM_SO	DSM-IV Social Phobia(Lifetime)	5207	485	4.2179758	1.0000000	5.0000000
DSM_SP	DSM-IV Specific Phobia(Lifetime)	5692	0	4.1918482	1.0000000	5.0000000
Sex	SEX	5692	0	0.4184821	0	1.0000000
Age	AGE	5692	0	43.3780745	17.0000000	98.0000000
educat	Education 4 categories(non imputed)	5685	7	2.6504837	1.0000000	4.0000000
marcat	Marital category	5689	3	1.6451046	1.0000000	3.0000000
str	Strata	5692	0	26.3787772	1.0000000	42.0000000
secu	Sampling Error Computing Unit	5692	0	1.5052706	1.0000000	2.0000000
finalp2wt	Final Part 2 weight	5692	0	1.0000001	0.1144058	10.1020733
racecat_	Race category(Imputed)	5692	0	3.4232256	1.0000000	4.0000000
inc_rsp	Respondent Income	4849	843	24573.59	0	125000.00

- Missing data on 4 variables is highlighted in red:
  - DSM_SO is a binary variable indicating a diagnosis of Social Phobia - coded 1 (YES) or 5 (NO)
  - EDUCAT is a categorical variable with 1-4 (lowest level of education to highest education)
  - MARCAT is a categorical variable with 3 levels: 1=married 2=previously married 3=never married
  - INC_RSP is a continuous variable containing personal income

Examination of Missing Data Pattern

```
proc mi data=ex1_ncsr nimpute=0 ;
run ;
```

The MI Procedure

Model Information

Data Set	WORK.EX1_NCSR
Method	MCMC
Multiple Imputation Chain	Single Chain
Initial Estimates for MCMC	EM Posterior Mode
Start	Starting Value
Prior	Jeffreys
Number of Imputations	0
Number of Burn-in Iterations	200
Number of Iterations	100
Seed for random number generator	596031000

Missing Data Patterns

Group	DSM_GAD	DSM_SO	DSM_SP	Sex	Age	educat	marcat	str	secu	finalp2wt	racecat_	inc_rsp	Freq	Percent
1	X	X	X	X	X	X	X	X	X	X	X	X	4355	76.51
2	X	X	X	X	X	X	X	X	X	X	X	.	843	14.81
3	X	X	X	X	X	X	.	X	X	X	X	X	2	0.04
4	X	X	X	X	X	.	X	X	X	X	X	X	7	0.12
5	X	.	X	X	X	X	X	X	X	X	X	X	484	8.50
6	X	.	X	X	X	X	.	X	X	X	X	X	1	0.02

- Six distinct groups with missing data rates ranging from 0.02 to 14.81%
- Arbitrary missing data pattern with DSM_SO, EDUCAT, MARCAT, INC_RSP requiring imputation

Imputation of Missing Data and Check of Process

- Red highlights show syntax for FCS imputation with specific model to impute MARCAT (marital status) and default models for other imputed variables (all variables used) , NBITER=5 requests 5 "burn-in" iterations
- Use of /DETAILS option shows imputation coefficients and other details of the process
- 5 imputed data sets are created in this process

```
proc mi data=ex1_ncsr out=outex1 seed=1112 nimpute=5 ;
class dsm_gad dsm_sp secu dsm_so educat sex marcat racecat_ ;
fcs nbiter=5 order=var
  logistic (marcat=dsm_gad dsm_sp str secu finalp2wt sex age racecat_ / details)
  logistic (educat/ details)
  logistic (dsm_so/ details)
  reg(inc_rsp/ details);
  var dsm_gad dsm_sp str secu finalp2wt sex age racecat_ marcat educat dsm_so
   inc_rsp ;
run;
```

11

Burn in iterations are run prior to imputation and allow the chain to stabilize before the filling in of values. Note that the variable MARCAT is assumed to be ordinal in this example. This allows use of the class variables in the imputation model however a comparison using the discriminant function method for imputation of marcat was done and results were very similar. For the purpose of this example, assume that MARCAT is an ordinal variable.

