# Review of *Modern Applied Statistics with S-PLUS*

by W. N. Venables and B. D. Ripley

David A. James

AT&T Bell Laboratories
Murray Hill, NJ 07974

April 14, 1995

The S environment (Becker, Chambers, and Wilks (1988) and Chambers and Hastie (1992)) and its commercial implementation, S-PLUS by the StatSci division of MathSoft, consist of a rich language for programming with data and graphics plus extensive software to conduct most commonly used statistical techniques (regression, analysis of variance, statistical graphics, etc.) The language allows users to easily extend the environment by expressing computations in the S language itself and, if needed, incorporating Fortran or C code.

Profs. Venables and Ripley, like other researchers, have recognized the fundamental changes to applied statistics brought by advances in computing, and that the S/S-PLUS environment is well suited for exploiting these advances. (The LISP-STAT environment, Tierney (1990), provides similar advanced capabilities.) In this book the authors set out to show how they use the S/S-PLUS environment and how it "allows penetrating analyses to be performed routinely."

There have been other recent books describing aspects of the S/S-PLUS environment. Everitt (1994) presents numerous examples of data analyses performed in S/S-PLUS, while Spector (1994) concentrates on describing the S language and only briefly summarizes in one chapter some of the tools for fitting some statistical models. Profs. Venables and Ripley cover both aspects in their book.

This book is meant to be an introduction to would-be users of S/S-PLUS and for class use. It is not meant to be a text on statistical theory, nor comprehensive, nor to replace the S or S-PLUS manuals. Nevertheless, the task the authors have undertaken is challenging: getting new S/S-PLUS users to quickly learn the fundamentals of the language, and to present a modern approach to data analysis through numerous examples from many areas of statistics. They succeed in this at the expense of exploratory data analysis during model building. (More on this later.)

The book consists of 15 chapters, 5 appendices, and it includes a diskette with data sets used in the examples and a library of functions the authors have developed to implement a wide range of techniques not currently available on either S or S-PLUS (e.g., neural networks, spatial statistics, a negative binomial family for generalized linear models); in addition they include some bug fixes to some S/S-PLUS code.

The installation of the software was easy, both on a UNIX system running S-PLUS 3.1, and on a personal computer running S-PLUS for Windows version 3.2. Although I did not attempt to stress test their software, I had no problems using it.

The first 5 chapters (about 150 pages) concentrate on the S language, its basic data structures, graphics, etc. These chapters include exercises, most of which deal with problems encountered in practice.

Chapter 1 is a quick introduction to S/S-PLUS; it describes how to invoke the environment and how to exit it; the chapter also shows how to access the library of functions they provide (its installation is described in Appendix A), and a sample session.

Chapter 2 describes the fundamentals of the S language: object names, vectors, matrices, arithmetic and logical expressions, reading data from ASCII files, etc. The last 4 sections in this chapter deal with advanced topics, and the authors wisely suggest skipping them on first reading. However later chapters rely heavily on this material, and users will need to go back and study in detail these sections sooner rather than later, since the authors routinely use fairly advanced S constructs.

Chapter 3 shows how to generate most common statistical graphs: scatterplots, barplots, conditional plots, scatterplot matrices for multivariate data, perspective plots, contour plots, etc. It also shows how to control the layout and attributes of graphics, e.g., line types, colors, plotting marks, character sizes, legends.

Chapter 4 deals with writing functions to extend the computations in S. The authors show quite effectively the ease with which users can write and modify S functions to tailor their analyses. The material on evaluation of S expressions by the S interpreter is a fairly advanced topic, and many readers may need to refer to Becker, Chambers and Wilks (1988) to better understand this material. Section 4.5 seems to stress recursion a bit too much; one may view recursion as incidental, a useful byproduct of the S rules for evaluating expression. This chapter also shows how to invoke Fortran and C code from S; readers have many good examples of how to do this in the library of functions the authors provide. Readers interested in learning the object-oriented style of S programming will need to refer to Chambers and Hastie (1992) or the S-PLUS manuals.

