Paper AS05

Beyond missing data patterns: augmenting PROC MI tables with summary of imputed data

Lamine Signate, Covance, Leiden, The Netherlands
Domenico Magazzù, Covance, Milan, Italy
Tommi Nurminen, Tale Ltd, Jyväskylä, Finland

ABSTRACT

SAS® procedure MI (PMI) can be used for summarising the frequencies of each unique missing data pattern (MDP) in the dataset along with the mean values of the analysis variables within each pattern. When an MDP summary is needed the option (nimpute=0) can be specified; however, when multiple imputations on MD are performed by PMI it may be informative to describe the imputed data in the context of data actually observed for each MDP. In this work, we present a sequence of programming steps for augmenting the MDP summary from PMI with mean values of imputed data. In the first step, MD are imputed with PMI according to a specified imputation model. Secondly, subjects are categorised by MDP in a data step; the accuracy of this 'manual' categorization is cross-checked against the MDP summary from PMI. Thirdly, mean values of completed data are calculated for each missing pattern.

INTRODUCTION

Longitudinal studies give rise to missing data due to dropouts and a range of other reasons. Some observed data may be censored due to e.g. use of rescue therapy or treatment switches.

To overcome the statistical issues inherent to the amount of missing data, appropriate statistical analysis and inference may require that missing data are imputed using some imputation model.

In multiple imputation (Little & Rubin, 2002) each missing value is replaced with a set of plausible values that reflect the uncertainty about the true value.

SAS® procedure PROC MI can be used for summarising the frequencies of each unique missing data pattern (MDP) in the dataset along with the mean values of the analysis variables within each pattern.

Most of the time author's present or discuss methodology of handling missing data but the actually imputed data in each pattern are not described. We propose to add the summary statistics (means) of the imputed data in the MDP summary from PROC MI and present the programming steps for achieving this.

ALGORITHM AND PROGRAMMING STEPS

CASE EXAMPLE

The dataset to exemplify our approach is from a non-interventional, single-arm study assessing the effectiveness of a specific pharmaceutical intervention in patients with axial spondyloarthritis in daily practice.

- Primary endpoint was the Change from Baseline to Week 52 in BASDAI (Bath Ankylosing Spondylitis Disease Activity Index).
- BASDAI was assessed at 4 time points: Baseline, week 12, week 24 and week 52.
- For the purposes of the current work, data were assumed to be missing at random (MAR)
- The MI imputation was done in 2 parts: intermittent missing data imputed using MCMC and monotone missing data imputed using regression model. Further details of the imputation model will not be reported as out of the scope for the current work.

SUMMARY OF MDP FROM PROC MI

The MDP for the 4 time points together with the mean result of the data observed in each MDP can be easily obtained by the following statements, after having transposed the data in order to have the BASDAI value at each visit as separate columns.

```
ods output misspattern=missp;
proc mi nimpute=0 data=miss_data;
    var week0 week12 week24 week52;
run;
```

SAS Code Fragment 1: Summarising missing data patterns.

The resulting output would be as follows, with additional little work needed to add 'O' to distinguish between intermittent (.) and monotonic missing (O) missing data and 'NC' to indicate not calculable statistics due to missing data. 'X' represents available data.

		Observed Missing Data pattern				BASDAI Mean			
Pattern	Subjects with Missing Data Pattern n (%)	Baseline	Week 12	Week 24	Week 52	Baseline	Week 12	Week 24	Week 52
1	413 (73.2)	X	X	X	X	6.0	3.5	3.3	3.1
2	62 (11.0)	X	X	X	O	6.5	5.1	5.5	NC
3	60 (10.6)	X	X	O	O	6.1	5.1	NC	NC
4	17 (3.0)	X	X	.	X	6.7	4.1	NC	4.4
5	7 (1.2)	X	.	X	X	6.2	NC	4.0	3.6
6	3 (0.5)	X	.	X	O	5.5	NC	5.7	NC
7	2 (0.4)	X	.	.	X	8.8	NC	NC	4.6

Table 1: Missing data patterns and observed means in the example dataset.

