Journal of Statistical Software
July 2006, Volume 16, Book Review 5. http://www.jstatsoft.org/

Reviewer: Rodney Sparapani
Medical College of Wisconsin

Statistical Analysis of Medical Data Using SAS

Geoff Der and Brian S. Everitt
Chapman & Hall/CRC, Boca Raton, Florida, 2006.
ISBN 1-58488-469-X. 428 pp. GBP 39.99.
`http://ftp.sas.com/samples/A60928`

Introduction

The audience of this book includes those who have very little or no experience with SAS. Only very basic computer familiarity is assumed. The goal of the book is to give you the skills necessary to perform statistical analyses with SAS v. 9.1 which is current as of this time. Much of the knowledge that you learn will most likely be useful in future versions of SAS and should allow you to function comfortably with legacy versions that are common today.

The authors attempt to explain statistical concepts and methodology without assuming a lot of background knowledge. However, the book is not aimed at those who have no statistical training. The assumption is that the reader has a basic understanding of statistics, but may not be familiar with a particular statistical idea or technique.

Chapter 1: An introduction to SAS

This chapter gives you a basic understanding of SAS necessary for simple tasks like creating SAS datasets from text files and manipulating SAS datasets such as subsetting or merging. If you are already familiar with SAS you can skim this chapter. For those whose only knowledge of SAS is based on v. 6, there are some new features discussed that you may find useful such as `PROC IMPORT` and the output delivery system (ODS).

Chapter 2: Describing and summarizing data

SAS has many tools for summarizing data. The authors mention a few with some extra emphasis in areas that biostatisticians will appreciate. Specifically, the capabilities of SAS to easily generate graphical figures such as histograms with super-imposed kernel density estimates (PROC UNIVARIATE and the HISTOGRAM statement), box-plots (PROC BOXPLOT) and bar charts (PROC GCHART).

Chapter 3: Basic inference

Simple hypothesis testing of continuous and categorical data are explained and the corresponding SAS code presented. The topics covered include the 2-sample t test, the Wilcoxon-Mann-Whitney rank sum test, the paired t test, the Wilcoxon signed rank test, the χ^2 test for independence in 2×2 and $r \times c$ contingency tables, Fisher's exact test, the Mantel-Haenszel test and McNemar's test.

Chapter 4: Scatterplots, correlation, simple regression and smoothing

More graphical displays are introduced. Scatterplots are discussed (PROC GPLOT) with some advanced features of overlaying text and graphics with ANNOTATE. The code and data that are used throughout the book can be found online at the URL above. This chapter discusses a very useful SAS macro found there: %plotmat, which can be used to place multiple scatterplots on one page of output.

Density estimation is an important recent addition to SAS (PROC KDE). This topic is introduced with corresponding 2-dimensional (PROC GCONTOUR) and 3-dimensional (PROC G3D) displays presented. Also introduced in this chapter are Pearson's correlation coefficient (PROC CORR), partial correlation coefficients, simple linear regression (PROC REG) and LOESS (PROC LOESS).

Chapter 5: Analysis of variance and covariance

Balanced/unbalanced ANOVA is introduced along with the some more in-depth topics like Scheffe's multiple comparison procedure, factorial experiments and type I/II/III sums of squares. Non-parametric ANOVA and ANCOVA are touched on as well as more graphical displays and the %template SAS macro which can be used to place multiple displays on one page of output. Other useful SAS features are introduced as well such as arrays, the logical IN operator and the character string concatenation (||) operator.

Chapter 6: Multiple regression

Multiple regression (PROC REG) is introduced including estimation, testing and diagnostics plots (predicted values, residuals and Cook's D). Some more advanced topics are also presented: variance inflation factors (VIFs), Mallow's C_p and forward/backward/stepwise selection.

Chapter 7: Logistic regression

Multiple logistic regression (PROC LOGISTIC) is introduced including estimation, testing and residual diagnostics (Pearson and deviance). Conditional logistic regression is also briefly presented.

Chapter 8: The generalized linear model

The generalized linear model (PROC GENMOD) is introduced. Poisson regression and errors that follow the Γ distribution are highlighted. Other general topics covered are residuals and overdispersion.

Chapter 9: Generalized additive models

For those who have a background in statistics, this chapter should be a pleasant surprise. For this demographic, the book has probably been pretty slow reading to this point. To learn SAS, you necessarily start with the basics which can be very uninteresting to read. But, for most readers, generalized additive models (GAM) will be a new topic and dispel the drudgery. Beyond the interest of the topic, the authors also use the chapter to introduce many other features of SAS to the reader; this intersection makes it the best chapter of the book.

