Handbook of
SAS® DATA Step Programming

Handbook of SAS® DATA Step Programming

Arthur Li

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20130226

International Standard Book Number-13: 978-1-4665-5239-5 (eBook - PDF)

To Dave

Contents

Preface

A common statistical programmer's task generally begins with reading one or more raw data sets into a statistical software and performing data management and manipulation, such as checking and modifying values of variables to ensure they satisfy the analytic quality, transposing data into a desired shape for a certain type of statistical analysis, merging multiple data sets by common variables, etc. Once the data has been assured of possessing the desired quality, the transformed data is ready for statistical analysis by using procedures that are provided by the statistical software. Data manipulation is an essential step to obtaining a reliable, statistical, analytical result because an analytical result that is based on unreliable data is not trustworthy; this is often referred to as "garbage in, garbage out." Successfully creating reliable data solely depends upon writing an accurate computer program.

In SAS®, data manipulation and management are mostly performed in the DATA steps; conducting statistical analysis and creating reports is carried out by using SAS procedures (PROC steps). A computer program that is used to perform data manipulation and analysis consists of a series of DATA or PROC steps or both and is written within the SAS programming language. The DATA and PROC steps consist of a series of SAS statements that are created by following the SAS language syntax. A PROC step is often easy to write because it is purely syntax driven; therefore, simply knowing how to follow the syntax from the documentation will sometimes be sufficient to help you accomplish the task successfully. On the other hand, in order to write an accomplished program in the DATA step, a programmer must be able to understand programming logic and to know how to implement and even create his or her own programming algorithm.

The focus of this book is not about learning statistical procedures but rather learning how best to manage and manipulate data by using the DATA step. Beginning programmers often tend to focus on learning syntax without focusing on programming logic and algorithms, which often results in common problems when they create a SAS data set. For example, the data set that they created is not what they originally intended to create—that is, there are more or less observations than intended or the value of the newly created variable was not retained correctly. These types of mistakes are most commonly committed because programming novices don't understand fundamental and unique SAS programming concepts, such as understanding the compilation and execution phases of the DATA step, what happens in the program data vector (PDV) during the DATA step execution, etc. This book will provide insight to readers that simply learning syntax will not solve all the problems that they'll

encounter; instead, they need to understand SAS processing in order to be successful programmers.

Another common problem novice programmers face is a lack of programming strategies. Therefore, when SAS programming novices encounter a new programming task, often they don't know where to begin and what steps will be involved in solving the problem. Most of the examples in this book begin with discussing the strategies and steps for solving the problems, then providing a solution, and in the end, providing a more detailed explanation for the solution.

An Overview of SAS Software

SAS was originally the acronym for Statistical Analysis System, which is an integrated software system that utilizes fourth-generation programming language to perform tasks like data management, report writing, statistical analysis, data warehousing, and application development. SAS has undergone various upgrades to its software system over the years. This book was written using Version 9.2.

The core component of the SAS system is Base SAS software, which consists of different modules such as DATA steps, SAS Base procedures, SAS macro facility, and Output Delivery System (ODS). Among the modules in Base SAS, this book covers the DATA step and some of the SAS Base procedures that relate to data management.

SAS provides multiple methods for starting and running your SAS program, which depends upon your operating system. The common method for most SAS users is to utilize the SAS Windowing Environment. Alternatively, you can also run your program by using an interactive or noninteractive line mode, as well as batch mode. Please refer to SAS documentation for these alternative methods.

SAS Windowing Environment (illustrated in Figure 0.1) consists of five windows where you can create and edit your SAS program and manage your SAS files, which includes Program Editor, Log, Output, Result, and Explorer windows. To create a SAS program in the SAS Windowing Environment, you can use the Program Editor window to write your SAS code. You can either submit your entire program at once or highlight only part of the program that you wish to run by clicking on the "Submit" button on the tool bar or by selecting "Run ◊ Submit" from the tool bar. Once the program is submitted, SAS will display messages such as the name of the newly created data set and error or warning messages in the Log window. If the program generates output, it will be displayed in the Output window. You can use the Result window to view and manage different output that is created from your program. To manage files that are stored in the SAS library, you can utilize the Explorer window.