Output from PROC MI with /Details Option

- Example of output from /details option on FCS statement, shows estimates for the imputed variable MARCAT for each of the 5 imputed data sets

Logistic Models for FCS Method

Imputed Variable	Effect	marcat	DSM_GAD	DSM_SP	secu	Sex	racecat_	Imputation 1	Imputation 2	Imputation 3	Imputation 4	Imputation 5
marcat	Intercept	1.000000	.	.	.	.	.	-0.013115	0.036111	0.011554	0.046493	-0.000150
marcat	Intercept	2.000000	.	.	.	.	.	1.129444	1.104309	1.113927	1.126601	1.108605
marcat	DSM_GAD	.	1.000000	.	.	.	.	-0.139653	-0.135746	-0.109061	-0.056286	-0.082520
marcat	DSM_SP	.	.	1.000000	.	.	.	-0.006082	-0.057440	-0.011175	0.021214	0.003464
marcat	str	.	.	.	.	.	.	0.152436	0.114497	0.180760	0.146378	0.080557
marcat	secu	.	.	.	1.000000	.	.	-0.015715	-0.038588	-0.029257	-0.056189	-0.028366
marcat	finalp2wt	.	.	.	.	.	.	-0.200221	-0.176056	-0.141248	-0.145683	-0.113177
marcat	Sex	.	.	.	.	0	.	-0.112413	-0.122972	-0.135173	-0.129839	-0.105792
marcat	Age	.	.	.	.	.	.	0.515771	0.480843	0.499158	0.503565	0.503839
marcat	racecat_	.	.	.	.	.	1.000000	0.242861	0.205465	0.195057	0.092029	0.261217
marcat	racecat_	.	.	.	.	.	2.000000	-0.455928	-0.470378	-0.507940	-0.457066	-0.542520
marcat	racecat_	.	.	.	.	.	3.000000	-0.029085	0.052478	0.048119	0.061939	-0.037141

Check of Imputed Data Sets

```
The MEANS Procedure
                              Analysis Variable : inc_rsp
  Imputation
      Number   N Obs       N          Mean        Std Dev        Minimum       Maximum
-------------------------------------------------------------------------------------
           1    5692    5692      24550.85       26523.82      -64205.22     125000.00
           2    5692    5692      24551.45       26831.98      -74659.84     125000.00
           3    5692    5692      24663.71       26714.12      -51309.18     125000.00
           4    5692    5692      24725.44       26885.61      -81236.09     125000.00
           5    5692    5692      24439.65       26752.89      -46573.77     125000.00
-------------------------------------------------------------------------------------
```

```
proc means data=outex1;
var inc_rsp;
class _imputation_;
run;
```

```
Table of _Imputation_ by DSM_SO
_Imputation_(Imputation Number)
         DSM_SO(DSM-IV Social Phobia(Lifetime))
Frequency|
Percent  |
Row Pct  |
Col Pct  |       1|       5|  Total
---------+--------+--------+
       1 |   1112 |   4580 |   5692
         |   3.91 |  16.09 |  20.00
         |  19.54 |  80.46 |
         |  20.08 |  19.98 |
---------+--------+--------+
       2 |   1108 |   4584 |   5692
         |   3.89 |  16.11 |  20.00
         |  19.47 |  80.53 |
         |  20.00 |  20.00 |
---------+--------+--------+
       3 |   1112 |   4580 |   5692
         |   3.91 |  16.09 |  20.00
         |  19.54 |  80.46 |
         |  20.08 |  19.98 |
---------+--------+--------+
       4 |   1101 |   4591 |   5692
         |   3.87 |  16.13 |  20.00
         |  19.34 |  80.66 |
         |  19.88 |  20.03 |
---------+--------+--------+
       5 |   1106 |   4586 |   5692
         |   3.89 |  16.11 |  20.00
         |  19.43 |  80.57 |
         |  19.97 |  20.01 |
---------+--------+--------+
Total        5539    22921    28460
            19.46    80.54   100.00
```

```
proc freq data=outex1;
tables _imputation_*(marcat educat dsm_so);
run;
```

13

Analysis of Completed Data Sets with SURVEYLOGISTIC and Print-Out of Estimates Data Set

```
proc surveylogistic data=outex1 ;
strata str ; cluster secu ; weight finalp2wt ;
class sex marcat racecat_  / param=ref ;
model dsm_so (event='1') =age sex marcat racecat_  ;
by _imputation_ ;
ods output parameterestimates=outestex1 ;
run ;

proc print data=outestex1 ;
run ;
```

