Measurement, Design, and Analytic Techniques in Mental Health and Behavioral Sciences

Lecture 8 (Jan 30, 2007): SAS Proc MI and Proc MiAnalyze

XH Andrew Zhou

azhou@u.washington.edu

Professor, Department of Biostatistics, University of Washington

Imputation methods in SAS Proc MI Procedure

- Regression method
- Predictive mean matching method
- Propensity score
- Logistic regression
- Discriminant function method
- MCMC Data Augmentation method

SAS Proc MI Procedure, cont

Table 1: Imputation Methods in PROC MI

Pattern of Missingness	Type of Imputed Variables	Recommended Methods
Monotone	Continuous	Regression method Predictive Mean Matching Propensity Score
Monotone	Ordinal categorical	Logistic Regression
Monotone	Nominal categorical	Discriminant function method
Arbitrary	Continuous	MCMC Data Augmentation

MI for continuous variables

- Assumption: data, $Y_1, \ldots, Y_p$, are from a continuous multivariate distribution and contain missing data values that can occur for any of the variables.
- Monotone missing-data pattern: regression method (multivariate normal), predictive mean matching (multivariate normal), propensity score (non-parametric), and MCMC data augmentation (multivariate normal).
- Arbitrary missing-data pattern: MCMC data augmentation (multivariate normal assumption).

Principe behind Regression method MI for monotone missing data

- The data, $Y_1, \ldots, Y_p$ (in that order) is said to have a monotone missing pattern if an individual has an observed value on a variable Y_j, all previous variables $Y_j, j < k$, are also observed for that individual.
- Impute a value for missing Y_j from the predictive distribution, $P(Y_j \mid Y_{obs}) = P(Y_j \mid Y_1, \ldots, Y_{j-1})$. Let $\theta = (\beta_0, \ldots, \beta_{j-1}, \sigma_j)$
- Note that

$$P(Y_j \mid Y_1, \ldots, Y_{j-1}) = \int P(Y_j \mid Y_1, \ldots, Y_{j-1}, \theta) P(\theta \mid Y_1, \ldots, Y_{j-1}) d\theta.$$

- One plan is to draw a value of θ from its posterior distribution $P(\theta \mid Y_1, \ldots, Y_{j-1})$, say θ_*, and then draw a value of missing Y_j from its conditional posterior distribution given the drawn value of θ, $P(Y_j \mid Y_1, \ldots, Y_{j-1}, \theta_*)$.
- Repeat this process m times to create m draws from the joint posterior distribution of Y_j and θ.
- Ignore the drawn values of θ gives m draws from the predictive distribution of Y_j.

Implementation on Regression method MI for monotone missing data

- For a variable Y_j with missing values, we fit a regression model with previous variables, $Y_1, \ldots, Y_{j-1}$, as independent covariates:

$$Y_j = \beta_0 + \beta_1 \, X_1 + \beta_2 \, X_2 + ... + \beta_{j-1} \, Y_{j-1}.$$

  using observations with observed values for variables $_1, \ldots, Y_j$.
- The fitted model includes the regression parameter estimates

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_{j-1})$$

  and the associated covariance matrix $\hat{\sigma}_j^2 V_j$,where V_j is the usual X'X inverse matrix derived from the intercept and variables $Y_1, \ldots, Y_{j-1}$.

Implementation, cont

- For each imputation, new parameters $\beta_* = (\beta_{*0}, \beta_{*1}, ..., \beta_{*(j-1)})$ and σ^2_{*j} are drawn from the posterior predictive distribution of the parameters. That is, they are simulated from $(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_{j-1}), \sigma^2_j$, and V_j. The variance is drawn as

$$\sigma^2_{*j} = \hat{\sigma}^2_j (n_j - j)/g$$

  where g is a $\chi^2_{n_j - j}$ random variate and n_j is the number of nonmissing observations for Y_j. The regression coefficients are drawn as

$$\beta_* = \hat{\beta} + \sigma_{*j} V'_{hj} Z$$

  where V_{hj}' is the upper triangular matrix in the Cholesky decomposition, $V_j = V'_{hj} V_{hj}$, and Z is a vector of j independent random standard normal variates.

