Studying Missing Data Patterns Using a SAS Macro

Theresa Schwartz, Division of Biostatistics, New York State Psychiatric Institute, New York, NY
Qixuan Chen, Department of Biostatistics, Columbia University, New York, NY
Naihua Duan, Division of Biostatistics, New York State Psychiatric Institute/Department of Biostatistics, Columbia University, New York, NY

Abstract/Introduction

Before diving into the analysis phase of any study, it is essential to examine the data thoroughly. Performing univariate analyses is a good first step, noting not only what data are observed, but just as important, what data are missing. This is especially important when working with complex survey data involving skip patterns. It is useful to check missing data patterns to examine whether the survey was administered and entered correctly. Missing data patterns can also assist in variable selection for statistical analyses as well as missing-data imputation.

We present here a SAS 9.2 macro, %missingPattern, that makes identifying missing data patterns fast and easy. The macro is designed to look at missing data in four ways: the proportion of subjects with each pattern of missing data, the number and percentage of missing data for each individual variable, the concordance of missingness in any pair of variables, and possible unit nonresponse. The user can customize these analyses by specifying which variables to include or exclude, and which output should be produced. We illustrate the use of our macro with a simple hypothetical example.

Up to 4 Patterns in Just 3 Easy Steps!

```
%macro missingPattern(datain=, varlist=, exclude=, missPattern1=,
dataout1=, missPattern2=, dataout2=, missPattern3=, dataout3=,
missPattern4=, dataout4=);
```

When calling the %missingPattern macro a minimum of three parameters must be specified: (1) input data set, (2) type(s) of missing pattern analysis to be produced, and (3) name(s) of output data set for each requested missing pattern analysis. The user can choose to output any one, any combination, or all four missing pattern analysis output data sets. The user can also subset the data being analyzed for missing patterns by excluding or including a list of variable names.

We outline a 3 step process for producing the desired output:

1. Specify the input data set in “datain=”
2. A subset of variables can be run through the macro:
    - To include a list of specific variables use “varlist=”.

- To exclude a list of variables use “varlist=” and specify “exclude=‘TRUE’”
- Otherwise leave “varlist=” blank for the default analysis on all of the variables in the data.

3. The macro creates up to 4 SAS output data sets. Each data set evaluates missing data in a different way.
    - Request a data set to be created by specifying “misspattern#=” to be ‘TRUE’.
    - Specify name of the output data set with “dataout#=”.

Macro Output: 4 Diagnostic Data Sets Explained Using a Hypothetical Example

The use of this SAS macro is illustrated with a hypothetical data set shown in Table 1. There are in total 11 observations and 6 variables in the example data. The first four variables are numeric variables and the last two are character variables. Missing data are present in variables “aa”, “bb”, “dd”, and “ff”, denoted as “.” for numeric variables, and as blank space for character variables

Table 1. Example Data

Obs	aa	bb	cc	dd	ee	ff
1	3	5	7	6	a	b
2	2	3	4	3	b	
3	4	5	6	8	r	t
4	2	4	6	.	d	y
5	3	.	4	.	e	j
6	2	.	4	.	f	
7	8	1	9	7	a	
8	.	2	4	6	b	
9	.	.	3	.	d	
10	.	.	2	.	n	u
11	.	.	10	.	k	

The first missing data pattern (*misspattern1=‘TRUE’, dataout1=DATA1)* provides the summary matrix summarized by the pattern of missing data. (An example is shown in Table 2.) A missing data indicator is created for each variable, with 1 for missing and 0 for observed; the indicator variable is named with a prefix “m_XXXXX”, where “XXXXX” denotes the name of the original variable. Each row of the output represents a unique pattern of these indicator variables, along with the number and proportion of subjects that exhibit such pattern. This output data set is useful in determining how much of the data are complete (how many subjects are missing no variables), how much are

incomplete or possibly unusable (how many subjects are missing a majority of variables, or all variables). Any patterns with a high proportion of subjects should be inspected further to determine if data entry possibly went awry, or if they are evidence of skip patterns. Table 2 displays DATA1 - the output for this analysis from our example data. There are 8 different missing data patterns in this data set. Five of the patterns are unique to one individual with the first, second and last pattern are observed in two individuals or 18.2% of the data. Since Table 2 is a SAS data set, it can be sorted in ascending or descending sequence by missPattern_prop, or NObs.

Table 2: Proportion of Subjects with Each Missing Data Pattern

Obs	m_aa	m_bb	m_cc	m_dd	m_ee	m_ff	NObs	missPattern_prop
1	0	0	0	0	0	0	2	18.1818
2	0	0	0	0	0	1	2	18.1818
3	0	0	0	1	0	0	1	9.0909
4	0	1	0	1	0	0	1	9.0909
5	0	1	0	1	0	1	1	9.0909
6	1	0	0	0	0	1	1	9.0909
7	1	1	0	1	0	0	1	9.0909
8	1	1	0	1	0	1	2	18.1818

The second missing data pattern (*misspattern2='TRUE', dataout2=DATA2)* provides the number and percent of missing data in each variable. When concerned about the completeness of an individual variable, this output data set can be useful. Table 3 displays DATA2 - the output for this analysis from our example data. In Table 3, we see that 36.4% of the sample subjects are missing variable "aa", 45.5% are missing variable "bb", none of the subjects is missing variable "cc", 54.5% are missing "dd", none is missing "ee" and 54.5% are missing "ff". If this were survey data it may be alarming to see about half of the respondents missing any single variable and this would need to be looked at further to check for data entry error, or evidence of skip logic. When there are a large number of variables, this SAS data set can be sorted by prop_miss to more easily identify the variables with high proportion of missing.