Chapter 5 covers data summaries, probability distributions (including their quantile and density functions, and generation of pseudo random numbers), histograms, stem-and-leaf plots, QQ-plots, boxplots, density estimation, functions for hypothesis testing (T-test, Chi-square, binomial, etc.), and some bootstrap methods and permutation tests. I found the discussion on QQ-plots inadequately brief, even though the authors consider QQ-plots "[o]ne of the best ways to compare the distribution of a sample `x` with a distribution...". (For a good discussion on interpreting QQ-plots, see, for instance, Chambers et al.(1983).)

The remaining chapters cover S/S-PLUS analyses from specific statistical areas; linear and generalized linear models, robust and resistant methods, non-linear models, "modern" regression, survival analysis, multivariate analyses, classification and regression trees, time series, and spatial statistics. These chapters are self-contained, except for the material on S formulas which is covered in Chapter 6 and used in later chapters. Each chapter starts with a brief summary of the statistical methodology before presenting the core S/S-PLUS material. Comparing this list with the table of contents in Chambers and Hastie (1992) we notice a substantial overlap in the topics, but it is clear (the authors state this in the preface) that they do not attempt to replace this book or the S-PLUS manuals. By and large, Venables and Ripley (1994) provide effective introduction to these topics, but more advanced S/S-PLUS users will need to refer to Chambers and Hastie (1992) or the S-PLUS manuals.

Chapter 6 covers the vast area of linear models; single and multiple regression, analysis of variance (fixed and random effects), design of experiments. Also, in this chapter they cover S formulas, which are expressions used to specify relationships between a response variable and a set of predictors, and this topic is a prerequisite for most subsequent chapters. The authors cover in detail crossed and nested models, the various possibilities when predictors include continuous and categorical (nominal and ordinal) data; they also describe various model parametrizations or contrasts (sum, treatment, Helmert, etc.) and how to specify them.

This chapter feels long. There is just too much important material, and probably it could be split into two chapters. Also, since the material on S formulas is general enough to specify models other than linear, it could reasonably be in a previous chapter.

Chapter 7 on generalized linear models contains good examples for binomial and poisson regression; also, it shows how to extend the S/S-PLUS `glm` computations to include a negative binomial family.

Chapter 8 on robust methods covers both univariate techniques as well as robust regression; in Section 8.3 again they nicely show how to extend the S computations, in this case by "robustifying" the linear model by creating a new class of robust linear model `rlm` class that builds on `lm` methods. However, the authors do not discuss transformations as a way to analyze long-tailed data, as in Mosteller and Tukey (1977).

Chapter 9 on non-linear models is a very nice complement to the Bates and Chambers chapter in Chambers and Hestie (1992). The examples are interesting and realistic, in particular the Stormer's viscometer example. Moreover, the example on fitting data to a mixture of gaussian distributions shows effectively the power of the S language to easily express and carry out complex statistical computations. The authors walk the reader through S code that computes maximum-likelihood estimates by minimizing the sum of the contributions from each observation to the (negative) likelihood function (for this they use the S function `ms`); they also show how to symbolically compute first and second derivatives, how to compute large-sample standard errors of the estimates, and finally how to compute quantiles of the mixed distribution by inverting the estimated mixed density function using the Newton-Raphson method.

Chapter 10 on modern regression briefly discusses smoothing and includes topics such as generalized additive models, projection pursuit, alternating conditional expectations, additive and variance stabilizing models, and neural nets. The commonly used loess method (Stone (1977), Cleveland (1979)) is mentioned here and in Chapter 15 but not discussed.

Chapter 11 deals with methods for lifetime data, both Cox's proportional hazard models (including time-varying coefficients) and parametric models. The software behind this chapter, `survival3`, was developed by Terry Thurneau at the Mayo Clinic and is freely available through the statlib server at Carnegie Mellon University; it will be available in a future release of S-PLUS. In the discussions on formulas to specify survival models, the distinction between factors and strata is not made clear. This is puzzling, since strata variables *are* factor variables. For instance, it is not clear from this chapter why when fitting a Cox proportional hazard model the following two expressions, `strata(ag) * log(wbc)`, and `ag * log(wbc)` differ, given that `ag` and `strata(ag)` both are factors.

