The SAS® System in a Longitudinal Environmental Health Intervention Study with Repeated Monitoring

WinKey Pan, Department of Environmental Health
University of Cincinnati

Keywords: Statistical Computing, Software Integration, SAS Macro Programming, Data Management and Processing

ABSTRACT

In this paper, we will focus on how SAS has been applied in managing, analyzing data, and preparing reports for a longitudinal neighborhood-wide environmental health intervention project. Special attention will be given to the following topics.

1. How PC SAS, mainframe SAS, and xBase interact with each other in data management.
2. How SAS Macro Language has been used in repeating similar data analysis and generating stepwise modeling procedures, e.g., stepwise GLM, SYSLIN.
3. How SAS's capability is enhanced by interacting with other statistical packages, e.g., BMDP.
4. How SAS is used as a statistical component in a generalized geographic information system (GIS).
5. How data has been analyzed.
6. How SAS and the personal computer based word processor, spreadsheet, and statistical graphics packages interact with each other in preparing report.
7. How SAS was used in developing new statistical procedures and methodologies.

INTRODUCTION

In this paper, we will present how SAS is integrated with other software in managing, analyzing data, and preparing reports for a three-year neighborhood-wide environmental health intervention study, the Cincinnati Soil Lead Abatement Demonstration Project (Clark 1989). This longitudinal project has ten sample collection phases. Blood, hand wipe, interior dust, exterior dust, and soil samples were collected for their lead results at least once in every six months from 225 children in 146 households. The study subjects, young children less than 4 years old and at least six months old who lived in six study neighborhoods, and their caregivers were interviewed by questionnaires at each clinic visit. Several other samples were also taken during the course of the project, e.g., paint, water, and dust fall, to enhance our understanding of the childhood lead exposure and its effectiveness of soil and dust lead abatement.

In addition, we will add other techniques and procedures developed for different projects as long as they are appropriate.

Special attention will be given to the following topics: 1. prepare and manage data for analysis and reporting; 2. the power of SAS Macro Language in repeated or stepwise data analysis and table preparation; 3. interacting with other statistical packages; 4. an integral part of a generalized geographical information system; 5. the detailed step by step description of our data analysis procedure; 6. report preparation through the integration of word processor, statistics and graphical package, and spreadsheet; 7. new procedure and methodology development.

HOW PC SAS, MAINFRAME SAS, AND XBASE INTERACT WITH EACH OTHER IN DATA PREPARATION AND MANAGEMENT

SAS has been the integral part of our data management. The final version of our data sets were saved and managed under SAS on both PC and mainframe computer. The SAS upload, download, r-submit subroutines
made data sharing between mainframe and PC easy. Our data were first entered into an Apple Macintosh based database, through a data entry system which was developed on FoxBase+/Mac (Pan 1992b and Pan 1993), transferred to PC dBase® data files, and then converted to PC SAS data sets. This FoxBase+/Mac - dBase - SAS configuration made both clinic operation and statistical data analysis easy: (1) data can be entered on site and retrieved as needed by the clinical investigators immediately with minimum training and (2) quality data are available for data analysis after they were entered.

HOW SAS MACRO LANGUAGE HAS BEEN USED IN REPEATING SIMILAR DATA ANALYSIS AND GENERATING STEPWISE DATA MODELING PROCEDURES, E.G., STEPWISE CATMOD, GLM, SYSLIN, AND TABLE GENERATION

In our studies, several data analyses must be performed on different sets of variables repeatedly, e.g., the summary statistics for each variable in each phase. We condensed pages of repeated SAS program statements into less than a page by re-writing them with SAS macro language. SAS macro language was also a timesaver in our data modeling. As an example, for CATMOD, GLM and SYSLIN, we can write a SAS macro as a function to repeatedly analyze data with all possible combinations of the independent variables and their interactions as the parameters to be passed. This was done by a sequence of SAS macro function calls. We reviewed these results at once when output is available.

In environmental health studies, e.g., environmental epidemiology, it is very instrumental to have a table of geometric means of all the exposure biomarkers and the environmental toxicant measurements listed in columns or rows side by side by the different levels of the confounding factors to facilitate identifying the source of exposure as well as the response and extent of exposure. A SAS macro program is developed to accomplish this task.

HOW SAS'S CAPABILITY IS ENHANCED BY INTERACTING WITH OTHER STATISTICAL PACKAGES, E.G., BMDP

Missing values imputation and repeated measures data analysis with continuous and categorical covariates are two important topics which we do not have good and handy solutions from SAS. To meet these needs, our programs called BMDP subroutines from SAS on our IBM mainframe computer, e.g., BMDP AM to do missing values imputation and BMDP 3V, 4V and 5V for repeated measures data analysis. Similar data analysis can be done on PC, but we need to convert data back and forth between PC SAS and PC BMDP.