AUGMENTING PROC MI WITH SUMMARY OF IMPUTED DATA

The following steps will fill in the 'NC' cells of Table 1 by the means of multiple imputed data.

1. Derive the missing data pattern (MDP) in a data step, taking as the reference Table 1, the output from PROC MI, to make sure the derivations are done correctly.

```
data miss_pattern;
    set miss_data;
    array missx (*) $ week0_miss week12_miss week24_miss week52_miss;
    array obsx (*) $ week0 week12 week24 week52;
    do i=1 to dim(missx);
           if missing(obsx[i]) then missx[i]='M'; /* missing values are noted by M */
           else missx[i]='X';
    end;
/* MDP derivation */
    miss_pat=cats (week0_miss, week12_miss, week24_miss, week52_miss);
run;
```

SAS Code Fragment 2: Manual derivation of missing data patterns.

2. Impute data according to any specified imputation model. Here we used PROC MI to do multiple imputation but note that this step can be extended to any imputation method including single-value imputation, e.g. LOCF or WOCF, as long as the resulting dataset contains complete data. Use *miss_pattern* as input dataset and create dataset *miout1* that will contain complete data and the MDP (*miss_pat*) derived above.

   Figure 1 includes a sample of *miout1:* 10 pairs of imputed values at Weeks 24 and 52 for a subject for whom these were missing. Week 0 and 12 did not have missing values, therefore the same observed value is repeated across imputations, while week 24 and 52 show different imputed values.

Imputation Number	WEEK0	WEEK12	WEEK24	WEEK52	WEEK0_MISS	WEEK12_MISS	WEEK24_MISS	WEEK52_MISS	MISS_PAT
1	5.8	2.4	1.2534	2.0428	X	X	M	M	XXMM
2	5.8	2.4	1.5803	3.1348	X	X	M	M	XXMM
3	5.8	2.4	5.0146	1.3532	X	X	M	M	XXMM
4	5.8	2.4	2.2591	3.3814	X	X	M	M	XXMM
5	5.8	2.4	5.2887	3.4578	X	X	M	M	XXMM
6	5.8	2.4	2.8209	2.9045	X	X	M	M	XXMM
7	5.8	2.4	3.3606	2.9605	X	X	M	M	XXMM
8	5.8	2.4	1.6949	1.721	X	X	M	M	XXMM
9	5.8	2.4	0.7946	0.0698	X	X	M	M	XXMM
10	5.8	2.4	0.9307	1.0539	X	X	M	M	XXMM

Figure 1: Miout1 dataset: a sample including a subject with missing values at Week 24 and 52.

3. Compute the mean of completed data (*miout1*) by miss_pat (output dataset=*means*) in order to replicate the results already provided by PROC MI (the numeric results on the right of Table 1) but also to summarize the imputed data (for the cells previously shown in Table 1 as 'NC').

```
proc sql;
   create table means
   as select     miss_pat,
                 mean(week0) as mean_week0,
                 mean(week12) as mean_week12,
                 mean(week24) as mean_week24,
                 mean(week52) as mean_week52
   from miout1
   group by miss_pat;
quit;
```

SAS Code Fragment 3: SQL step calculating the mean of BASDAI by missing data pattern.

4. Compute the frequencies (n) and percent (%) by miss_pat using PROC SQL and store them in *freq* dataset. This replicates the statistics already provided by PROC MI and shown on the 2nd column of Table 1, but allows to keep information on each subject's MDP. (Subject-level information on MDP cannot be retrieved from PROC MI but is needed for subsequent data merging steps.) The corresponding n and % from PROC MI, by MDP, can be used for checking the accuracy of the data steps derivations).

This can be done by an SQL step directly from miout1 on _imputation_=1 and using 'distinct' clause.

```
proc sql;
   create table cnt1
   as select miss_pat, count(distinct usubjid) as count, joinkey=1
   from miout1 where _imputation_=1 group by miss_pat;

   create table bign
   as select count (distinct usubjid) as bign, joinkey=1
   from miout1 where _imputation_=1;

   create table freq as select a.*, b.bign ,count/bign *100 as percent
   from cnt1 a
   left join bign b on a.joinkey=b.joinkey;
quit;
```

SAS Code Fragment 4: SQL step calculating the n and % of subjects in each missing data pattern.