PROC GAM provides smoothing via LOESS, cubic splines and thin plate splines. Other important SAS knowledge is imparted as well, such as the RETAIN statement, the INPUT statement with the trailing @ symbol for reading multiple observations from a single row of data, interpolation and plotting of data via the SYMBOL statement, an introduction to ODS for exporting output to other formats besides plain text and the introduction of the %panelplot (yet another SAS macro useful for combining multiple displays).

One minor quibble in an otherwise well-written chapter. On page 247, the authors show how to delete all observations from a SAS dataset that are associated with the city of Chicago as follows:

```
if city=:'Chicago' then delete;
```

Although, this code will actually work, it gives the reader the wrong impression. The exact comparison operator = would be more appropriate since it compares all of the characters in a string, whereas the comparison operator =: only compares the beginning of a character string. For example, suppose that we wanted to delete the observations for all cities that start with Ch; in that case, the “begins with” operator would be very handy.

Chapter 10: Nonlinear regression models

Nonlinear regression (PROC NLIN) is introduced. This seems like a good chapter to skip until the need for this type of analysis arises.

Chapter 11: The analysis of longitudinal data I

Some nice features for graphically displaying longitudinal data are discussed such as mean and standard deviation plots and box plots that are also produced by the SYMBOL statements interpolation (I=) option. Other topics include handling missing data and data transformations.

Chapter 12: The analysis of longitudinal data II: Models for normal response variables

The discussion of longitudinal data continues with mixed models for repeated measures data (`PROC MIXED`). Some advanced topics are introduced such as methods for modeling the denominator degrees of freedom in the F test, the estimating and testing of the covariance structure of the random effects, prediction of random effects and dropout assumptions.

Chapter 13: The analysis of longitudinal data III: Non-normal responses

Two strategies are introduced: Generalized estimating equations (GEE, `PROC GENMOD`) and non-linear mixed models (`PROC GLIMMIX`). Although, GEE has been around for a while, it has only gained more attention as software packages like SAS have built in support for it. Similarly, non-linear mixed models have been popular with Bayesians since they were supported by popular Bayesian software packages like **BUGS**. Now, frequentists have access to both of these with SAS, although non-linear mixed models will take considerably more CPU time. The authors should be commended for tackling such a modern topic that many readers will not be familiar with.

Chapter 14: Survival analysis

Survival analysis is a topic that is central to biostatistics, yet peripheral to statistics in general. That makes it an ideal topic for this book. The concepts of the survivor function and the hazard function are introduced. The Kaplan-Meier estimator of the survivor function is described as well as it's calculation and corresponding inference (`PROC LIFETEST`). Cox's proportional hazards model (`PROC TPHREG`) is also introduced as well as more advanced topics such as time-varying covariates and diagnostics with residuals.

Chapter 15: Analysing multivariate data: Principal components and cluster analysis

The methods of principal components (`PROC PRINCOMP`) and cluster analysis (`PROC CLUSTER`) are introduced. The graphical display of cluster analysis via dendograms (`PROC TREE`) is also discussed. This is a chapter that might be skipped on a first reading.

Comment

The purpose of the book is not to replace the help that SAS provides online and/or in their manuals. Rather, the purpose of the book is to familiarize the reader with certain `DATASTEP` statements and `PROC` procedures. Therefore, the reader will have an idea how to go about their task and, if necessary, what to look for online and/or in the manuals for help. Even restricting ourselves to the SAS/BASE, SAS/STAT and SAS/GRAPH products, SAS is sufficiently complicated that there is a niche market for a book that can hold a newbie's hand while also being an occasional reference for the more experienced.

My biggest complaint about the book is that I didn't like the title. The words "medical data" have many connotations and initially gave me the wrong impression. "Biomedical

Data" or "clinical data" would have been better choices perhaps; or even working in the more common term "biostatistics". Although the title of a book is important, the content is far more important; and where the content is concerned the book is a good introduction to SAS for those wanting to analyze biomedical data.

What's missing? Obviously, you can't put in every topic since that is what the whole bookshelf of SAS manuals is for. But, there are certain things that all SAS users struggle with such as SAS dates/times, floating-point precision, the use of the SAS macro facility or `FIRST./LAST.` indicators for repeated keys. However, I found myself struggling to come up with things that the authors did not at least mention. That is a good sign. I recommend this book as either a text for a course, or a companion, for those who are new to SAS.

Reviewer:

Rodney Sparapani
Medical College of Wisconsin
Center for Patient Care and Outcomes Research
8701 Watertown Plank Rd.
PO Box 26509
Milwaukee, WI 53226, United States of America
E-mail: rsparapa@mcw.edu

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 16, Book Review 5 *Published:* 2006-07-26
July 2006