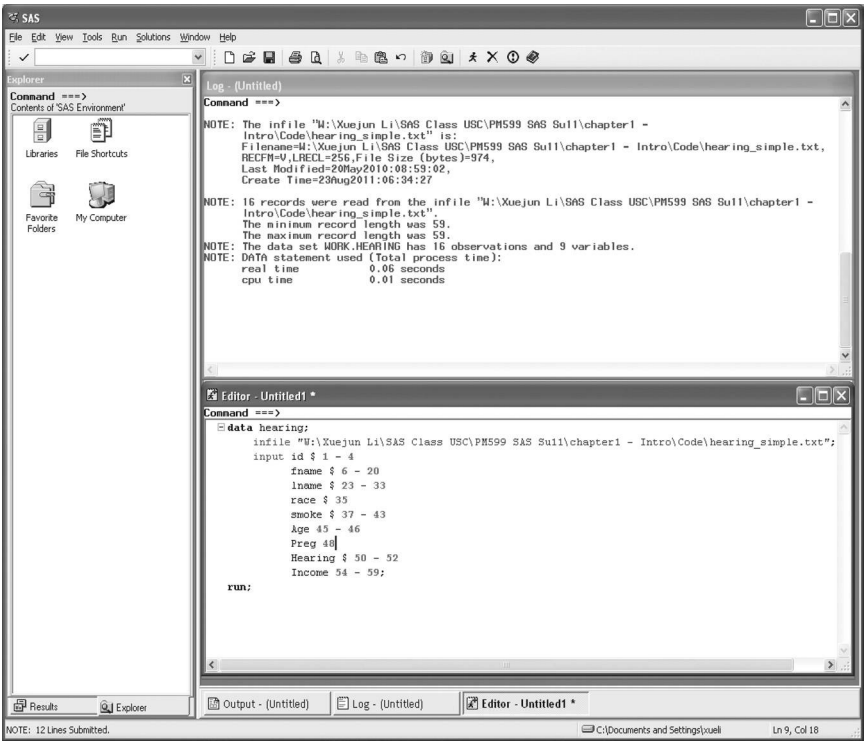

FIGURE 0.1
SAS windowing environment.

SAS Help and Documentation

One of the best resources for learning the SAS language and searching for SAS programming help is *SAS Help and Documentation*, which is part of your SAS software installation. To open *SAS Help and Documentation*, click on the "Help" icon on the tool bar, which will bring up the documents.

There are two windows within *SAS Help and Documentation*. The left window, which contains four tabs (Contents, Index, Search, and Favorites), can help you navigate the different topics. The contents of the selected topic are displayed on the right side of the window. The SAS documentations are grouped by categories under the Contents tab. The Contents tab can often be more useful compared to the other tabs if you know how each topic is categorized.

Most of the references in this book are based on *SAS Help and Documentation*. Every topic within this book is under the "SAS Products ◊ Base SAS" main branch. Within this main branch, if you are looking for documents for

non-statistical procedures, it will be under "Base SAS 9.2 Procedure Guide." The topic under "SAS 9.2 Language Reference Concepts" provides detailed documentation about SAS system concepts, SAS DATA steps, and file concepts. Documents that are related to the language elements, such as SAS data set options, formats, functions, informats, statements, and system options, can be found under "SAS 9.2 Language Reference: Dictionary ◊ Dictionary of Language Elements." This search path for SAS documentation through the HELP menu was introduced with SAS Version 9.2. If you are using another version of SAS, such as 9.3, the documentations might be organized differently. Thus, instead of listing the entire search path for a specific documentation, only the name of the last node, such as "Dictionary of Language Elements," is provided as a reference in the book. If you cannot locate the referenced documentation from the HELP menu, an alternative and effective method is to perform a Google search on the name of the documentation (just make sure to include "SAS" in the search string).

How Best to Navigate This Book

The contents of this book are grouped into ten chapters. These ten chapters can be grouped into three main sections.

The first section of the book includes Chapter 1 and Chapter 2, which serve as prerequisite reading for the main section of the book. Chapter 1 provides an introduction and overview of the SAS language, which includes reading data into the SAS system, some Base SAS procedures, creating variables, and subsetting data sets. Some of the topics in these chapters will also be expanded in detail in the later chapters. Chapter 2 discusses how to conditionally create variables based on existing variables, such as how best to use the IF-THEN/ELSE statement or the SELECT statement to create variables.

The second section of the book includes Chapters 3 through 6, which is also the core component of the entire book. These four chapters describe how essential it is to understand the PDV in order to write an accurate program in the DATA step. The topic covered in each chapter is built upon the concepts covered in a previous chapter. Chapter 3 provides an overview about DATA step processing, how to retain variables by using the RETAIN and SUM statements, conditional processing in the DATA step, and how to use the DATA step debugger. Chapter 4 covers BY-group processing within the DATA step and some applications relating to longitudinal data sets. Chapter 5 introduces topics on explicit loops in the DATA step and compares the differences between implicit and explicit loops. Chapter 6 covers array processing, which includes one- to multi-dimensional arrays. Many applications that utilize array processing are related to understanding the explicit loop.