Obs	_Imputation_	Variable	Class Val0	DF	Estimate	StdErr	WaldChiSq	Prob ChiSq
1	1	Intercept		1	-1.1798	0.1169	101.8355	<.0001
2	1	Age		1	-0.0199	0.00265	56.4076	<.0001
3	1	Sex	0	1	0.0718	0.0808	0.7905	0.3739
4	1	marcat	1	1	0.0411	0.0929	0.1955	0.6584
5	1	marcat	2	1	0.4008	0.1335	9.0087	0.0027
6	1	racecat_	1	1	-0.5421	0.1523	12.6680	0.0004
7	1	racecat_	2	1	-0.2980	0.1340	4.9476	0.0261
8	1	racecat_	3	1	0.2743	0.1754	2.4440	0.1180
9	2	Intercept		1	-1.2279	0.1141	115.7528	<.0001
10	2	Age		1	-0.0188	0.00295	40.6859	<.0001
11	2	Sex	0	1	0.0857	0.0718	1.4245	0.2327
12	2	marcat	1	1	0.0191	0.0929	0.0422	0.8373
13	2	marcat	2	1	0.4057	0.1369	8.7772	0.0031
14	2	racecat_	1	1	-0.5156	0.1452	12.6195	0.0004
15	2	racecat_	2	1	-0.2814	0.1245	5.1058	0.0238
16	2	racecat_	3	1	0.2825	0.1505	3.5246	0.0605

ETC.

14

Note that each imputed data set or _IMPUTATION_ =1,2,3,4,5 has separate observations in this output data set. Use of the BY statement will trigger a warning in the log but since this is the entire data set not a subpopulation, it is statistically appropriate.

Pooled Results from PROC MIANALYZE

```
proc mianalyze parms(classvar=classval)=outestex1;
class sex marcat racecat_;
modeleffects intercept age sex marcat racecat_ ;
run ;
```

The MIANALYZE Procedure

Model Information

PARMS Data Set	WORK.OUTESTEX1
Number of Imputations	5

Variance Information

Parameter	sex	marcat	racecat_	Variance: Between	Variance: Within	Variance: Total	DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
intercept				0.001539	0.012492	0.014339	241.07	0.147860	0.135953	0.973529
age				0.000000490	0.000007014	0.000007601	669.03	0.083802	0.080068	0.984239
sex	0			0.000069390	0.006112	0.006195	22143	0.013624	0.013530	0.997301
marcat		1		0.000297	0.008000	0.008357	2198.5	0.044555	0.043525	0.991370
marcat		2		0.000244	0.016721	0.017014	13491	0.017521	0.017365	0.996539
racecat_			1	0.002015	0.025534	0.027952	534.57	0.094693	0.089901	0.982337
racecat_			2	0.000375	0.016570	0.017020	5732.2	0.027133	0.026756	0.994677
racecat_			3	0.000905	0.024860	0.025946	2282.7	0.043689	0.042699	0.991533

Parameter Estimates

Parameter	sex	marcat	racecat_	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
Intercept				-1.241283	0.119744	-1.47716	-1.00540	241.07	-1.274731	-1.179812	0	-10.37	<.0001
age				-0.018759	0.002757	-0.02417	-0.01335	669.03	-0.019933	-0.018156	0	-6.80	<.0001
sex	0			0.082079	0.078710	-0.07220	0.23636	22143	0.071814	0.093748	0	1.04	0.2971
marcat		1		0.043083	0.091415	-0.13619	0.22235	2198.5	0.019078	0.067519	0	0.47	0.6375
marcat		2		0.408207	0.130437	0.15253	0.66388	13491	0.392911	0.434289	0	3.13	0.0018
racecat_			1	-0.556629	0.167189	-0.88506	-0.22820	534.57	-0.627261	-0.515636	0	-3.33	0.0009
racecat_			2	-0.289776	0.130461	-0.54553	-0.03402	5732.2	-0.317222	-0.265301	0	-2.22	0.0264
racecat_			3	0.306468	0.161078	-0.00941	0.62234	2282.7	0.274285	0.345390	0	1.90	0.0572

15

Results show averaged estimates (from imputed data sets and PROC SURVEYLOGISTIC), variance information including between, within, and total variance plus Relative Increase in Variance due to missing data, Fraction Missing Information (due to missing among all variables in analysis) and Relative Efficiency (how efficient is imputation by variable).

Red highlights indicate significant predictors of having Social Phobia, adjusted for complex survey design (SURVEYLOGISTIC) and variability due to imputation process. These results would be interpreted as usual for binary outcome with logistic regression but recognizing the use of the SURVEY procedure and imputation variability.

