Paper 88

Mini Macros for Major Management of Longitudinal Data

Philip J. d'Almada, EDS, Atlanta, Georgia

ABSTRACT

A SAS® program running on Windows NT and utilizing SAS macroprocessing was designed and developed by the author in order to present the data managers of a longitudinal database with a tool to monitor the monthly additions to this longitudinal database. This program was finally developed in its current version when data inconsistencies, that were discerned in the database, prompted a structured and consistent yet analytic perusal of the database for such crucial inconsistencies.

The program was designed with four macros that each respectively (i) creates the data tracking variables, (ii) prints the tracking data, (iii) completes a crosscheck of chronologically successive data sets, and (iv) invokes the previously mentioned three macros for each of the two types of SAS data sets that would be monitored. This is accomplished by the end-user keying the month-order, the three-letter abbreviation for the month, and the year of interest for both the two months of interest.

This paper was designed to address reasonably recent SAS macro users.

INTRODUCTION

A public health, longitudinal data base (PHLD) is complemented monthly with data submitted to the receiving location. These data are either new records or old records that are editted in some fashion. The incoming data are PRODAS files and the output, analytic database is in the form of two SAS data sets. The final SAS data sets are then made available, monthly, to the end-user location. Until this series of macros were developed there was no consistent assessment of the data quality perhaps for justifiable reasons when the system was originally set up. During data preparation for analysis, inconsistencies were discerned in the database by the author and this prompted a structured and consistent yet analytic perusal or mining of the database for such crucial inconsistencies. The author developed the structure for analysis and designed, developed and implemented the software that isolated these inconsistencies. Consequently, a SAS program utilizing SAS macroprocessing was designed and developed by the author in order to present the data managers or end-users of the PHLD with a tool to monitor the monthly additions to this data base.

METHOD

The program was designed with four macros that each respectively (i) creates the data tracking variables, (ii) prints the tracking data, (iii) completes a crosscheck of chronologically successive data sets, and (iv) invokes the previously mentioned three macros for each of the two types of SAS data sets that would be monitored, that is, the "exact" and the "merged" data sets, hereafter identified as the "EX" and "MR" data sets, respectively. This is accomplished by the end-user keying the month-order, the three-letter abbreviation for month, and the year of interest for both the two months of interest.

In the first macro, "DATTRK" , three variables are created as follows:

- BD_X identifies a record as missing a birth date (1), or not (0),
- NO_0 identifies a record from a followup form as having no associated initial form (1), or not (2),
- FPAT identifies a record as being the first record for that person (1), or not (2).

In addition, a FORMAT is placed on the variable, FORM, to identify records derived from initial (i) or followup (f) forms. The variables, STATUS and OK, are also included in the tracking process. The variable, STATUS, indicates Active records (A) and records for which a Required field is missing (R), and these types of records should be the only types included in the "combined" data set, that is, the PRODAS precursor to the EX and MR SAS data sets. The variable, OK, in the EX data set, indicates whether the record exists exclusively in the EX data set (0) or has a match in the MR data set (1). Tracking is implemented by generating a frequency distribution on these six variables for both the EX and the MR data sets. In addition, and conditional only upon the EX data set, a distribution of sites is generated for those records that have no match in the MR data set and are derived from followup forms with no associated initial form, thus causing these records of this type to be the first record for that person. A second distribution of sites is generated for the remaining

records derived from followup forms. Thus, the total number of followup forms by site that have no associated initial forms is given by summing across these two distributions by site, and the number of persons per site for whom there is no initial form is given in the first distribution by site. For each month selected, these three distributions from the EX data set are saved in three data sets, and the first distribution from the MR data set is saved in one data set.

In the second macro, “PRTPRG”, the data from the three distributions mentioned above and developed from the first user-specified month of the PHLD data are each concatenated with the data from the corresponding distributions developed from the second user-specified month of the PHLD data, then printed for both chronologically successive data sets on separate pages for each type of distribution. The information for each month is identified and set off by a created variable, MONTH, with value the month-order and abbreviated name of the month. The first page of EX or MR distributions is footnoted with a legend explaining the values of the six tracking variables. The second and third pages of EX distributions are footnoted with descriptives that appropriately identify the values of those four of the six tracking variables by which these distributions were developed.

The third macro, “XCHK”, is used in a crosscheck of the successive EX/MR data sets. This is done through a distribution of records that exist in the one data set but not in the other and use of five of the six tracking variables with the variable, OK, being omitted. Thus, two pages per data set type, EX and MR, are produced, and a legend identifying the variable values is footnoted on each page. One distribution of much larger numbers will reflect the addition of new records to the database, and the other very small distribution will reflect attrition in the database.

The fourth macro, “SASTYP”, controls the execution of the former three macros, and the iterative execution for the EX and MR types of data sets. What is actually controlled in the execution of these three macros includes the specification of the string of variables used in the production of the tables, the content of the footnotes, and the text in the titles. In this version of the program, the end-user implements quality management only through specification of parameters for this fourth macro. These parameters indicate any pair of chronologically successive months from the public health, longitudinal database.

CONCLUSION

The worth of SAS macroprocessing is demonstrated in this exercise to produce a simple yet sufficiently sophisticated methodology to monitor a longitudinal database for anomalies that may occur without notice. Three macro programs form the essence of this methodology that provides the end-user with little but sufficient information to assess the condition of the database for the current month. The implementation of this methodology is through a fourth macro allowing for a user-friendly definition of any pair of months for which the databases are to be compared. Other procedures or properties of the SAS system could be invoked, such as PROC REPORT, that would provide a more elaborately designed layout in the output.

ACKNOWLEDGEMENTS

The author appreciates the support of his wife, Shelley, and the perseverance of his children, Seth, Samuel and Abigail. The author also appreciates the support of his employer, EDS, and is grateful to be able to work on the project from which this paper is derived.

CONTACT INFORMATION

The mailing address, work telephone number and e-mail address to reach the author are given below:

EDS, c/o CDC,
1600 Clifton Rd., E-45,
Atlanta, GA 30333.

404-639-6120

pxd2@cdc.gov

[Sugi24-fin1.doc]