Journal of Statistical Software

January 2014, Volume 56, Book Review 1. http://www.jstatsoft.org/

Reviewer: Xiangxiang Meng, Wayne Thompson
SAS Institute Inc.

Handbook of SAS® DATA Step Programming

Arthur Li
Chapman & Hall/CRC, Boca Raton, FL, 2013.
ISBN 978-1-46655-239-5. 275 pp. USD 59.95.
http://www.crcpress.com/product/isbn/9781466552388

The SAS DATA step is one of the most widely used statistical programming languages for data manipulation, data set management and statistical summary. "Handbook of SAS DATA Step Programming" by Arthur Li provides a thorough introduction to the statements and functionalities of the SAS DATA step. The book can be used as an introductory tutorial for beginning SAS programmers, or as a reference book for experienced users.

Unlike other tutorials for SAS programming, the book focuses on DATA step. Without using any SAS macro language or PROC SQL, the book shows the comprehensive capabilities of DATA step to manipulate and summarize data, including data input/output, table merge and concatenation, column modification and derivation, and calculation of summary statistics. The book not only shows how to program in a DATA step, but also explains in details how the DATA step works when it executes a sequence of statements through an implicit loop of observations. What happens inside the SAS program data vector (PDV) is explained in full detail for many important elements of the DATA step, such as the RETAIN statement and the BY processing.

The first two chapters of the book serve as a tutorial for beginners. Instead of introducing DATA step as a statistical or data manipulation tool, Chapter 1 describes it as a programming language, and well explains the building blocks and basic syntax of SAS programming language and SAS DATA step. The most widely used SAS statements and SAS/BASE procedures are then introduced with examples that are simple and easy to use. Chapter 2 follows the style of Chapter 1 and continues to introduce some commonly used features of DATA step, such as IF-THEN-ELSE and DO-END for conditional and group processing of SAS statements. In general, the functionality of DATA step is very comprehensive and complex. Instead of covering every statement of DATA step, these two chapters try to describe what a DATA step can do by using examples that are frequently used by a SAS programmer to process real data, such as clinical trial data from pharmaceutical industry.

For users who have programed in SAS DATA step and want to better understand how it works, Chapter 3 is an excellent overview of the underlying mechanism of DATA step. DATA step executes the statements in an implicit loop through observations, with one observation loaded

into a program data vector after another. The underlying mechanism of DATA step cannot be well described without explaining how the data is processed in PDV. In this chapter, the data flow in PDV is well explained with two statements, RETAIN and SUM. These two statements are actually the most basic but commonly used statements that requires a PDV to save some information from one observation to be used by the next observation. And they are the building blocks for many complex data step applications, such as last-value-carry-forward and data aggregation. Debugging using PUT statement and SAS debugger are also introduced in this chapter. We recommend a careful reading of the PDV examples in this chapter (such as Figure 3.5, 3.6, 3.7 for RETAIN statement). Understanding how the PDV works is the best way to improve programming efficiency using SAS DATA step.

Chapter 4 to 6 introduce three important concepts in SAS DATA step: by-group processing, do-loop, and variable array. Each of these concepts are explained with examples to show the underlying data processing in the program data vector. Most of the examples in these chapters are well known tips and tricks for DATA step, such as example 4.3 for removing duplicated data, and example 5.5 for random number generation.

The next four chapters cover the details of some important DATA step statements and concepts that have been briefly introduced in first two chapter, such as SET/MERGE for table concatenation and joint (Chapter 7), the INFILE/INPUT statements for reading external data source (Chapter 8), common SAS functions and call routines (Chapter 9), and some important SAS/BASE procedures (Chapter 10). Although reading data directly from a flat table becomes a standard way to import data source into SAS, reading data stored in a non-flat or compressed type of source file is still common, and Section 8.2 provides an overview of the INPUT statement for reading such text files. SAS date/time functions, character functions, and INPUT/PUT functons are powerful add-ons to DATA step, and they are introduced in Section 9.2 to 9.4. the TRANSPOSE procedure is another important feature that requires careful reading. Readers are recommended to read through the six examples in this section to understand how this procedure can transpose a data from one shape to another. Last but not least, SAS user-defined format is a unique SAS feature to transform or translate the information in a column without duplicating the column. For example, example 10.17 illustrates how to use grouping formats to categorize two continuous variables AGE and INCOME, and how to use the formatted continuous variables in the FREQ procedure to produce cross-tabulations. The examples in this section well describe SAS formats as a derivation or mapping tool to transform variable efficiently.

In summary, we applaud Anthur for introducing SAS DATA step as a programming language, and explaining how things work behind the grammar of DATA step. Many users of DATA step have a strong statistical background, but not all of them really try to understand DATA step as a programming language. The "Handbook of SAS DATA Step Programming" covers this gap, and helps statisticians improving their programming efficiency using SAS DATA step.

Reviewer:

Xiangxiang Meng, Wayne Thompson
SAS Institute Inc.

100 SAS Campus Drive
Cary, NC 27513, United States of America
E-mail: Xiangxiang.Meng@sas.com, Wayne.Thompson@sas.com

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 56, Book Review 1 *Published:* 2014-01-06
January 2014