5. Create a shell keeping one occurrence of each MDP (column *miss_pat*) and the indicators of whether each analysis variable was missing or observed for the particular MDP.

```
proc sort nodupkey data=miout1
    out=shell (keep= week0_miss week12_miss week24_miss week52_miss);
    by miss_pat;
run;
```

SAS Code Fragment 5: Creating a shell for a data summary table by missing data patterns.

6. Merge all data: n, %, means of observed and imputed data by each MDP. The means of imputed data are annotated with square brackets; this requires that numeric means are rounded and converted to characters. (The additional SAS code to distinguish between intermittent and monotone missing data is not shown).

```
data final;
    merge means freq shell;
    by miss_pat;
    length weekc0 weekc12 weekc24 weekc52 $30;
    array missx (*) $ week0_miss  week12_miss  week24_miss  week52_miss;
    array visx (*) $ mean_week0 mean_week12 mean_week24 mean_week52;
    array vischarx (*) $ weekc0 weekc12 weekc24 weekc52;
    do i=1 to dim(missx);
           if missx[i] ne 'M' then vischarx[i]=put(round(visx[i],0.1),4.1);
           else vischarx[i]='[' || put(round(visx[i],0.1),4.1) || ']';
    end;
run;
```

SAS Code Fragment 6: Merging n, %, and means of observed and imputed data and annotating the imputed cells.

8. Report results by decreasing frequencies of MDP

		Observed Missing Data Pattern				BASDAI Mean over Imputed Data			
Pattern	Subjects with Missing Data Pattern n (%)	Baseline	Week 12	Week 24	Week 52	Baseline	Week 12	Week 24	Week 52
1	413 (73.2)	X	X	X	X	6.0	3.5	3.3	3.1
2	62 (11.0)	X	X	X	O	6.5	5.1	5.5	[4.7]
3	60 (10.6)	X	X	O	O	6.1	5.1	[4.3]	[4.1]
4	17 (3.0)	X	X	.	X	6.7	4.1	[4.2]	4.4
5	7 (1.2)	X	.	X	X	6.2	[4.1]	4.0	3.6
6	3 (0.5)	X	.	X	O	5.5	[4.8]	5.7	[5.0]
7	2 (0.4)	X	.	.	X	8.8	[5.6]	[4.9]	4.6

Table 2: Missing data patterns with means of observed and imputed data.

CONCLUSION

Presenting a summary including the means for the observed data and the imputed data by MDP may help to evaluate the appropriateness of imputed data and thus the overall conclusions.

The output makes it easy to compare the behaviour of model-based imputations with some old-school methods such as LOCF.

We presented a sequence of programming steps to augment the MDP summary from PROC MI with summaries of imputed data. Some of the programming steps could be eliminated if the pattern number related to each study subject could be retrieved directly from PROC MI.

REFERENCES

1- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (Second ed.). London: Wiley
2- European Medicines Agency (2010). Guideline on Missing Data in Confirmatory Clinical Trials.
3- SAS Institute Inc. (2011). SAS/STAT® 9.3 User's Guide. Cary, NC: SAS Institute Inc.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Lamine Signate, Senior Statistical Programmer
Covance International Holdings B.V.
Schipholweg 111, 2316 XC Leiden, The Netherlands
Work Phone: +31 71 711 25 10
Fax: +31 71 891 00 25
Email: Lamine.signate@covance.com
Web: Covance.com

Domenico Magazzu, Senior Statistical Programmer
Covance Clinical and Periapproval Services Ltd Italian Branch
Via Montecuccoli 20/1, 20147 Milan, Italy
Work Phone: +39 02 8978941
Fax: +39 02 897894217
Email: domenico.magazzu@covance.com
Web: Covance.com

Tommi Nurminen, Biostatistician
Tale Ltd, Metsontie 7, FI-40400 Jyväskylä
Work Phone: +358 405 781 958
Email: tommi@tomminurminen.com