Summary of Example 1

- The experimental FCS imputation method offers a new approach and increased flexibility for imputation of categorical and continuous variables with arbitrary missing data patterns
- Prior to this experimental feature, imputation of categorical variables required a monotone pattern or relaxing of MVN assumption
- Arbitrary missing data patterns are common in “real world” data sets so very useful new option in PROC MI
- With complex sample design data, SURVEY procedures in the second step of the MI process are required, important for correct variance estimation

Example 2: MCMC Method with Diagnostic Plots

- The second example demonstrates how to carry out the MI process using a subset of categorical and continuous variables from the NHANES 2005-2006 data set, another nationally representative complex sample , focused on health and nutrition issues
- As in the analysis of the NCS-R data set, the standard errors should be adjusted to account for stratification, clustering, and other complex sample features
- This example also uses a DOMAIN statement for correct analysis of a subpopulation of those Male and Mexican using SURVEY commands along with a BY statement for the multiple imputations (BY _IMPUTATION_)

Examination of Missing Data Pattern

- Grid indicates a monotone pattern with missing on BMXBMI (1.82%), mix of continuous and categorical variables will be used to impute BMXBMI
- PROC MI with NIMPUTE=0 to produce grid
- Note that some variables such as SEQN, MALE_MEXICAN will not be used in imputation because they are utility variables (case id and domain indicator)
- Our analysis goal is to examine mean BMI in the Male/Mexican subgroup and compare to mean BMI in the non-Male/Mexican group

Missing Data Patterns

Group	SEQN	RIAGENDR	RIDRETH1	WTMEC2YR	SDMVPSU	SDMVSTRA	BMXBMI	male_ mexican	Freq	Percent
1	X	X	X	X	X	X	X	X	5237	98.18
2	X	X	X	X	X	X	.	X	97	1.82

Missing Data Patterns

	Group Means							
Group	SEQN	RIAGENDR	RIDRETH1	WTMEC2YR	SDMVPSU	SDMVSTRA	BMXBMI	male_mexican
1	36353	1.520336	2.867863	40937	1.505442	50.556616	28.511957	0.099866
2	36657	1.494845	3.041237	34134	1.484536	50.010309	.	0.082474

18

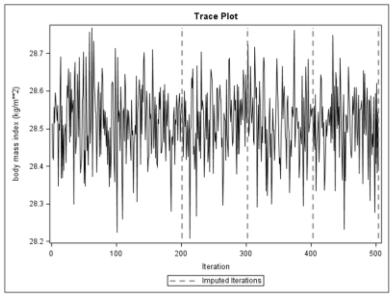
Covariates include SEQN (CASEID), RIAGENDR (GENDER), RIDRETH1 (RACE/ETH), WTMEC2YR (MEDICAL EXAM WEIGHT FOR 2 YRS), SDMVPSU (MASKED PSU), SDMVSTRA (MASKED STRATA). Mix of continuous and categorical covariates.

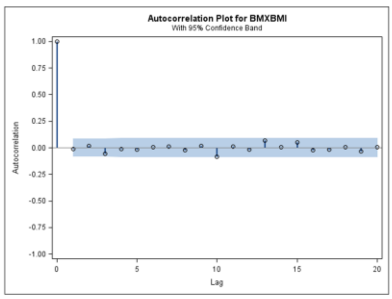
Impute Missing Data using MCMC Method

```
proc mi data=nhanes0506 nimpute=4 seed=555 out=imp_ex2 ;
mcmc plots=( trace(mean(bmxbmi)) acf(mean(bmxbmi)) );
var wtmec2yr sdmvstra sdmvpsu riagendr ridreth1 bmxbmi;
run ;
```

- Imputation uses continuous and categorical covariates to impute a continuous variable (use of categorical predictors with no missing is innocuous though this method assumes all variables are multivariate normal)
- MCMC plots statement requests plots to evaluate convergence of the Markov Chains, plots are informative about convergence status
- Trace and ACF (autocorrelation) plots provide a way to evaluate patterns among parameter estimates across iterations, look for no obvious patterns or large positive/negative autocorrelations
- The order of the variables in the VAR statement is important -fully observed variables first followed by variable (BMXBMI) with missing data

MCMC Diagnostic Plots





• The Trace Plot for BMI shows no apparent pattern among parameter estimates in the four iterations (iterations indicated by vertical dotted lines). The first 200 points on the x axis (before th dotted vertical line) represent the “burn-in” iterations.

