Modern Data Analysis in S-PLUS

References and On-line information

The main reference is Venables & Ripley (1999) and its on-line complements on statistics available at several sites via

`http://www.stats.ox.ac.uk/pub/MASS3/sites.html`

The scripts used in the demos will be at

`http://www.stats.ox.ac.uk/pub/bdr/NewOrleans99/`

References

This is an overview: in particular the books will give references to the primary sources.

Density estimation

Simonoff (1996) provides an excellent overview of density estimation, with sections on computational resources including S-PLUS functions.

Bowman & Azzalini (1997) provide a gentle introduction to (kernel) density estimation, and provide (on-line) an S-PLUS library with many examples.

Local likelihood methods are discussed in a (recent / forthcoming) book by Loader.

There are several closely-related proposals (see Simonoff, 1996, pp. 67–70, 90–92) for parametric fitting of smooth functions to log-densities. `logspline` implements that by Kooperberg & Stone (1992), although a later version described in Stone *et al.* (1997) is promised to replace it.

Green & Silverman (1994) provide a fairly gentle introduction to splines.

Visualization

Ripley (1996, Chapter 9) gives an overview of multivariate visualization methods, including Sammon mapping (Sammon, 1969), multidimensional scaling (Shepard, 1962a,b; Kruskal, 1964a,b) and exploratory pursuit regression (Kruskal, 1969; Friedman & Tukey, 1974; Huber, 1985; Friedman, 1987; Jones & Sibson, 1987).

Regression and Beyond

The Scottish Hill races dataset is discussed in Atkinson (1988) and Staudte & Sheather (1990) as well as Venables & Ripley (1999). It is interesting to see how much those authors missed.

Marazzi (1993) is probably the best source for robust and resistant regression. `lmRobMM` is a wrapper for his FORTRAN code.

Davison & Hinkley (1997) is by far the best practical guide to the bootstrap.

The CPUs data come from Ein-Dor & Feldmesser (1987): I use it as a case study of how *not* to do regression. The dangers of a simple Box–Cox transformation came from a remark to us by a senior British statistician who should have known better.

Hastie & Tibshirani (1990) is the most extensive reference for additive models, and covers `bruto`. MARS is developed in Friedman (1991). ACE is by Breiman & Friedman (1985) and AVAS by Tibshirani (1988).

Random Effects and Mixed Models

There is an extensive literature, much of it not best read by beginners. Diggle *et al.* (1994) is a good overview, and Davidian & Giltinan (1995) perhaps the best source for technical details (but needs to be read very carefully to avoid confusion).

The gasoline data are from Prater (1956) and the rabbits' blood pressure data are discussed in Ludbrook (1994).

Survival Analysis

Kernel smoothing as implemented in `muhaz` is from Mueller & Wang (1994): there are many other variants.

HEFT is from Kooperberg *et al.* (1995).

The `hazcov` plot is by the software of Gray (1994, 1996).

P-splines are expuonded by Eilers & Marx (1996), and used by Therneau's `survival5` code, available in S-PLUS 2000.

Multivariate Analysis

Correspondence analysis goes back to Fisher (1940) at least. Greenacre (1992); Gower & Hand (1996) provide good reviews.

The theory and formulae for discriminant analysis are all in Ripley (1996) in sections 2.4, 2.5 and 3.1.

The Iris dataset is (of course) from Fisher (1936): the Cushing's syndrome one from Aitchison & Dunsmore (1975, Tables 11.1–3).

References

Aitchison, J. and Dunsmore, I. R. (1975) *Statistical Prediction Analysis*. Cambridge: Cambridge University Press.

Atkinson, A. C. (1988) Transformations unmasked. *Technometrics* **30**, 311–318.

Bowman, A. and Azzalini, A. (1997) *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford: Oxford University Press.

Breiman, L. and Friedman, J. H. (1985) Estimating optimal transformations for multiple regression and correlations (with discussion). *Journal of the American Statistical Association* **80**, 580–619.

Davidian, M. and Giltinan, D. M. (1995) *Nonlinear Models for Repeated Measurement Data*. London: Chapman & Hall.

Davison, A. C. and Hinkley, D. V. (1997) *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

Diggle, P. J., Liang, K.-Y. and Zeger, S. L. (1994) *Analysis of Longitudinal Data*. Oxford: Clarendon Press.

Eilers, P. H. and Marx, B. D. (1996) Flexible smoothing with B-splines and penalties. *Statistical Science* **11**, 89–121.

