

Preface

None of us are formally trained as statisticians. We arrived at our common interest in applications of statistics to biomedical data analysis from quite different backgrounds. We each first encountered multiple regression as graduate students. For one of us it was a collaboration with a political science graduate student and volunteer from Common Cause to analyze the effects of expenditures on the outcomes of political campaigns in California* that sparked the interest. For another, it was a need to quantify and unravel the multiple determinants of the way the heart responded to a sudden change in blood pressure.[†] And, for the third, it was graduate coursework exposure to correlation, regression, and analysis of variance and covariance framed as specific cases of the general linear model then applied to a diverse array of disciplines, from psychology to nursing research. This book in applied analysis of real problems, as opposed to developing new statistical theories.

Our experience using the regression-based methods convinced us that these methods are very useful for describing and sorting out complex relationships, particularly when one cannot precisely control all the variables under study. Indeed, it is this capability that has made multiple regression and related statistical techniques so important to modern social science and economic research and, more recently, biomedical and clinical research. These statistical techniques offer powerful tools for understanding complex physiological, clinical, and behavioral problems, particularly when dealing

*Glantz S, Abramowitz A, and Burkart M. Election outcomes: whose money matters? *J Politics*. 1976;38:1033-1038.

†Slinker BK, Campbell KB, Ringo JA, Klavano PA, Robinette JD, and Alexander JE. Mechanical determinants of transient changes in stroke volume. *Am J Physiol*. 1984;246:H435-H447.

with intact organisms, which explains why multiple regression and related methods have become indispensable tools in our research.

The techniques we apply to a wide variety of disciplines in the health and life sciences, social and behavioral sciences, engineering, and economics. Indeed, multiple regression methods are much more widely used in many disciplines other than the health and life sciences. We hope that people working in those other disciplines will also find this book a useful applications-oriented text and reference.

This book also grows out of the very positive experience associated with writing the introductory text—*Primer of Biostatistics* (also published by McGraw-Hill). It has been very gratifying to see that text make introductory statistical methods comprehensible to a broad audience and, as a result, help improve the quality of statistical methods applied to biomedical problems. This book seeks to go the next step and make more advanced multivariable methods accessible to the same audience.

In writing this book, we have tried to maintain the spirit of *Primer of Biostatistics* by concentrating on intuitive and graphical explanations of the underlying statistical principles and using data from actual published research in a wide variety of disciplines to illustrate the points we seek to make. We concentrate on the methods that are of the broadest practical use to investigators in the health and life sciences. We also use real data in the problems at the end of each chapter.

This book should be viewed as a follow-on to *Primer of Biostatistics*, suitable for the individual reader who already has some familiarity with basic statistical terms and concepts, or as a text for a second course in applied statistics.

Multiple regression can be developed elegantly and presented using matrix notation, but few readers among those we hope to reach with this book can be expected to be familiar with matrix algebra. We, therefore, present the central development of the book using only simple algebra and most of the book can be read without resorting to matrix notation. We do, however, include the equivalent relationships in matrix notation (in optional sections or footnotes) for the interested reader and to build the foundation for a few advanced methods—particularly principal components regression in Chapter 5, the full-information maximum likelihood (FIML) estimation in the context of missing data described in Chapter 7, and mixed models estimated via maximum likelihood for repeated-measures data described in Chapter 10—that require matrix notation for complete explanation. We also include an appendix that covers all necessary notation and concepts of matrix algebra necessary to read this book.

This focus on practical applications led us to present multiple regression as a way to deal with complicated experimental designs. Most investigators are familiar with simple (two-variable) linear regression, and this procedure can be easily generalized to multiple regression, which simply

has more independent variables. We base our development of multiple regression on this principle.

Our focus on practical applications also led us to present analysis of variance as a special case of multiple regression. Most presentations of analysis of variance begin with classical developments and only mention regression implementations briefly at the end, if at all. We present traditional development of analysis of variance, but concentrate on the regression implementation of analysis of variance. We do so for two reasons. First, the regression approach permits a development in which the assumptions of the analysis of variance model are much more evident than in traditional formulations. Second, and more important, regression implementations of analysis of variance can easily handle unbalanced experimental designs and missing data. Although one can often design experiments and data collection procedures in other disciplines to ensure balanced experimental designs and no missing data, this situation often does not occur in the health and life sciences, particularly in clinical research.