• The ACF Plot also indicates no apparent pattern with a good mix of small autocorrelations (positive and negative) for the 20 lagged time points (except for lag=0, as expected). The NLAG= option defaults to 20 lagged points but can be changed.

• These results suggest little concern about the convergence of the MCMC iterative approach for imputation of BMXBMI.

20

Analysis of Imputed Data Sets with PROC SURVEYMEANS

- In Step 2, PROC SURVEYMEANS with a DOMAIN statement (using an indicator of being a Mexican Male) is used to produce a means analysis for each of 4 imputed data sets along with a BY statement for each imputed data set (not a random variable, multiple complete data sets)
  - use of DOMAIN rather than a BY statement for subgroup analysis is correct way to analyze subgroups in survey data sets, unconditional approach preserves full design variable information and random variability
- Use of the ODS output option creates a data set for use in Step 3 (PROC MIANALYZE)

```
proc surveymeans data=imp_ex2 ;
strata sdmvstra ; cluster sdmvpsu ; weight wtmec2yr ;
var bmxbmi ;
domain male_mexican ;
by _imputation_ ;
ods output domain=outmeans ;
run ;
```

Listing of Data Set Produced by PROC SURVEYMEANS

- List output of the data set called “outmeans” shows the mean and SE for BMI from the DOMAIN analysis of Male/Mexican or not Male/Mexican (by each of 4 imputed data sets): therefore 8 different means for 2 domain values*4 imputed data sets are set for use in PROC MIANALYZE

Imputation	DomainLabel	male_ mexican	Var Name	VarLabel	N	Mean	StdErr
1	male_mexican	0	BMXBMI	body mass index (kg/m**2)	4803	28.472426	0.233170
1	male_mexican	1	BMXBMI	body mass index (kg/m**2)	531	28.085561	0.330402
2	male_mexican	0	BMXBMI	body mass index (kg/m**2)	4803	28.470355	0.236915
2	male_mexican	1	BMXBMI	body mass index (kg/m**2)	531	28.054206	0.327459
3	male_mexican	0	BMXBMI	body mass index (kg/m**2)	4803	28.454714	0.233496
3	male_mexican	1	BMXBMI	body mass index (kg/m**2)	531	28.070451	0.342631
4	male_mexican	0	BMXBMI	body mass index (kg/m**2)	4803	28.466996	0.234309
4	male_mexican	1	BMXBMI	body mass index (kg/m**2)	531	28.090376	0.338132

Pooling Results in PROC MIANALYZE

- Prior to use in PROC MIANALYZE, the data set must by sorted by the DOMAIN and _IMPUTATION_ variables, this is due to the need to analyze the means for Male/Mexican=1 or 0 over the 4 imputed data sets
- The BY statement in PROC MIANALYZE provides means and standard errors for BMI by the values of the DOMAIN variable

```
proc sort ;
  by male_mexican _imputation_ ;
run ;

proc mianalyze data=outmeans ;
by male_mexican ;
modeleffects mean ;
stderr stderr ;
run ;
```

23

Output from PROC MIANALYZE

- Results from PROC MIANALYZE show that mean (se) BMI for those Male/Mexican is 28.08 (0.23) while for those not Male/Mexican it is 28.47 (0.34)
- The SE's are corrected for the complex sample and the imputation variability while using a correct DOMAIN statement for analysis of subgroups

```
male_mexican=0
Parameter Estimates

Parameter  Estimate     Std Error    95% Confidence Limits      DF        Minimum       Maximum
mean       28.466123    0.234645     28.00623     28.92602    1.47E6     28.454714     28.472426

male_mexican=1
The MIANALYZE Procedure
Parameter Estimates

Parameter  Estimate     Std Error    95% Confidence Limits      DF        Minimum       Maximum
mean       28.075148    0.335209     27.41815     28.73215    340089     28.054206     28.090376
```

Summary of Example 2

- The second example demonstrates use of diagnostic plots to evaluate convergence of the iterative MCMC process
- The covariates used in the imputation are both continuous and categorical and though the MCMC method assumes MVN, use of categorical predictors without missing data is harmless to violation of this assumption
- Use of PROC SURVEYMEANS and a DOMAIN statement for an unconditional analysis of a random variable plus the BY statement for use with the _IMPUTATION_ variable (fixed sample size per data set)
- Also demonstrates use of the BY statement in PROC MIANALYZE to obtain means and standard errors for each level of the DOMAIN variable used in the means analysis

