Journal of Official Statistics
Vol. 7. No. 1, 1991, pp. 127–133

Book Reviews

Books for review are to be sent to the Book Review Editor Gösta Forsman, Department of Mathematics, University of Linköping, S-581 83 Linköping, Sweden.

Aitkin, M., Anderson, D., Francis, B., and Hinde, J., Statistical Modelling in GLIM. Oxford University Press, Oxford, 1989. ISBN 0-19-852204-5; ISBN 0-19-852203-7. (Pbk) xi + 374pp., £35.

The publication of Nelder and Wedderburn's famous paper (1972) concerning the theory on which the analysis of generalized linear models (GLM) is based, has stimulated, during recent years, many other papers on the same subject. Each of these papers aims at developing specific aspects regarding GLM. The development of many of the aspects of GLM (fitting, validation, model criticism, parametric link function, etc.), and of both the theory and applications is also largely due to the GLIM package, i.e., the statistical package for Generalized Linear Interactive Modelling developed by the Working Party of the Royal Statistical Society.

The first version of this package dates back to 1973. Only in 1983 with McCullagh and Nelder's book (the second more up to date edition was published in 1989) was an organic treatment reached, including the treatment of every problem relating to GLM. The book by Aitkin, Anderson, Francis, and Hinde supports, extends, and completes many of the theoretical aspects already dealt with in McCullagh and Nelder, which is often cited by Aitkin et al. The title of this book suggests a two fold objective: (a) to give an exposition of the principles of statistical modelling and (b) to describe the application of these principles using GLIM. The first objective is reached by detailed discussions of the examples, with great emphasis on the direct use of the likelihood function for inference and the computing systems used for model fitting. All of the examples are analyzed using GLIM.

The GLIM instructions (when brief) are also presented with the analyzed material in the text. The instructions appear in the Appendix, in the form of a MACRO, when they are of general interest and applicability or used in several contexts in the book.

Chapter 1 and the numerous examples fulfil the authors' goal of presenting practical applications of GLIM. Chapter 1 is particularly useful for new GLIM users since it guides them more easily than a manual through the characteristics and structure of the package. Nonetheless, this chapter can also be interesting for people who already use GLIM and wish to deepen

their knowledge of it. The chapter displays the possibilities of fitting different models in GLIM. Moreover, applications to different types of data show the various possibilities of data handling and function evaluation enhancing the flexibility of the reading, displaying, and graphical facilities in GLIM. The part devoted to MACRO text and files handling, sorting, and tabulation of data is also extensive.

The following chapters introduce non-standard procedures for solving specific problems and show the versatility of the package in contexts other than the classical statistical modelling. Chapter 2 includes a general introduction to the principles of statistical modelling with simple applications to different types of variables (categorical, continuous, discrete counts, etc.). Maximum likelihood estimation and likelihood ratio testing are then presented together with a discussion on model fitting strategies (choice of a model and selection of a model among different fitted models). A quite large part is devoted to the different aspects of model criticism: misspecification of the probability distribution, misspecification of the link function, and occurrence of outliers or influential observations. Chapter 3 presents the problems concerning model fitting when the dependent variable is normally distributed: regression analysis for prediction, calibration, and ANOVA. In particular, the authors consider the normal distribution as a member of the Box-Cox transformation family. In this way concepts such as profile likelihood function and parametric link function are easily introduced. These concepts are then examined thoroughly in the rest of the book.

Chapter 4 deals with the analyses of binomial response data which is discussed using different transformations and link functions. The use of the binomial model for analyzing categorical explanatory variables and contingency tables is shown.

Particular attention is given to the issue of model criticism for binary data and to the use of the profile and conditional likelihood in 2×2 tables. Topics which have recently been of great interest in the literature are also briefly discussed such as over-dispersion and effects of variable omission.

Many of the methodologies presented in these chapters are reconsidered and developed in Chapter 5, where the relationship between multinomial and Poisson distributions is investigated.

