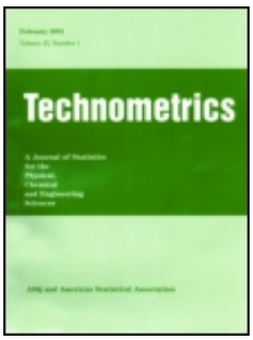

Technometrics

ISSN: 0040-1706 (Print) 1537-2723 (Online) Journal homepage: https://www.tandfonline.com/loi/utch20

Regression and ANOVA: An Integrated Approach Using SAS Software

Richard F Gunst

To cite this article: Richard F Gunst (2003) Regression and ANOVA: An Integrated Approach Using SAS Software, Technometrics, 45:2, 170-171, DOI: 10.1198/tech.2003.s159

To link to this article: https://doi.org/10.1198/tech.2003.s159

Published online: 01 Jan 2012.

Regression Modeling Strategies, by Frank E. HARRELL, Jr., New York: Springer-Verlag, 2001, ISBN 0-387-95232-2, xxii + 568 pp., $79.95.

The Preface of this book starts in a very promising way by stating: "There are many books that are excellent sources of knowledge about individual statistical tools (survival models, general linear models, etc.), but the art of data analysis is about choosing and using multiple tools." The author goes on to say that "this argues the need for a better balance in the literature and in statistical teaching between *techniques* and problem solving *strategies*." Most statisticians (including myself) would agree with the author. Unfortunately, strategies are more difficult to teach than tools. This is because effective strategies are learned by and large through experience, not by taking a course or reading a book. The other difficulty in teaching and writing about strategies is that, as the author notes, data analysis is as much an art as it is a science.

The book comprises of 568 pages distributed in 20 chapters:

1. Introduction
2. General Aspects of Fitting Regression Models
3. Missing Data
4. Multivariate Modeling Techniques
5. Resampling, Validating, Describing, and Simplifying the Model
6. S-PLUS Software
7. Case Study in Least Squares Fitting and Interpretation of a Linear Model
8. Case Study in Imputation and Data Reduction
9. Overview of Maximum Likelihood Estimation
10. Binary Logistic Regression
11. Logistic Model Case Study 1: Predicting Cause of Death
12. Logistic Model Case Study 2: Survival of Titanic Passengers
13. Ordinal Logistic Regression
14. Case Study in Ordinal Regression, Data Reduction, and Penalization
15. Model Using Nonparametric Transformations of X and Y
16. Introduction to Survival Analysis
17. Parametric Survival Models
18. Case Study in Parametric Survival Modeling and Model Approximation
19. Cox Proportional Hazards Regression Model
20. Case Study in Cox Regression

The material seems to be divided into two parts. Chapters 1 through 5 are of a general nature and are not linked to any statistical software. These chapters discuss practical issues in the modern analysis of data, such as missing data, data reduction, and the bootstrap. The author covers a lot of ground, and the pace could be intimidating to someone without a solid knowledge of generalized linear models and experience in their application.

The rest of the book makes heavy use of S-PLUS. Chapter 6 includes a very short description to S-PLUS and introduces a library of S functions called Design that is used frequently throughout the rest of the book. The seven extensive case studies included are analyzed entirely with S-PLUS, and a list of S-PLUS functions is included at the end of every chapter. All the datasets are available from the author's website, so the results can be easily duplicated, particularly by S-PLUS users with some experience. Those not well versed in S-PLUS will have a harder time using these techniques, because not all are readily available in other languages.

Thirteen of the chapters have a list of interesting problems that can be used in the classroom. The reference list of 466 entries is impressive and up to date. A useful feature throughout the book is the numbers in the margins referring to annotations at the end of each chapter. These are very useful for identifying references in specific areas. This book is an ambitious, and mostly successful, attempt to disseminate effective strategies for the use of regression techniques. Many of the examples are from the medical area, in which the author has worked for many years and has accumulated a wealth of experience. It is written in a clear and direct style. However, the amount of information can sometimes be overwhelming. This book can be frustrating for someone seeking to learn new techniques. The author's intention was to write not about techniques, but rather about effective strategies to apply them. This text could be used in a graduate-level course directed to students with a strong background in applied statistical modeling.

Regression Modeling Strategies is definitely a valuable reference for modern applications of commonly used regression techniques. Data analysts, particularly users of S-PLUS, with experience in the application of these tools will benefit the most from this book.

Esteban WALKER
University of Tennessee

Regression and ANOVA: An Integrated Approach Using SAS Software, by Keith E. MULLER and Bethel A. FETTERMAN, Cary, NC: SAS Institute, 2002, ISBN 1-58025-890-5, xi + 565 pp., $65.95.