Table 3. Proportion of Subjects Missing Each Individual Variable

Obs	var	num_miss	prop_miss
1	aa	4	36.3636
2	bb	5	45.4545
3	cc	0	0.0000

Obs	var	num_miss	prop_miss
4	dd	6	54.5454
5	ee	0	0
6	ff	6	54.5455

The third missing data pattern *(misspattern3='TRUE', dataout3=DATA3)* provides the pair-wise concordance between any two variables in the input data set or from a variable list specified in the "varlist=" and "exclude=" options. All of these variables are compared in a pair-wise fashion to determine concordance of missingness. When two variables record the same thing, such as race/ethnicity, this output gives the user a comparison of missingness that is useful in deciding between the variables. Perfect concordance in two variables would be evident if the last column (prop_concordance) in the output data set was 100. Five percentages are also output: (1) P00: % of subjects with both var1 and var2 observed; (2) P01; % of subjects with var1 observed but var2 missing; (3) P10: % of subjects with var1 missing but var2 observed; and (4) P11: % of subjects with both var1 and var2 missing. The summary measure prop_concordance (= P00 + P11) presents the % of data that var1 and var2 are missing or observed together. Table 4 is the output table for DATA3, produced by the third type of missing pattern analysis from our example.

Table 4. Pairwise Missingness

Obs	var1	var2	P00	P01	P10	P11	prop_concordance
1	cc	ee	100	0	0	0	100
2	bb	dd	45.455	9.0909	0	45.4545	90.909
3	aa	bb	45.455	18.1818	9.0909	27.2727	72.727
4	aa	dd	36.364	27.2727	9.0909	27.2727	63.636
5	aa	ff	36.364	27.2727	9.0909	27.2727	63.636
6	aa	cc	63.636	0	36.3636	0	63.636
7	aa	ee	63.636	0	36.3636	0	63.636
8	bb	cc	54.545	0	45.4545	0	54.545
9	bb	ee	54.545	0	45.4545	0	54.545
10	bb	ff	27.273	27.2727	18.1818	27.2727	54.545
11	cc	dd	45.455	54.5455	0	0	45.455
12	cc	ff	45.455	54.5455	0	0	45.455
13	dd	ee	45.455	0	54.5455	0	45.455
14	dd	ff	18.182	27.2727	27.2727	27.2727	45.455
15	ee	ff	45.455	54.545	0	0	45.455

The fourth missing data pattern *("misspattern4='TRUE', dataout4=DATA4)* checks the data for unit nonresponse. When a questionnaire is administered and a subset of sampled individuals do not complete the questionnaire, the only observed information for these individuals is the survey design variables measured for everyone irrespective of their response status. This type of missing data is called unit nonresponse in surveys. The macro checks whether the most extreme missing pattern (smallest number of variables observed) matches the theoretical pattern for unit nonresponse – namely, for this missing data pattern, all variables observed are also observed in all other missing data patterns. If such a candidate missing data pattern is found, DATA4 is outputted with the candidate unit nonresponse, for the investigators to ascertain whether all variables observed in this missing data pattern are indeed design variables instead of survey response variables. Our data exhibits possible unit nonresponse as seen in example data set shown in Table 5.

Please note that a specific survey dataset might or might not include a candidate unit nonresponse. If no candidate unit nonresponse is found, the macro outputs a statement "There are no unit nonresponses!" to the log window and DATA4 is not created.

Table 5. Possible Unit Nonresponse

Obs	m_aa	m_bb	m_cc	m_dd	m_ee	m_ff	NObs	num_Var_missing
1	1	1	0	1	0	1	2	4

Conclusion

The %missingPattern macro presented here can be used as an instrumental first step in checking the integrity of a data set. Instead of producing pages and pages of frequencies and means for individual variables (which will be difficult to review and interpret), this macro can be used (within seconds) to validate data entry, check completion of individual surveys or variables, as well as check for pair-wise variable concordance and unit non-response. The user can define a subset of variables to include or eliminate from this analysis to further look at skip patterns in scales, or patterns within sections of a survey. We have shown here the usefulness of the %missingPattern macro using a simple hypothetical dataset, but this can be used on any data set.

Acknowledgement

The authors thank Professor Trivellore E. Raghunathan (University of Michigan) and Professor Alan M. Zaslavsky (Harvard University) for their helpful discussion in the development of this SAS macro. This work is supported by 1U01 MH087981 – 01 from National Institutes of Health.

To download the %missingPattern macro please visit http://www.columbia.edu/~qc2138/