Implementation, cont

- The missing values are then replaced by

$$\beta_{*0} + \beta_{*1}\, y_1 + \beta_{*2}\, y_2 + \ldots + \beta_{*(j-1)}\, y_{j-1} + z_i\, \sigma_{*j}$$

  where $y_1, \ldots, y_{j-1}$ are the covariate values of the first $j-1$ variables and z_i is a simulated standard normal deviate.
- The process is repeated sequentially for variables with missing values.

SAS code

```
proc mi data=MonotoneData seed=501213;
   class female;
   monotone reg (mh1 mh2 mh3 mh4/details);
   var  female age mh1 mh2 mh3 mh4 ;
   run;
```

SAS Output

Missing Data Patterns

Group	FEMALE	AGE	mh1	mh2	mh3	mh4	Freq	Percent
1	X	X	X	X	X	X	759	86.25
2	X	X	X	X	X	.	92	10.45
3	X	X	X	X	.	.	27	3.07
4	X	X	X	.	.	.	2	0.23

Regression Models for Monotone Method

Imputed Variable	Effect	FEMALE	Obs-Data
mh2	FEMALE	Female	-0.01240
mh2	AGE		-0.00792
mh2	mh1		0.44922

SAS Output, cont

```
                  Regression Models for Monotone Method

Imputed                      ----------------Imputation-----------
Variable Effect     1          2          3           4          5

mh2     Intercept 0.070107   0.024755  0.068154  -0.075028 -0.016925
mh2     FEMALE    -0.085586  -0.041030 -0.015157 0.041547  0.018400
mh2     AGE       0.005786   -0.080487 -0.057540 -0.044029 -0.021239
mh2     mh1       0.487531   0.425315  0.432769  0.459250  0.412827
```

SAS Output, cont

```
Imputed     --------------------------------Imputation---------
 Variable   Effect   1          2           3          4       5

mh3 Intercept -0.004942  -0.001726  0.079478   -0.056868  0.029256
mh3  FEMALE   -0.071074  -0.028655  -0.003361  -0.046291 -0.050447
mh3  AGE      -0.052435  -0.042099  -0.016780   0.040931  0.024660
mh3  mh1       0.209715   0.174411  0.121701    0.165569  0.160420
mh3  mh2       0.419566   0.461516  0.466346    0.426716  0.478057
```

SAS output, cont

Multiple Imputation Variance Information

Variable	Between	Within	Total	DF
	-----------------Variance-----------------			
mh2	0.000011092	0.010507	0.010520	875.59
mh3	0.000096510	0.012258	0.012374	852.58
mh4	0.000738	0.013334	0.014219	457.69

Multiple Imputation Variance Information

Variable	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
mh2	0.001267	0.001266	0.999747
mh3	0.009448	0.009403	0.998123
mh4	0.066389	0.064068	0.987349

SAS output, cont

Multiple Imputation Parameter Estimates

Variable	Mean	Std Error	95% Confidence Limits		DF
mh2	10.503899	0.102568	10.30259	10.70521	875.59
mh3	10.921351	0.111239	10.70302	11.13969	852.58
mh4	11.443991	0.119244	11.20966	11.67832	457.69

Predictive mean matching method for monotone missing data

- Predictive mean matching is similar to the regression method except that it imputes each missing value from a set of observed values whose predicted values are closest to the predicted value for the missing value from the simulated regression model.
- For a missing value of variable Y_j, we follow the same procedure as in the regression method to obtain new parameters $\beta_* = (\beta_{*0}, \beta_{*1}, ..., \beta_{*(j-1)})$ and σ^2_{*j}.
- Compute predicted value of Y_j for individual i as

$$y_{i*} = \beta_{*0} + \beta_{*1}\, y_{i1} + \beta_{*2}\, y_{i2} + ... + \beta_{*(j-1)}\, y_{i(j-1)}\, \sigma_{*j}$$

where $y_{i1}, \ldots, y_{i(j-1)}$ are the covariate values of the first $j-1$ variables for individual i.

Predictive mean matching method, cont

- Choose a set of j_0 observations with observed Y_j whose corresponding predicted values are closest to y_{j*}.
- Impute the missing value of Y_j by a value randomly drawn from these j_0 observed values.
- The process is repeated sequentially for variables with missing values.
- Predicted mean matching method ensures that imputed values are plausible and may be more appropriate than the regression method if the normality does not hold.

