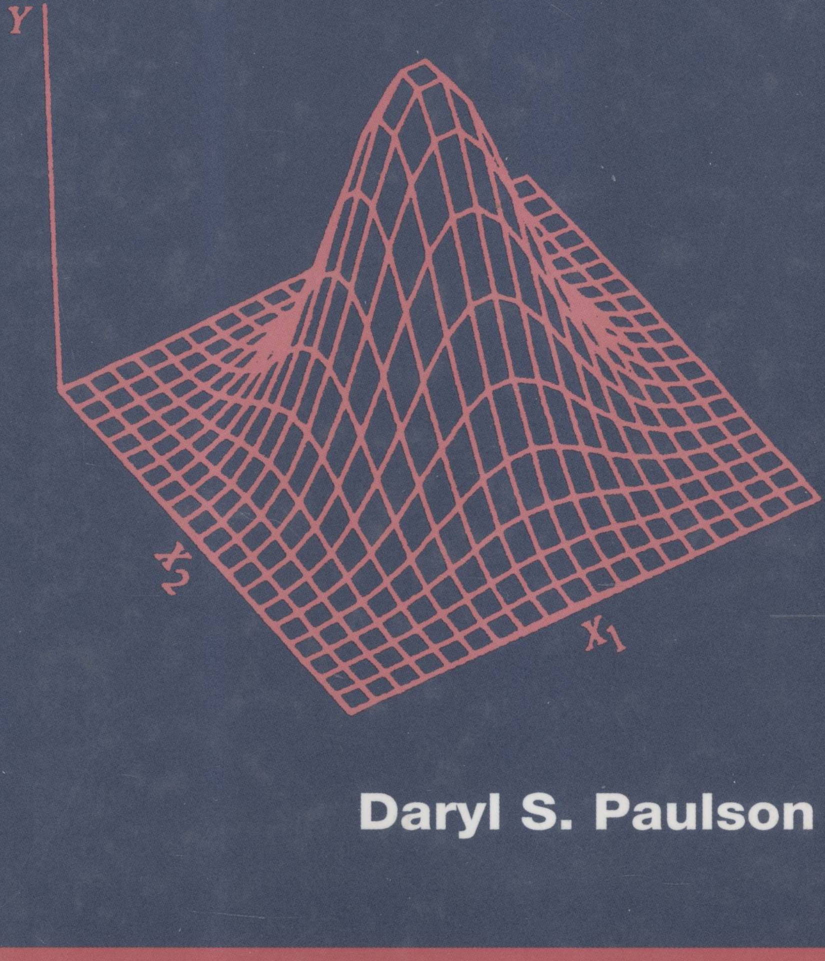
Chapman & Hall/CRC Biostatistics Series

HANDBOOK OF REGRESSION AND MODELING

Applications for the Clinical and Pharmaceutical Industries



Daryl S. Paulson

Chapman & Hall/CRC
Taylor & Francis Group

Chapman & Hall/CRC Biostatistics Series

Handbook of Regression and Modeling

Applications for the Clinical and Pharmaceutical Industries

Daryl S. Paulson
BioScience Laboratories, Inc.
Bozeman, Montana, U.S.A.

Chapman & Hall/CRC
Taylor & Francis Group
Boca Raton London New York

Chapman & Hall/CRC is an imprint of the
Taylor & Francis Group, an informa business

Chapman & Hall/CRC
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

International Standard Book Number-10: 1-57444-610-X (Hardcover)
International Standard Book Number-13: 978-1-57444-610-4 (Hardcover)

Library of Congress Cataloging-in-Publication Data

Paulson, Daryl S., 1947-
Handbook of regression and modeling : applications for the clinical and pharmaceutical industries / Daryl S. Paulson.
p. ; cm. -- (Biostatistics ; 18)
Includes index.
ISBN-13: 978-1-57444-610-4 (hardcover : alk. paper)
ISBN-10: 1-57444-610-X (hardcover : alk. paper)
1. Medicine--Research--Statistical methods--Handbooks, manuals, etc. 2. Regression analysis--Handbooks, manuals, etc. 3. Drugs--Research--Statistical methods--Handbooks, manuals, etc. 4. Clinical trials--Statistical methods--Handbooks, manuals, etc. I. Title. II. Series: Biostatistics (New York, N.Y.) ; 18.
[DNLM: 1. Clinical Medicine. 2. Regression Analysis. 3. Biometry--methods. 4. Drug Industry. 5. Models, Statistical. WA 950 P332h 2007]

R853.S7P35 2007
610.72'7--dc22 2006030225

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Chapman & Hall/CRC Biostatistics Series

Handbook of Regression and Modeling

Applications for the Clinical and Pharmaceutical Industries

Chapman & Hall/CRC Biostatistics Series

Series Editor
Shein-Chung Chow, Ph.D.
Professor
Department of Biostatistics and Bioinformatics
Duke University School of Medicine
Durham, North Carolina, U.S.A.