Practical applications also dictated our selection of the specific topics in this book. We include a detailed discussion of repeated-measures designs—when repeated observations are made on the same experimental subjects—because these designs are very common, yet they are rarely covered in texts at this level. We also include detailed treatment of nonlinear regression models because so many things in biology are nonlinear, and logistic regression and survival analysis because many things in clinical research are measured on dichotomous scales (e.g., dead or alive). We also devote considerable attention to how to be sure that the data fit the assumptions that underlie the methods we describe and how to handle missing data. Like any powerful tools, these methods can be abused, and it is important that people appreciate the limitations and restrictions on their use.

This edition adds an explanation of maximum likelihood estimation, a new chapter on missing data, including multiple imputation, and a completely rewritten presentation of repeated-measures analysis of variance that uses modern maximum likelihood approaches (while keeping the classical ordinary least squares approach as optional sections). Other additions include a discussion of robust variance estimation for regression models in Chapter 4 and how to address small or zero cells in the context of logistic regression and how to model repeatedly measured binary dependent variables in Chapter 12.

In addition to a year-long comprehensive course, this book can be used in a range of intermediate courses taught in the areas of applied regression and analysis of variance. For example,

1. Applied linear models, one semester: Chapters 1–5, 8, and 9.
2. Applied linear models, two semesters: Chapters 1–11.

3. Multiple linear regression: Chapters 1–6, with possible use of Chapters 7, 11, and 12, depending on the course goals.
4. ANOVA (or Experimental Design): Chapters 1, 8–10.
5. Regression with a qualitative dependent variable and survival models: Chapters 12–14.

This book can also be used as an applied reference book and source for self-study for people who, like us, come to more advanced statistical techniques after they have finished their formal training.

The universal availability of computer programs to do the arithmetic necessary for multiple regression and analysis of variance has been the key to making the methods accessible to anyone who wishes to use them. We use four widely available programs, Minitab, SAS, SPSS, and Stata to do the computations associated with all the examples in the book. We present outputs from these programs to illustrate how to read them, and include input control language files or commands, and data sets in the appendices. We also include a comparison of the relevant capabilities of these programs at the time of this writing and some hints on how to use them. This information should help the reader to learn the mechanics of using these (and similar) statistical packages as well as reproduce the results in all the examples in the book. Of course, computer programs are being constantly updated and so the reader is advised to use the most recent versions of statistical computing packages and to carefully review their documentation to learn their capabilities, defaults, and options.

As already mentioned, the examples and problems in this book are based on actual published research from the health and life sciences. We say “based on” because we selected the examples by reviewing journals for studies that would benefit from the use of the techniques we present in this book. Many of the papers we selected did not use the methods we describe or analyze their data as we do. Indeed, one of our main objectives in writing this book is to popularize the methods we present. Few authors present their raw data, and so we worked backwards to simulate what the raw data may have looked like based on the summary statistics (i.e., means and standard deviations) presented in the papers. We have also taken some liberties with the data to simplify the presentation for didactic purposes, generally by altering the sample size or changing a few values to make the problems more manageable. We sought to maintain the integrity of the conclusions the authors reach, except in some cases where the authors would have reached different conclusions using the analytical techniques we present in this book. The bottom line of this discussion is that the examples should be viewed for their didactic values as *examples*, not as precise reproductions of the original work. *Readers interested in the original work should consult the papers we used to generate the examples and problems, and compare the approaches of the original authors with those we illustrate. Keep in*

mind, however, that in most cases, we are dealing with simulated or altered data sets.

We thank our teachers and mentors, including Julien Hoffman, Richard Brand, David Burns, Steve Gregorich, Earl Jennings, Bill Koch, Ken Campbell, "Doc" Stanford, Marty LeWinter, and Mike Patetta. Charles McCulloch supplied helpful suggestions on what to add to the third edition. We are grateful to Estie Hudes for reviewing much of the new content on ANOVA and giving us helpful suggestions for making the material more clear (of course, any errors are our own). We appreciate our colleague Eric Vittinghoff taking the time to educate us on cross-validation approaches and for generously sharing with us his Stata program for performing cross-validation of linear regression models. Thanks also go to Samantha Dilworth Johnson who created the electronic versions of the data sets and tested many of the syntax files which appear in Appendix B and Jonathan Leff for helping prepare the index for the third edition. We also thank those who made raw data available to us to develop some of the examples. Lastly, we thank our families for giving us the time to do this book. It has been gratifying to have the first two editions of this book so well received, and we have high hopes that the third will be a significant force within the health and life sciences community to popularize multivariate statistical techniques, which have proved so powerful for people in the social and economic sciences.

You, the reader, will be the judge on how well we have succeeded.

STANTON A. GLANTZ

BRYAN K. SLINKER

TORSTEN B. NEILANDS