Example 3 – Imputation of Longitudinal Data

- The final example demonstrates how to use multiple imputation for longitudinal data
- HRS (Health and Retirement Survey) 2004-2006 data is used in this example to examine relationship between total assets in 2004 and 2006 predicted by education level of the financial respondent
- Discussion of correct data structure for accounting for dependence between repeated records per individual, how to build this into the imputation step
- HRS is a complex sample survey, again use SURVEY procedures in the analysis of completed data sets in Step 2 of the MI process

Structure of Longitudinal HRS Data

- Data on Total HH Assets from 2004 and 2006 is collected in “long” or multiple records per HH financial respondent
- Analysis goal is the examine impact of education of HH financial respondent on total assets for the two years of interest
- Missing data on total assets from 2004, requires imputation
- Issue with imputing missing data is that records are inherently correlated due to repetition
- One solution is to use a one record per individual data set for imputation with differently named variables for each time point
  - captures impact of time and allows us to use the measurements at different points in time in the imputations

27

Longitudinal Data Structure

- The current data structure has 2 records per individual with different values for TOTALASSETS and WEIGHT for 2004 and 2006 (YR)
- The values for the other variables are time invariant so we will need to create new variables for total assets in 2004 and 2006 and weights for 2004 and 2006

HHIDPN	SECU	STRATUM	EDCAT	WEIGHT	TOTALASSETS	YR
3010	1	40	2	4394	756000	2004
3010	1	40	2	4528	914000	2006
10001010	2	1	2	9084	450000	2004
10001010	2	1	2	8706	1000000	2006
10003030	2	1	4	0	20500	2004
10003030	2	1	4	0	12000	2006
10004010	2	1	4	5111	1973000	2004
10004010	2	1	4	5422	1832000	2006
10013010	2	1	2	5564	500	2004
10013010	2	1	2	5315	50	2006

Create a One Record Per Individual Data Set

- Use of arrays to turn the multiple record data set into 1 record per individual

```
data onerec ;
array ta [2] totalassets2004 totalassets2006  ;
retain totalassets2004 totalassets2006 ;
array wgt [2] wgt2004 wgt2006 ;
retain wgt2004 wgt2006 ;
set hrs2004_2006 ;

by hhidpn yr ;
ta[yr] =totalassets ;
wgt[yr]=weight ;
if last.hhidpn then output ;
drop totalassets weight yr ;

proc print data=onerec (obs=5 ) ;
run ;
```

Obs	totalassets2004	totalassets2006	wgt2004	wgt2006	HHIDPN	SECU	STRATUM	EDCAT
1	756000	914000	4394	4528	3010	1	40	2
2	450000	1000000	9084	8706	10001010	2	1	2
3	20500	12000	0	0	10003030	2	1	4
4	1973000	1832000	5111	5422	10004010	2	1	4
5	500	50	5564	5315	10013010	2	1	2

Imputation of Missing Data

```
proc mi nimpute=0 ;
run ;
```

Missing Data Patterns

Group	totalassets2004	totalassets2006	wgt2004	wgt2006	HHIDPN	SECU	STRATUM	EDCAT	Freq	Percent
1	X	X	X	X	X	X	X	X	7993	97.03
2	.	X	X	X	X	X	X	X	245	2.97

```
proc mi data=onerec nimpute=3 seed=765 out=outimp_ex3 ;
     class edcat ;
     monotone regression (totalassets2004=totalassets2006 wgt2004 wgt2006 stratum secu edcat) ;
     var totalassets2006 wgt2004 wgt2006 stratum secu edcat totalassets2004 ;
run ;
```

Parameter Estimates

Variable	Mean	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Mu0
totalassets2004	413078	18737	376304.0	449851.3	883.54	409117	415255	0

t for H0: Mean=Mu0	Pr > |t|
22.05	<.0001

30

Reverse Data Set Structure to Multiple Records

```
data multrec ;
set outimp_ex3 ;
array ta [2] totalassets2004 totalassets2006 ;
array wgt [2] wgt2004 wgt2006 ;
do i = 1 to 2 ;
  weight = wgt[i] ;
  totalassets=ta[i] ;
  year_int=i  ;
  if year_int=1 then year_int=2004 ;
  if year_int=2 then year_int=2006 ;
output ;
end ;
```