The fitting of a multinomial logit model with ordered response categories and the related problems of common slopes for the regression linear trend over response categories are discussed.

Chapter 6 deals with survival analysis. GLIM allows the statistical analysis of survival data only by means of specifically written MACROs. This requires both a deep theoretical knowledge of the statistical methods and a good level of expertise with the GLIM language. As a consequence this chapter is very interesting from a pedagogical point of view.

The authors show the reader how to write the likelihood function in the presence of censored observations both to produce the Kaplan-Meier estimators, and to fit parametric and semiparametric models (Gamma, Weibull, Extreme value distribution, Cox proportional hazard model, etc.). Also, problems concerning survival analysis when the failure may be from one of several causes (competing risks), and when the explanatory variables are not fixed in time (time-dependent variables), are briefly mentioned.

The mathematics which was not necessary for understanding the topics presented in the chapters is wisely postponed to Appendix 1.

Appendices 2 and 3, containing concisely the characteristics of the GLIM directives and system defined structure, could perhaps be omitted.

Appendix 4 lists the datasets and the GLIM MACROs used throughout the book. This is very useful, both as a pedagogical support for the teacher and as a source of data for the researcher who needs to test new procedures on well known data sets.

This is a book suitable for an advanced course in applied statistics. It could also be suggested for non-GLIM users who are interested in understanding the role of models in data analysis.

The book is a result of various revisions

of the material used in the courses on Statistical Modelling held at the Centre of Applied Statistics of Lancaster from 1982–87, and consequently is an excellent pedagogical tool.

References

McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models, Second edition. London: Chapman and Hall.

Nelder, J.A. and Wedderburn, R.W.M. (1972). Generalized Linear Models. Journal of the Royal Statistical Society, ser. A, 135, 370–384.

Adriano Decarli
Università degli studi di Milano
Milano
Italy

Gastwirth, J.L., Statistical Reasoning in Law and Policy, Volumes 1 and 2. Academic Press, New York, N.Y., 1988. Vol. 1: ISBN 0-12-277160-5. xxii + 20 (index) + 466pp. Vol. 2: ISBN 0-12-277161-3. xiii + 29 (index) + 459pp., $84.50 (vol. 1 + 2).

According to the author, this two volume set is "designed to serve as a text book for students of law and public policy as well as a resource for individuals who are engaged in the planning, analysis and interpretation of statistical studies or in the assessment of the validity of the results of such studies."

Volume 1 (Chapters 1–8) is entitled "Statistical Concepts and Issues of Fairness." Volume 2 (Chapters 9–14) is entitled "Tort Law, Evidence, and Health." Each volume is close to 500 pages. Each chapter (and many sections) is followed by a section on "Problems" (generally two or three exercises followed by answers to one or more exercises); "Notes," which are elaborations of comments made in the text, or relationships of ideas mentioned in the text to specific cases; "References" to specific citations made in that chapter (references are separately listed by "books," and "articles"). Chapter 12 is also followed by a listing of "cases," a listing in which methodology discussed in that chapter was utilized (brief summaries of each of the cases is also provided). Each volume culminates with a Case Index, a Name Index, and a Subject Index.

The books provide an unusual integration of statistical theory and application to the legal setting. A listing of chapter topics by methodology as determined by this reviewer, is given below.

Chapter 1 – Descriptive statistics
Chapter 2 – Probability and random variables
Chapter 3 – Estimation and hypothesis testing: binomial and normal distributions
Chapter 4 – Applications of the binomial distribution to the legal setting
Chapter 5 – Statistical procedures for comparing sample proportions
Chapter 6 – Applications of Chapter 5 to the legal setting
Chapter 7 – Comparing two distributions
Chapter 8 – Correlation and regression
Chapter 9 – Sample survey methods
Chapter 10 – Poisson models for rare events
Chapter 11 – Non-parametric methods
Chapter 12 – Bayesian methods
Chapter 13 – Statistical methods used in medicine: clinical trials, competing risks, observational studies
Chapter 14 – Applications of health related statistical methods to the legal setting.