Ein-Dor, P. and Feldmesser, J. (1987) Attributes of the performance of central processing units: A relative performance prediction model. *Communications of the ACM* **30**, 308–317.

Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics (London)* **7**, 179–188.

Fisher, R. A. (1940) The precision of discriminant functions. *Annals of Eugenics (London)* **10**, 422–429.

Friedman, J. H. (1987) Exploratory projection pursuit. *Journal of the American Statistical Association* **82**, 249–266.

Friedman, J. H. (1991) Multivariate adaptive regression splines (with discussion). *Annals of Statistics* **19**, 1–141.

Friedman, J. H. and Tukey, J. W. (1974) A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers* **23**, 881–890.

Gower, J. C. and Hand, D. J. (1996) *Biplots*. London: Chapman & Hall.

Gray, R. J. (1994) Hazard estimation with covariates: algorithms for direct estimation, local scoring and backfitting. Technical Report 784Z, Dana-Farber Cancer Institute, Division of Biostatistics. [Available from `ftp://farber.harvard.edu/stats/gray/784Z.ps.Z`].

Gray, R. J. (1996) Hazard rate regression using ordinary nonparametric regression smoothers. *J. Comp. Graph. Statist.* **5**, 190–207.

Green, P. J. and Silverman, B. W. (1994) *Nonparametric Regression and Generalized Linear Models. A Roughness Penalty Approach*. London: Chapman & Hall.

Greenacre, M. (1992) Correspondence analysis in medical research. *Statistical Methods in Medical Research* **1**, 97–117.

Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models*. London: Chapman & Hall.

Huber, P. J. (1985) Projection pursuit (with discussion). *Annals of Statistics* **13**, 435–525.

Jones, M. C. and Sibson, R. (1987) What is projection pursuit? (with discussion). *Journal of the Royal Statistical Society A* **150**, 1–36.

Kooperberg, C. and Stone, C. J. (1992) Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics* **1**, 301–328.

Kooperberg, C., Stone, C. J. and Truong, Y. K. (1995) Hazard regression. *J. Amer. Statist. Assoc.* **90**, 78–94.

Kruskal, J. B. (1964a) Multidimensional scaling by optimizing goodness-of-fit to a nonmetric hypothesis. *Psychometrika* **29**, 1–29.

Kruskal, J. B. (1964b) Non-metric multidimensional scaling: a numerical method. *Psychometrika* **29**, 115–129.

Kruskal, J. B. (1969) Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new 'index of condensation'. In *Statistical Computation*, eds R. C. Milton and J. A. Nelder, pp. 427–440. New York: Academic Press.

Ludbrook, J. (1994) Repeated measurements and multiple comparisons in cardiovascular research. *Cardiovascular Research* **28**, 303–311.

Marazzi, A. (1993) *Algorithms, Routines and* S *Functions for Robust Statistics*. Pacific Grove, CA: Wadsworth and Brooks/Cole.

Mueller, H. G. and Wang, J. L. (1994) Hazard rates estimation under random censoring with varying kernels and bandwidths. *Biometrics* **50**, 61–76.

Prater, N. H. (1956) Estimate gasoline yields from crudes. *Petroleum Refiner* **35**, 236–238.

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press.

Sammon, J. W. (1969) A non-linear mapping for data structure analysis. *IEEE Transactions on Computers* **C-18**, 401–409.

Shepard, R. N. (1962a) The analysis of proximities: multidimensional scaling with an unknown distance function I. *Psychometrika* **27**, 125–139.

Shepard, R. N. (1962b) The analysis of proximities: multidimensional scaling with an unknown distance function II. *Psychometrika* **27**, 219–246.

Simonoff, J. S. (1996) *Smoothing Methods in Statistics*. New York: Springer-Verlag.

Staudte, R. G. and Sheather, S. J. (1990) *Robust Estimation and Testing*. New York: John Wiley and Sons.

Stone, C. J., Hansen, M., Kooperberg, C. and Truong, Y. K. (1997) Polynomial splines and their tensor products in extended linear modelling. *Annals of Statistics* **25**, 1371–1470.

Tibshirani, R. (1988) Estimating transformations for regression via additivity and variance stabilization. *Journal of the American Statistical Association* **83**, 394–405.

Venables, W. N. and Ripley, B. D. (1999) *Modern Applied Statistics with S-PLUS*. Third Edition. New York: Springer-Verlag. (First edition 1994, second edition 1997.).