Chapter 12 and 13 cover multivariate analyses and tree-based models, respectively. Diagnostics for tree-based models are not covered as extensively as in the Clark and Pregibon chapter in Chambers and Hastie (1992); for instance there is no mention to `burl.tree, hist.tree, tile.tree`, etc.

Like Chapter 6, Chapter 11 on survival analyses and Chapter 14 on time series cover extensive topics at the expense of preliminary analyses. For instance there is no mention of exploratory data analyses to assess the various assumptions (e.g., distributional, independence, proportional hazards) prior to model fitting, as done in Freeny and Nair (1994), Mosteller and Tukey(1977), Chambers et al. (1983), etc. The concept of residuals and its generalizations under censoring are briefly presented; however these residuals are hard to analyze properly and many readers may need more examples.

The last chapter on spatial statistics has only two sections, and one is left with the feeling that there is a lot of material the authors left out, especially looking at Prof. Ripley's extensive research in this area.

It seems somewhat unfortunate that in dealing with a "modern approach" to data analysis, the book consistently leaves out exploratory analyses prior to fitting models to data. S/S-PLUS provide extensive tools for this, e.g., `plot.factor`, `plot.design`, `interaction.plot`, but these are not even mentioned. In the low birth weight example they consider 9 predictors, and one would expect some to be highly correlated, but there's no preliminary exploratory analyses conducted on this set. In practice one would like to explore the data, and together with subject-matter intuition and knowledge, generate plausible models to fit. Probably this is not an issue for advanced readers, but others could greatly benefit from Profs. Venables and Ripley's experience. If analyses preliminary to fitting models were left out due to lack of space, then perhaps some analyses could be outlined as exercises?

The book (first printing) seems to suffer a larger-than-usual number of typos, and lines of computer output missing in a couple of places. The index could also be improved.

In summary, the authors successfully present in one book material currently disseminated in various sources: They cover the fundamentals of the S language, and they illustrate how to carry out numerous statistical techniques in the S/S-PLUS environment (including some techniques, like spatial statistics, not covered anywhere else). Since the book covers many statistical methods, it sacrifices some depth for breath. Overall, many Technometrics readers may find this book useful, even though it does not cover the S-PLUS material on statistical process control. The software they provide is equally useful and seems to be robust.

## References


Becker, R. A., Chambers, J. W. & Wilks, A. R. (1988). *The New S Language*, Wadsworth, Pacific Grove, California.

Chambers, J. W., Cleveland, W. S., Kleiner, B. & Tukey, P. A. (1983). *Graphical Methods for Data Analysis*, Wadsworth, Pacific Grove, California.

Chambers, J. W. & Hastie, T. (eds) (1992). *Statistical Models in S*, Wadsworth, Pacific Grove, California.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scaterplots, *Journal of the American Statistical Association* **74**: 829–836.

Everitt, B. S. (1994). *A Handbook of Statistical Analyses using S-PLUS*, Chapman & Hall, London.

Freeny, A. E. & Nair, V. N. (1994). Methods for assessing distributional assumptions in one and two sample problems, *in* J. Stanford & S. Vardeman (eds), *Probabilistic and Statistical Methods in the Physical Sciences*, Academic Press, New York, chapter 7.

Mosteller, F. & Tukey, J. W. (1977). *Data Analysis and Regression: a Second Course in Statistics*, Addison-Wesley, Reading, Massachusettes.

Spector, P. (1994). *An Introduction to S and S-plus*, Duxbury, Belmont, CA.

Stone, C. J. (1977). Consistent nonparametric regression, *The Annals of Statistics* **5**: 595–620.

Tierney, L. (1990). *LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics*, Wiley, New York.