A daunting task awaits any authors who attempt to write a book on the broad topics of regression and analysis of variance (ANOVA). Choices must be made on the topical coverage, theoretical details, breadth of examples, and target audience. The authors of this book have chosen to emphasize the specification of linear fixed-effects models in both uncontrolled (regression) and controlled (design of experiments) settings, theoretical results without proofs, clinical/biostatistical applications, and an audience of graduate statistics students and statistical consultants.

The authors use this text for an intermediate linear models course. As such, its strengths in alternative model specifications of fixed effects and extensive use of matrix properties, both full rank and less than full rank, are noteworthy. Matrix specification of hypotheses, estimators, and test statistics is very well done. Careful use of matrix transformations to relate alternative model specifications is also a strength. Fundamentals of the multivariate normal distribution, linear model assumptions, and distributional properties of least squares estimators for full-rank models and solutions to normal equations for less-than-full-rank models are given in initial chapters and a technical appendix. In addition to its value in a linear models ANOVA course, this text can also be a useful supplement to a regression analysis course or a course on the design and analysis of experiments, provided that these courses stress matrix representations of models and estimators.

As suggested in the previous paragraph, readers of this book must be facile with matrix operations. Although the short technical appendix can serve as a refresher, the heavy emphasis on modeling, estimation, and testing presentations in matrix notation requires that the reader be comfortable with matrix operations and properties. Coupled with this requirement is the benefit of extensive SAS/IML code to perform the matrix operations under discussion. The algebraic matrix manipulations and the IML code reinforce one another.

Missing in the book's topical outline are regression topics, such as scatterplot smoothing, robust model fitting, nonparametric modeling, and many contemporary approaches to graphics. Missing from a comprehensive coverage of linear models from designed experiments are nesting, random effects (apart from randomized block designs), a careful discussion of unbalanced and missing-data estimation and testing, and diagnostic issues for designs with a small number of (or no) replicates. Noting these omissions is not intended as a criticism, but only a note on the book's limitations.

Following a short introductory chapter, Chapters 2 and 3 succinctly outline model specification, assumptions, estimation, and distributional properties of estimators and inferential statistics. These chapters provide a quick review of theoretical model and estimation results for those who have already taken courses in ANOVA and regression. SAS/IML code is provided for much of the model specification and estimation formulas.

Chapters 4–11 constitute the coverage of regression modeling, defined to be linear models with interval- or ratio-scaled predictors that have little or no measurement error. The restriction to "continuous" predictors is artificial and serves to suggest a limitation on the use of categorical predictors other than in analysis of covariance models (Chap. 16). Extensive delineation is made between intercept and no-intercept models and between uncorrected versus corrected (for the intercept) sums of squares and ancillary statistics such as R^2. This might reflect the authors' experiences in clinical and biostatistical applications, but the amount of emphasis on intercept versus no-intercept issues seems excessive given other important topics that could have been covered. The order of topics also appears unnatural in the context of

good regression practice. The discussion of regression model specification, (Chap. 4) is immediately followed by hypothesis testing (Chap. 5) and various forms of correlation estimation and testing (Chap. 6). All of this precedes discussion of data and regression diagnostics (Chaps. 7 and 8) and variable transformations (Chap. 10). Coverage of polynomial regression (Chap. 9) and variable selection (Chap. 11) concludes the general treatment of regression modeling.

ANOVA models that result from designed experiments with one or two fixed-effects factors are viewed as special cases of regression models in Chapters 12, 13, 14, and 16. Very extensive and careful discussions of different methods of specifying ANOVA models are shown to provide equivalent results in overall tests of effects, but very different results in comparative tests of individual effects. This treatment is very good. Reference cell coding, cell mean coding, ANOVA coding, and effect coding are all analyzed as special cases of regression models that use indicator (dummy variable) specification of predictors.

Chapter 15 provides an introduction to random-effects models by primarily discussing a model for a randomized complete-block design. The relationship between this ANOVA model and the multivariate normal distributions with a compound-symmetry covariance matrix is very effectively exploited. Chapter 17 concludes the text with a short discussion of power.

The emphasis throughout on the scientific relevance of model specification and inference is a major strength of this text. So too is the extensive use of matrix results to delineate different forms of model specification. The use of general solutions to the normal equations in inferential statistics permits a focus on the key issues in estimation and testing rather than a need to focus on special cases that depend on how main effects and interactions are specified.

The text emphasizes IML code to reinforce the matrix manipulations that are so valuable in understanding the estimation and testing issues. But this has a drawback, because most modern computer programs do not use, for example, matrix inverses to solve systems of equations. At times the text reads as though it was written to support familiarity with IML coding. The authors properly state that the text can be used without regard to the software, including SAS. Although there is adequate use of Proc Reg and Prog GLM in the text, no discussion of issues like type I–IV sums of squares accompanies the various forms of model specification. The emphasis is on using contrast statements to test hypotheses.