An interesting thing about these chapters is that while I have classified them methodologically, the statistical procedures discussed in the chapters are woven into the text of legal cases in ways that make the flow of statistical ideas very natural.

For example, in Chapter 5 where the author discusses comparing sample proportions, he makes the point that sometimes we are comparing proportions from independent samples in 2 × 2 tables, and other times we are comparing non-independent sample proportions. In the former case we are comparing independent binomial proportions whereas in the latter case he con-

siders the hypergeometric distribution as having generated the 2 × 2 table. For the latter case he takes up "Fisher's exact test." He then discusses the actual case of "Johnson vs. Perini" (No. 76-2259, D.C.D.C. June 1, 1978). There, "concern was whether blacks were discharged at a higher rate than whites" (discharged from an organization). He gives the data shown below.

	Terminated (cause)	Not Terminated	Total	Fraction Terminated
Black	1	8	9	0.111
White	1	46	47	0.021
Total	2	54	56	0.036

Using Fisher's exact test the author finds that the probability of the observed data is 0.2747, and the probability of any more extreme tables is 0.023, so the overall probability of observing as many or more blacks terminated for cause than in the actual data is 0.2747 + 0.023 = 0.298, which is not close to significance. (Of course the sample sizes of those, black or white, who were terminated are very small, raising some suspicions about reliability. Note also that there is a typo in the text here.) The point, however, is that examples like this are an integral part of the text throughout the two volumes, as compared with legal examples inserted to illustrate methodology. The books are written to appeal to the practitioner trying to bring quantitative methods to bear on legal issues; rather than to appeal to statisticians looking for applications of particular methods.

The volumes make ample reference to germane texts on statistical methodology, such as Fleiss (1981), "Statistical Methods for Rates and Proportions," and Siegel (1956), "Non-parametric Statistics for the Behavioral Sciences," and to other books relating statistics to law, such as Baldus and Cole (1980), "Statistical Proof of Discrimination," Connolly and Peterson (1980), "Use of Statistics in Equal Employment Opportunity Litigation," and Finkelstein (1978), "Quantitative Methods in the Law."

This delightfully easy-to-read text does an excellent job of presenting methods of statistical reasoning to the legal practitioner, which is all its title claims to do. On occasion, it does not provide sufficient detail for a practitioner to actually understand how to analyze an important problem. For example, Chapter 12, Volume 2 takes up the forensic statistics issue of DNA fingerprinting (pp. 698–699). But here, the treatment of this new and exciting approach to determining lack of paternity, or innocence of any accused person whose DNA in body fluids did not match those of the victim is not sufficient for the reader to understand where the statistical problems really are, or how to carry out appropriate statistical tests.

It would also not be very easy for the legal practitioner to carry out procedures which are based upon methods of Bayesian inference, by using Chapter 12 alone. In fact the author suggests that in some situations "some prior knowledge should not be used," and that "the Bayesian approach is most useful when one knows a substantial amount about the process generating the data and can formulate a sound prior distribution" (p. 738). In fact, available prior information should always be used, but one must think carefully about whether the appropriate constraints have been imposed upon the prior distribution. Moreover, the Bayesian approach may be used to great advantage even when very little information is available about the data generating mechanism, and perhaps in such situations problems can be formulated for legal purposes in far more effective ways than could otherwise be accomplished. For example, a prior distri-

bution for log odds of an event (such as whether discrimination was practiced by some firm) should be centered about zero, to establish a neutral position for the court. This is the case of minimal prior information. The further this value is from zero the greater is the weight of evidence in favor of, or against the discrimination allegation.

While this reviewer would have preferred the author to recommend a broader basis of statistical inference, the text is very well written and should be considered as a welcome reference text for all statisticians who serve as expert witnesses in court, or who serve as consultants to the legal profession.

S. James Press
University of California
Riverside, CA
U.S.A.

Hosmer, D.W. Jr. and Lemeshow, S., Applied Logistic Regression. John Wiley & Sons, New York, 1989, ISBN 0-471615536. xiii + 307pp., £35.50.