HOW SAS IS USED AS A STATISTICAL COMPONENT IN A GENERALIZED GEOGRAPHIC INFORMATION SYSTEM (GIS)

In our project, we found that it is very productive to combine a geographical information system (GIS) with statistical packages as a generalized GIS to facilitate our investigation. The following are some of the applications that we have considered (Pan et al. 1991): (a) Identify the spatial distributions of environmental toxicants in the study areas. (b) Investigate the spatial relations between environmental toxicants and human exposure biomarkers. (c) Study the exposure pathways of human subjects to environmental toxicants. (d) Study the migration pathways of environmental toxicants. (e) Propose environmental remediation strategies and plans. (f) Manage and monitor the remediation. (g) Measure the effectiveness of remediation and evaluate different remediation procedures. (h) Study the rate of re-contamination after remediation. (i) Retrieve and present study outcomes, especially as the digitized GIS maps.

HOW DATA HAS BEEN ANALYZED

A data analysis protocol was developed to analyze our data in which all steps and programs involved in data analysis must be documented. (a) Data Clean-up: All data in a data file has to be converted to the correct
data type and format before data analysis was started. All data will be visually checked after the conversion. Then, SAS DATA STEP is used to check if the data linkage identifiers are correctly coded or not. This is to ensure that all data files can be merged together for later analysis. (b) Check Data Coding and Its Documentation: For any variable in a data file, its data codes, range, and unit of measurement must be documented in the associated data dictionary before data were entered. If any code assigned to a variable was found not defined in the data dictionary, data originator will be contacted to obtain its definition. If it is an error, the established protocol will be followed for making the correction. (c) Summary Statistics: The summary statistics, e.g., mean, standard deviation, samples quantiles and sample frequencies, of all interested variables were calculated and reported as tables. After we have basic understanding of our data, we moved to the next step. (d) Exploratory Data Analysis is performed to investigate the Biological and Ecological relationships among variables in the available data sets. We did histograms, Box plots, scatter plots, correlation analysis, and simple regression to see if data demonstrated the known or any unknown characteristics. (e) Modeling and Testing the Effectiveness of Intervention: Linear models and generalized linear models were used to test treatment effects and the effectiveness of intervention. (f) Exposure pathway analysis: The migration path way of environmental toxicant and the human exposure pathway to the environmental toxicant were analyzed through structural equation models, models with both sides of variables are random, etc.

HOW SAS AND THE PERSONAL COMPUTER BASED WORD PROCESSOR, SPREADSHEET, AND STATISTICAL GRAPHICS PACKAGES INTERACT WITH EACH OTHER IN PREPARING REPORT

Study outcomes and findings through summary statistics and statistical data analysis are the integral parts of projects reports. Statistical graphics are also important in conveying the study results to the readers.

Several SAS programs were written to generate tables which can be easily read into word processor. A procedure was also developed to convert SAS data files so Postscript® based statistical graphics can be generated. Spreadsheet and word processor are also used to format tables and to prepare the integrated reports. In our studies, a report is the electronic integration of several documents generated by different software packages.

HOW SAS WAS APPLIED IN DEVELOPING NEW STATISTICAL PROCEDURES AND METHODOLOGIES.

Several statistical procedures were developed to facilitate data analysis or to meet project's needs. SAS IML and other SAS procedures were used to carry out the numerical calculation or obtain the empirical results. For example, a measurement errors model with random coefficient was developed to obtain the consensus values of a set of unknown standards which were analyzed by several laboratories (Pan etc. 1992c). Another measurement errors model with random coefficient was developed to re-calibrate the sample chemical results obtained under one calibration curve to the equivalent results as those samples were analyzed under another calibration curve (Pan etc. 1992a). Currently, we are working on longitudinal data models which has both sides of variables random.

CONCLUSION AND COMMENT

SAS® is an excellent statistics software package as recognized by many statisticians. However, the needs in each applied statistics area is changing and getting more demanding. Due to the exploded development in computer hardware and software, we can do a lot more data analysis and produce much nice looking report than ten years ago. Similarly, our colleague and co-workers are also raising their expectation on us even more quickly. To cope with the changing technology and society, we found we should be
INNOVATIVE  Develop new procedures to efficiently analyze data, generate tables, plot statistical graphs, and prepare reports.

ADAPTIVE  Adapt to the change in computer hardware and software as well the societal change, i.e., the change from mainframe computer to workstation and personal computer, the change in computer networking, the change in the availability of high quality input and output devices.

AN INTEGRATOR  Capable of integrating several different hardware and software together to produce the best outcome in the most cost effective fashion. For example, our data management system which takes care of data entry is based on Apple Macintosh®, our data analysis is performed on either mainframe or PC with SAS and BMDP®, tables are generated by SAS and formatted by Microsoft Excel for Macintosh®, statistical graphics are generated by KaleidaGraph® on Macintosh, and our reports are prepared on Macintosh with Microsoft Word® which integrate the output generated by all the above software into a single document. Report is printed on an Apple LaserWriter II NT®. The GIS we used is a PC based system PC ARC/INFO®. We planned, purchased and set up all the PC and Macintosh based hardware and software.

ACKNOWLEDGMENT

Support for this research came in part from US EPA Grant No.: V00 5035-01. Drs. R. Bornschein and S. Clark are the co-principal investigators.

SAS is a registered trademark or trademark of SAS Institute Inc. in the USA and other countries. Other brand and product names are registered trademarks or trademarks of their respective companies.

REFERENCES


AUTHOR ADDRESS

WinKey Pan, Ph. D. Department of Environmental Health, University of Cincinnati, 3223 Eden Avenue, ML 0056 Cincinnati, OHIO 45267-0056