- Restructure data set for rest of multiple imputation process, with completed data sets taking the dependence between individual records into account, the 2nd and 3rd steps can be done using the multiple records data set

Check of Multiple Records Data Set

- Double check shows the data is now in original format but has imputed data on the variable called "Totalassets"
- Analysis of the multiple record file can proceed

HHIDPN	weight	totalassets	year_int
3010	4394	756000	2004
3010	4528	914000	2006
10001010	9084	450000	2004
10001010	8706	1000000	2006
10003030	0	20500	2004
10003030	0	12000	2006
10004010	5111	1973000	2004
10004010	5422	1832000	2006
10013010	5564	500	2004
10013010	5315	50	2006

Analyze using PROC SURVEYREG

- Use SURVEYREG for analysis of Total Assets predicted by Education, controlling for year
- Repeat the regression for each multiple imputation iteration (3 data sets) and account for the complex sample design
- Data step with use of the compress function to prepare the output data set for PROC MIANALYZE (removes white space in variable "parameter")

```
proc surveyreg data=multrec ;
   strata stratum ; cluster secu ; weight weight ;
   class edcat year_int ;
   model totalassets=edcat year_int / solution  ;
   by _imputation_ ;  ods output parameterestimates=outest_ex3 ;
run ;

proc print data=outest_ex3 ;
Run;

data outest_ex3 ;
   set outest_ex3 ;
   parameter=compress (parameter) ;
run ;
```

MIANALYZE for Combining Results

The MIANALYZE Procedure

Model Information

PARMS Data Set	WORK.OUTEST_EX3
Number of Imputations	3

Variance Information

Parameter	Variance Between	Variance Within	Variance Total	DF	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
intercept	1176794	11348071508	11349640567	1.05E8	0.000138	0.000138	0.999954
edcat1	1647.467192	9713891546	9713893743	391E11	0.000000226	0.000000226	1.000000
edcat2	25648	10190790851	10190825049	178E9	0.000003356	0.000003356	0.999999
edcat3	28369030	10450157859	10487983232	153761	0.003620	0.003620	0.998795
year_int2004	4669269	590973248	597198940	18403	0.010535	0.010532	0.996502

Parameter Estimates

Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum
intercept	1183719	106535	974915	1392524	1.05E8	1182495	1184562
edcat1	-944390	98559	-1137562	-751218	391E11	-944421	-944344
edcat2	-797630	100950	-995488	-599773	178E9	-797814	-797519
edcat3	-673088	102411	-873811	-472364	153761	-677121	-667050
year_int2004	-124131	24438	-172031	-76230	18403	-125808	-121692

Parameter Estimates

Parameter	Theta0	t for H0: Parameter=Theta0	Pr > |t|
intercept	0	11.11	<.0001
edcat1	0	-9.58	<.0001
edcat2	0	-7.90	<.0001
edcat3	0	-6.57	<.0001
year_int2004	0	-5.08	<.0001

```
proc mianalyze parms=outest_ex3 ;
  modeleffects intercept edcat1 edcat2
  edcat3 year_int2004 ;
run ;
```

34

Interpretation of results is similar to any linear regression but mention of the imputed data sets and use of PROC SURVEYREG is expected.

Example 3 Summary

- Longitudinal data imputation requires recognition of the dependence of repeated records per unit of analysis and accounting for “over time” effects
- One way to address issue is to restructure the data to a one record per individual with differently named variables and impute this data set
- Then, change back to a multiple record per individual data set for analysis of completed files and use of PROC MIANALYZE for pooling the results
- This example uses PROC SURVEYREG with a dummy variable for year to account for multiple records per individual

Presentation Summary

- This presentation has covered three areas of interest to analysts of complex sample design data sets with missing data
  - Use of the FCS imputation method for imputation of arbitrary missing data
  - Use of diagnostic tools to evaluate the MCMC convergence status while imputing continuous variables with mixed covariates
  - Imputation of longitudinal data with use of data re-structuring concepts and imputation while accounting for time varying variables

- The examples are intended to provide practical guidance to analysts using all types of data sets but particularly those using complex sample design data sets

Author Contact Information

- Your comments and feedback are welcome and thank you for attending today!

- Patricia Berglund
- pberg@umich.edu