Contents:

1. Introduction to the logistic regression model
2. The multiple logistic regression model
3. Interpretation of the coefficients of the logistic regression model
4. Model-building strategies and methods for logistic regression
5. Assessing the fit of the model
6. Application of logistic regression with different sampling models
7. Logistic regression for matched case-control studies
8. Special topics

Readership: Graduate students in biostatistics and epidemiology, and applied statisticians

In the 1960s there appeared a seminal paper in which logistic regression was applied in the analysis of data from a heart study. Since then logistic regression has become the standard method for regression analysis of dichotomous data in many fields, particularly in the health sciences. The inclusion of logistic regression routines in major computer software packages such as GLIM, SPSS, BMDP, and SAS has led to widespread use of the method. (Important differences between these packages, when noted, are reported in the book for the reader's benefit.) Extensive data sets from epidemiological research are given in an appendix. These data sets serve as the source of the numerous illustrative examples and the exercises accompanying each chapter. The interested reader with access to a software package having logistic regression capabilities should have no difficulty in reproducing the analyses presented in the text.

The regression perspective is used throughout in the book (i.e., not the contingency table point of view). The reader is assumed to be familiar with regression analysis and Mantel-Haenzel methods in contingency table analysis. Mathematical and statistical subtleties are avoided and the book should serve as a straightforward introduction to the subject.

I find particularly commendable the importance placed by the authors on modelling strategies, parameter interpretation, and, in particular, assessing model adequacy. The intention is to enable readers to understand what they are doing when modelling the relationship between a dichotomous response (dependent) variable and a set of covariates. Slope coefficients in linear regression have a relatively straightforward interpretation, whereas proper interpretation of coefficients in a logistic model depends on being able to place meaning on the difference between logits. Therefore, the authors discuss substantive interpretation of such differences on a case-by-case basis.

Sven Berg
University of Lund
Lund
Sweden

Kaufman, L. and Rousseauw, P.J., Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, Interscience Publication, New York, 1990. ISBN 0-471-87876-6, xiv + 342 pp., $59.95.

If you were a young statistician in the early 70s the area to be in was Cluster Analysis. Researchers in disciplines ranging from psychology to archeology and from anthropology to astronomy were more than willing to apply existing and newly developed clustering methods to their data in the hope of uncovering those mythical qualities "structure" or "pattern," but perhaps even more in the hope of publishing the first application of this increasingly fashionable methodology in the literature of their particular subject. Papers on cluster analysis appeared at the rate of somewhere close to 1000 per year, several books on the topic were published, and the first critical reviews by increasingly sceptical statisticians were seen.

By the 80s the bubble had all but burst. Accounts of cluster analysis, both theoretical and applied were few and far between, and a greater sense of realism as to what could be achieved by the methods seemed to prevail. Consequently I was a little surprised to find a new book on the subject appearing in the first year of the 90s. According to its authors, *Finding Groups in Data*, is an applied book for the general user, and their aim is to make cluster analysis available to people who do not necessarily have a strong mathematical or statistical background. In many respects they succeed in this aim; the book contains no difficult mathematics and their descriptions of all the familiar clustering techniques such as k-means, Ward's method, and group-average are clear and concise. Each chapter also has a set of exercises which should be useful to students as well as applied researchers.

Unfortunately the authors fall into the trap, only too common these days, of trying to publish a description of their clustering programs, in parallel with their exposition of clustering methodology. Sections describing how to use the programs "insert the floppy disk in drive A and type A: DAISY" belong in a programming manual *not* a book such as this. (Incidentally all the programs are female; as well as DAISY there are PAM, FANNY, AGNES, DIANA and CLARA – a curious piece of sexism?) The programs themselves implement a number of interesting methods, for example, CLARA is a relatively simple adaptation of the usual k-means algorithm, but can be applied to large data sets, and FANNY which instead of creating a simple partition of a data set, assigns membership coefficients to create a "fuzzy" clustering. The programs are written for PCs and will at some future date be incorporated into the S-Plus system.