SAS code

```
proc mi data=MonotoneData seed=501213;
 class female;
 monotone regpmm(mh4=female age female*age mh1 mh2 mh3 mh3*mh3/details);
 var female age mh1 mh2 mh3 mh4;
 run;
```

SAS Output

Regression Models for Monotone Predicted Mean Matching Method

Imputed Variable	Effect	Imputation 1	2	3	4	5
mh4	FEMALE	-0.023746	0.091637	0.128823	0.115147	0.071817
mh4	AGE	-0.051876	-0.129313	-0.079178	-0.085499	-0.052836
mh4	AGE*FEMALE	0.051017	0.034009	-0.042764	-0.011488	-0.003801
mh4	mh1	0.051014	0.109213	0.064231	0.045100	0.033609
mh4	mh2	0.224346	0.107300	0.211499	0.236282	0.198416
mh4	mh3	0.304298	0.345190	0.381311	0.303170	0.353700
mh4	mh3*mh3	-0.004622	0.069952	-0.028568	-0.018023	0.036400

SAS Output, Cont

```
  Multiple Imputation Variance Information
-----------------Variance------------------
Variable          Between          Within      Total       DF

mh4              0.001213        0.013640     0.015096   275.41

Multiple Imputation Variance Information

                Relative        Fraction
                Increase        Missing        Relative
 Variable      in Variance    Information    Efficiency

 mh4              0.106750        0.100628     0.980271
```

SAS Output, Cont

```
  Multiple Imputation Parameter Estimates

 Variable     Mean    Std Error   95% Confidence Limits   DF

 mh4        11.466582  0.122864  11.22471   11.70845  275.41

Multiple Imputation Parameter Estimates
                                                    t for H0:
Variable   Minimum    Maximum      Mu0    Mean=Mu0   Pr > |t|

 mh4 11.417578   11.515219     0       93.33     <.0001
```

Propensity score method for monotone missing data

- The propensity score method uses the following steps to impute values for each variable Y_j with missing values:
- Create an indicator variable R_j with the value 0 for observations with missing Y_j and 1 otherwise.
- Fit a logistic regression model

$$\mathrm{logit}(p_j) = \beta_0 + \beta_1\, Y_1 + \beta_2\, Y_2 + ... + \beta_{j-1}\, Y_{j-1}$$

  where $p_j = Pr(R_j = 0|Y_1, \ldots, Y_{j-1})$ and $logit(p) = log(p/(1-p))$.
- Create a propensity score for each observation to estimate the probability that it is missing.
- Divide the observations into a fixed number of groups (typically assumed to be five) based on these propensity scores.

Propensity score method, cont

- Apply an approximate Bayesian bootstrap imputation to each group. In group k, suppose that Y_{obs} denotes the n_1 observations with nonmissing Y_j values and Y_{mis} denotes the n_0 observations with missing Y_j. The approximate Bayesian bootstrap imputation first draws n_1 observations randomly with replacement from Y_{obs} to create a new data set $Y_{obs}*$. The process then draws the n_0 values for Y_{mis} randomly with replacement from $Y_{obs}*$.
- Steps 1 through 5 are repeated sequentially for each variable with missing values.
- The goal of the propensity score method was to impute the missing values on the response variables. The method uses only the covariate information that is associated with whether the imputed variable values are missing. It does not use correlations among variables. It is effective for inferences about the distributions of individual imputed variables, such as an univariate analysis, but it is not appropriate for analyses involving relationship among variables, such as a regression analysis. It can also produce badly biased estimates of regression coefficients when data on predictor variables are missing (Allison 2000).

SAS code

```
proc mi data=MonotoneData seed=501213;
   monotone propensity  (mh2 mh3 mh4/details);
    var  mh1 mh2 mh3 mh4;
    run;
```

SAS Output

```
Monotone Model Specification
  Imputed
  Method                        Variables

  Propensity( Groups= 5)        mh2 mh3 mh4

  Missing Data Patterns

Group  mh1    mh2    mh3    mh4          Freq      Percent
  1    X      X      X      X             759        86.25
  2    X      X      X      .              92        10.45
  3    X      X      .      .              27         3.07
  4    X      .      .      .               2         0.23
```