The final section of the book includes Chapters 7 to 10. These four chapters are not built in sequence and can be read independently. Chapter 7 covers multiple methods for combining data sets vertically and horizontally. Although some examples in this chapter demonstrate an understanding of the PDV, readers should be able to grasp and understand most of the materials with or without having read the core section of the book. Chapter 8 covers data input and output. This chapter covers different input methods of reading and writing text files in the DATA step. Chapter 9 covers DATA step functions and call routines, which represent one of the many strengths of SAS software. This chapter introduces a few categories of SAS functions or call routines, such as date and time functions, character functions, and functions for converting variable types. Chapter 10 covers some useful SAS procedures that relate to data management, which include the SORT, COMPARE, TRANSPOSE, FORMAT, and OPTIONS procedures.

A novice programmer with minimum SAS background should read this book cover to cover in the order in which the book is presented. An intermediate SAS programmer who is interested in learning the concept of DATA step processing can begin reading Chapters 3 to 6 without reading the first two prerequisite chapters.

A Note on SAS Output and Log Font

The output and logs that are generated from SAS procedures use the SAS Monospace font. However, the font used for SAS output and logs in this book is `Lucida Sans Typewriter`. The drawback of using the `Lucida Sans Typewriter` font is that grid lines for the table displays, such as the output from PROC FREQ, do not have the grid line "look." Here's an example of output from the PROC FREQ using the `Lucida Sans Typewriter` font:

Output using `Lucida Sans Typewriter`:

```
                    The FREQ Procedure
                                        Cumulative   Cumulative
  Preg     Frequency     Percent     Frequency      Percent
  ffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffffff
     0            19       63.33            19        63.33
     1            11       36.67            30       100.00
                  Frequency Missing = 4
```

The "*f*" symbols in the output above appear to be a straight line in the SAS Monospace font. In this situation, the "*f*" symbol is replaced with

a dash ("-") to achieve an easier-to-comprehend display of the output, like the one below:

Output using `Lucida Sans Typewriter` *with modification:*

```
                    The FREQ Procedure
                                    Cumulative   Cumulative
Preg     Frequency     Percent     Frequency      Percent
- - - - - - - - - - - - - - - - - - - - - - - - - - - - -
   0            19       63.33            19        63.33
   1            11       36.67            30       100.00
                  Frequency Missing = 4
```

Additional Components of This Book

Here is a list of materials that readers can download from the publisher's Web site (http://www.crcpress.com/product/isbn/9781466552388):

- All the data sets used in the book, as well as all the programs
- Exercise data sets and their solutions
- Slides for demonstrating the contents of the PDV for some examples in Chapters 3 through 6

The Author

After receiving an MS in Biostatistics from the University of Southern California (USC), **Arthur Li** embarked on a career as a biostatistician at the City of Hope National Medical Center, in Los Angeles County, California. Li is also a part-time statistical programming instructor at USC. He has given numerous presentations and seminars on DATA step programming and statistical analysis using SAS software at SAS conferences throughout the United States.

Acknowledgments

This book would not have been possible without the tremendous support of my mentor, Prof. Stanley Azen, who provided me the opportunity to teach my SAS class at the Department of Preventive Medicine at the University of Southern California (USC) for the past five years. The accumulated course materials bear an imprint upon this book. I would also like to express my sincere gratitude to Prof. Jim Gauderman, my first SAS teacher, who taught me much while I was his student and his teaching assistant. I also have to thank Prof. Roberta Mckean-Cowdin, who enriched my hands-on SAS experience from the many research projects that were provided by her.

I would also like to thank Xiaolong Li, Han Tun, and Wuchen Zhao for reviewing the contents and testing all the programs in this book.

I would like to acknowledge the enthusiastic and delightful people from the SAS community who provided invaluable support throughout the years for my recognition in the SAS community: MaryAnne DePesquo, Perry Watts, Sunil Gupta, Ron Cody, Kirk Lafler, Peter Eberhardt, and my dear friend Nate Derby. Among them, Sunil Gupta, Ron Cody, and Peter Eberhardt provided many helpful technical suggestions for this book. I am especially deeply indebted to Perry Watts, who spent enormous chunks of her personal time to provide detailed critiques from cover to cover regarding organization, structure, references, and programs/exercises, resulting in a much more precise product.

I am also grateful for the help I received from Taylor & Francis: Rob Calver (Senior Acquisitions Editor), Rachel Holt (Senior Editorial Assistant), Kathryn Everett (Project Coordinator), and Amy Rodriguez (Production Editor).

I would lastly like to express my deepest thanks and gratitude to my beloved Zaccagnino family, especially my "crazy" but lovable in-laws, Dan and Sally Zaccagnino. The Zaccagnino family's endless support makes my success possible. Finally, this book could not have been completed without my partner, David Zaccagnino, who not only edited the ESL (English as a second language) errors in this book but most importantly provided me with endless patience and encouragement.