An emphasis that is not opportune is the focus on drawing conclusions from regression diagnostics using hypothesis testing. The use of multiple comparisons notwithstanding, most diagnostic procedures in regression and ANOVA are not intended to be restricted to inferential procedures. It is also unfortunate that the text emphasizes minimum Cp statistics for subset selection, rather than recognizing the common occurrence of alternative subsets having Cp values very close to the minimum. The notion of a single "best" subset permeates all of the variable selection methods discussed.

Regression and ANOVA is very good in the areas of strength mentioned. It can be a very useful resource for select topics in regression and ANOVA modeling, even though it is not comprehensive enough in either subject to be used as a stand-alone text. It will be especially valuable for those seeking a thorough treatment of specifying and drawing inferences on fixed-effects linear models.

Richard F. GUNST
Southern Methodist University

Statistical Analysis of Designed Experiments (2nd ed.), by Helge TOUTENBURG, New York: Springer-Verlag, 2002, ISBN 0-387-98798-4, xv + 500 pp., $79.95.

Helge Toutenburg describes this text as a "resource/reference book which contains statistical methods used by researchers in applied areas." It contains 10 chapters:

1. Introduction
2. Comparison of Two Samples
3. The Linear Regression Model
4. Single-Factor Experiments With Fixed and Random Effects
5. More Restrictive Designs (RBD, Latin Squares)
6. Multifactor Experiments
7. Models for Categorical Response Variables
8. Repeated-Measures Model
9. Cross-Over Design
10. Statistical Analysis of Incomplete Data

Three appendices "Matrix Algebra," "Theoretical Proofs," and "Distributions & Tables," are included in an attempt to make the book self-contained.

The writing style is very much that of a reference book. No overarching themes or particular viewpoints infuse the text. Methods suited for nominal, ordinal, and interval data are presented, which means that the book is not (as if often the case) an elaboration on ANOVA. Instead, the author describes nonparametric, categorical, and other methods as needed (and does not mention randomization tests). Output from SAS, S-PLUS, and SPSS is used to illustrate examples, but these are few and in no way do they prepare the reader to use these packages to undertake the analyses described. In most cases the theory is described in a shorthand style that gets to the point without overburdening the reader with mathematical detail; however, the model-based descriptions are rather "mathy" for many practitioners. For the most part, the book reads like an advanced undergraduate text; but there are too few examples and far too few exercises for someone to use this as a text without supplementing the text with some rich examples and exercises. Instructors with experience in DOE will be able to bridge that gap and select topics according to their interests and goals.

A plus is that the author includes thorough discussions of generalized linear models (categorical data analysis) and repeated-measures designs. A minus for those engaged in industrial experimentation is the complete absence of any discussion of fractional factorials, response surfaces, and sequential experimentation. The authors emphasize hypothesis testing as opposed to confidence interval estimation. This reviewer prefers the latter in many cases.

The Preface describes a book that "tries to bridge the gap between applications and theory dealing with designed experiments." Most of the examples presented are from dentistry, medicine, and agriculture (and mention is made of pharmaceutical research). The importance of confirming assumptions is affirmed, but very little is offered to aid the reader. For example, probability plots (normal or otherwise) are not mentioned. In general, there is little reliance on graphical methods, and the notion of data exploration as a means of acquiring knowledge is given short shrift compared with models and formal tests. Consistent with the book's title, little direction is offered to researchers who might wish to design experiments or studies. However, given data, one can find appropriate analysis methods for most cases arising in practice.

Statistical Analysis of Designed Experiments might be a useful, self contained reference for those who want a quick description of the underlying theory and practice for a large assortment of standard DOE problems. Be warned, however, that it is not strictly speaking a "how-to" book, and it requires a level of notional sophistication and comfort that may challenge some practitioners.

Peter WLUDYKA
University of North Florida

Nonparametric Analysis of Longitudinal Data in Factorial Experiments, by Edgar BRUNNER, Sebastian DOMHOF, and Frank LANGER. New York: Wiley, 2002, ISBN 0-471-44166-X, xvii + 261 pp., $94.95.

This book presents nonparametric methods of analyzing the longitudinal data by formulating meaningful effects and hypotheses in factorial experiments. The methods presented are valid even for noncontinuous distribution functions, particularly in treating ties, count data, and ordinal data. The situations with missing data and singular covariance matrices are also considered. Special attention is given to situations in which sample sizes are small. Many examples from human and veterinary medicine, pharmacology, and forestry are given in the Introduction, accompanied by numerous graphical illustrations. The SAS codes and macros in SAS-IML for carrying out analyses using the methods described in the book are included. Internet addresses for downloading the available macros are also listed. Some theoretical backgrounds are also provided for interested readers.

This book is very readable. The first author is a leading researcher in this field and has contributed to the development of the methods described. The book is aimed at the needs of applied statisticians, biometricians, researchers, and students for analyzing the longitudinal data nonparametrically. This book