SAS Output, cont

```
                  The MI Procedure
   Logistic Models for Monotone Propensity Scores Method
Imputed       ---------------------------Imputation-------------
Variable Effect    1           2          3         4          5
mh3      Intercept -3.51882 -3.51201 -3.50779 -3.50700 -3.50916
mh3      mh1       -0.46665 -0.49263 -0.52075 -0.55083 -0.50884
mh3      mh2       -0.17100 -0.11080 -0.04464 0.02610  -0.07257


   Logistic Models for Monotone Propensity Scores Method

Imputed     -------------------------------Imputation--------
Variable Effect   1        2          3         4            5

mh4   Intercept-1.85764 -1.85422  -1.85924  -1.86018  -1.85019
mh4   mh1       -0.00275 -0.01209  -0.00534  -0.00968  -0.02738
mh4   mh2       -0.12763 -0.12103  -0.07728  -0.05221  -0.13160
mh4   mh3      -0.14692  -0.12658  -0.19541  -0.21437 -0.07963
```

SAS Output, cont

Multiple Imputation Variance Information

	------------------Variance------------------			
Variable	Between	Within	Total	DF
mh2	0.000034433	0.010528	0.010569	870.67
mh3	0.000237	0.012206	0.012490	771.54
mh4	0.001831	0.013573	0.015770	161.89

Multiple Imputation Variance Information

Variable	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
mh2	0.003925	0.003917	0.999217
mh3	0.023274	0.022997	0.995422
mh4	0.161863	0.147546	0.971337

SAS Output, cont

Multiple Imputation Parameter Estimates

Variable	Mean	Std Error	95% Confidence Limits		DF
mh2	10.502560	0.102806	10.30078	10.70434	870.67
mh3	10.933486	0.111761	10.71410	11.15288	771.54
mh4	11.481636	0.125577	11.23365	11.72962	161.89

Multiple Imputation Parameter Estimates

Variable	Minimum	Maximum	Mu0	t for H0: Mean=Mu0	Pr > |t|
mh2	10.494955	10.510718	0	102.16	<.0001
mh3	10.917712	10.956439	0	97.83	<.0001
mh4	11.429372	11.548524	0	91.43	<.0001

Monte Carlo Markov Chain (MCMC) method

- MCMC methods are used to generate pseudo-random draws from multidimensional and otherwise intractable probability distributions via Markov chains. A Markov chain is a sequence of random variables in which the distribution of each element depends only on the value of the previous one.
- In MCMC simulation, one constructs a Markov chain long enough for the distribution of the elements to stabilize to a stationary distribution, which is the distribution of interest. By repeatedly simulating steps of the chain, the method simulates draws from the distribution of interest.

MCMC Method in Bayesian inference

- MCMC has been applied as a method for exploring posterior distributions, $p(\theta \mid y)$, in Bayesian inference. That is, a Markov chain in θ with ergodic distribution $p(\theta \mid y)$ is set up.
- Gibbs sampler. Let $\theta = (\theta_1, \ldots \theta_p)$ denote the parameter vector. The Gibbs sampler is obtained by iteratively, for $j = 1, \ldots, p$, generating from the conditional posterior distributions

  $$\theta_j^{(t+1)} \sim p(\theta_j \mid \theta_1^{(t+1)}, \ldots, \theta_{j-1}^{(t+1)}, \theta_{j+1}^{(t)}, \ldots, \theta_p^{(t)}).$$

  If practicable it is advisable to generate from higher dimensional conditionals.
- The Gibbs sampler is most useful when the complete conditional posterior distributions, $p(\theta_j \mid \theta_i, i \neq j, y)$, take the form of some well-known distributions, allowing random variate generation.
- Data augmentation algorithm is a special type of Gibbs when $p = 2$.

Metropolis-Hastings algorithm

- For many important statistical applications, the complete conditional posterior distributions may not have well-known distributions.
- Other alternative Markov chains are needed. One of them is Metropolis-Hastings algorithm.

Imputation methods for discrete variables in SAS

- Under monotone missing data pattern, SAS implemented two MI procedures.
- Logistic regression for ordinal data
- Discriminant function method for nominal data

SAS code

```
proc mi data=exam3 out=outmi seed=501213;
 class npcerad ;
monotone discrim (npcerad=mmselast npgender educ npdage/details);
 var mmselast npgender educ npdage npcerad ;
 run;
```