The opening sentence of Chapter 1 of Kaufmann and Rosseauw's book begins, "Cluster analysis is the art . . . "; and this is, of course, one of the problems since, like many other artforms, the beauty is often only in the eye of the beholder. One person's view of a data set, that it consists of a number of interesting "clusters," is only too frequently seen by others, as an optimistic interpretation of what they consider random scatter. This is reflected in the author's advice for choosing the number of clusters given on page 38, "retain the clustering that appears to give rise to the most meaningful interpretation." Unfortunately, it is only too easy for over enthusiastic researchers, convinced a priori of the presence of clusters in their data, to be able to give a "meaningful interpretation" to almost any solution that some clustering algorithm produces. In this text scant attention is paid to this problem, which is a serious ommission.

In many respects *Finding Groups in Data* is an attractively produced book, and it would clearly be fun to have the programs on one's own PC. But as an introduction to cluster analysis and the associated possible pitfalls it perhaps does less well than some of the texts produced in the 70s.

Brian S. Everitt
Institute of Psychiatry
London
England

Ross, S.M., Introduction to Probability Models. Fourth Edition, Academic Press, San Diego, CA, 1989. ISBN 0-12-598464-2, xiv + 514 pp., $44.50

1. Introduction to Probability Theory
2. Random Variables
3. Conditional Probability and Conditional Expectation
4. Markov Chains
5. The Exponential Distribution and the Poisson Process
6. Continuous-Time Markov Chains
7. Renewal Theory and Its Applications
8. Queueing Theory
9. Reliability Theory
10. Brownian Motion and Stationary Processes
11. Simulation.

For those readers familiar with the third edition, the author's preface indicates that:

> For the most part, the fourth edition is quite similar in spirit to the third. It differs from the third primarily by the addition of material in the chapters on discrete time Markov chains, continuous time Markov chains, renewal theory, and simulation. For instance, examples concerning communication protocols and waiting times for patterns were added to the Markov chain chapter, as was some additional material on the reverse chain. The chapter on continuous time Markov chains includes additional material on birth and death processes as well as new material on computing the transition probability function. The renewal theory chapter includes, among other new things, a section on computing the renewal function, an application of results from the 2-state continuous time Markov chain to compute the renewal function when the interarrival distribution is the convolution or the mixture of two exponentials, and an application to quality control. There is additional material in the Simulation chapter concerning ways of efficiently generating nonhomogeneous Poisson processes, and the use of "conditioning" to reduce the variance when trying to estimate the average customer delay in a variety of queueing models.

Ross's text is well written and easy to read. Sometimes it is even fun to read. His approach is traditional, which is to say that it is dull. It does as well or better than any other available text at accomplishing the objective the author enunciates in the Preface:

> This text is intended as an introduction to elementary probability theory and stochastic processes. It is particularly well-suited for those wanting to see how probability theory can be applied to the study of phenomena in fields such as engineering, management science, the physical and social sciences, and operations research.

But it does so using artificial examples. Has probability theory never been applied in the study of real phenomena? Of course it has. So why do we have to talk about Max and Patty flipping pennies? And why do examples that deal with the likes of telegraph signal processes and managers of supermarkets seem so fake?

Most students learn by example. Some students are better than others in this regard. An engineer designing a spare parts replacement program may see that something she once studied in a probability course is relevant, but I am not optimistic. I would be more optimistic if the course used *actual* case studies from engineering. This optimism is based on my (weakly held) feeling that it is easier for students to see the relevance of real examples, and on my (strongly held) opinion that students learn better when presented with real problems.

In a purely mathematics course with no claim toward applications, a course based on this text would be fine. I would also use this book for a course in applications, but I would want to supplement it with real problems taken from the substantive disciplines of engineering, business, science, etc.

Donald A. Berry
University of Minnesota
Minneapolis
U.S.A.

Printed in March 1991.