Department of Statistics
National Cheng-Kung University
Tainan, Taiwan

1. *Design and Analysis of Animal Studies in Pharmaceutical Development,* Shein-Chung Chow and Jen-pei Liu
2. *Basic Statistics and Pharmaceutical Statistical Applications,* James E. De Muth
3. *Design and Analysis of Bioavailability and Bioequivalence Studies, Second Edition, Revised and Expanded,* Shein-Chung Chow and Jen-pei Liu
4. *Meta-Analysis in Medicine and Health Policy,* Dalene K. Stangl and Donald A. Berry
5. *Generalized Linear Models: A Bayesian Perspective,* Dipak K. Dey, Sujit K. Ghosh, and Bani K. Mallick
6. *Difference Equations with Public Health Applications,* Lemuel A. Moyé and Asha Seth Kapadia
7. *Medical Biostatistics,* Abhaya Indrayan and Sanjeev B. Sarmukaddam
8. *Statistical Methods for Clinical Trials,* Mark X. Norleans
9. *Causal Analysis in Biomedicine and Epidemiology: Based on Minimal Sufficient Causation,* Mikel Aickin
10. *Statistics in Drug Research: Methodologies and Recent Developments,* Shein-Chung Chow and Jun Shao
11. *Sample Size Calculations in Clinical Research,* Shein-Chung Chow, Jun Shao, and Hansheng Wang
12. *Applied Statistical Design for the Researcher,* Daryl S. Paulson
13. *Advances in Clinical Trial Biostatistics,* Nancy L. Geller
14. *Statistics in the Pharmaceutical Industry, 3rd Edition,* Ralph Buncher and Jia-Yeong Tsay
15. *DNA Microarrays and Related Genomics Techniques: Design, Analysis, and Interpretation of Experiments,* David B. Allsion, Grier P. Page, T. Mark Beasley, and Jode W. Edwards
16. *Basic Statistics and Pharmaceutical Statistical Applications, Second Edition,* James E. De Muth
17. *Adaptive Design Methods in Clinical Trials,* Shein-Chung Chow and Mark Chang
18. *Handbook of Regression and Modeling: Applications for the Clinical and Pharmaceutical Industries,* Daryl S. Paulson

Preface

In 2003, I wrote a book, *Applied Statistical Designs for the Researcher* (Marcel Dekker, Inc.), in which I covered experimental designs commonly encountered in the pharmaceutical, applied microbiological, and healthcare-product-formulation industries. It included two sample evaluations, analysis of variance, factorial, nested, chi-square, exploratory data analysis, nonparametric statistics, and a chapter on linear regression. Many researchers need more than simple linear regression methods to meet their research needs. It is for those researchers that this regression analysis book is written.

Chapter 1 is an overview of statistical methods and elementary concepts for statistical model building.

Chapter 2 covers simple linear regression applications in detail.

Chapter 3 deals with a problem that many applied researchers face when collecting data of time–serial correlation (the actual response values of y are correlated with one another). This chapter lays the foundation for the discussion on multiple regression in Chapter 8.

Chapter 4 introduces multiple linear regression procedures and matrix algebra. The knowledge of matrix algebra is not a prerequisite, and Appendix II presents the basics in matrix manipulation. Matrix notation is used because those readers without specific statistical software that contains "canned" statistical programs can still perform the statistical analyses presented in this book. However, I assume that the reader will perform most of the computations using statistical software such as SPSS, SAS, or MiniTab. This chapter also covers strategies for checking the contribution of each x_i variable in a regression equation to assure that it is actually contributing. Partial F-tests are used in stepwise, forward selection, and backward elimination procedures.

Chapter 5 focuses on aspects of correlation analysis and those of determining the contribution of x_i variables using partial correlation analysis.

Chapter 6 discusses common problems encountered in multiple linear regression and the ways to deal with them. One problem is multiple collinearity, in which some of the x_i variables are correlated with other x_i variables and the regression equation becomes unstable in applied work. A number of procedures are explained to deal with such problems and a biasing method called ridge regression is also discussed.

Chapter 7 describes aspects of polynomial regression and its uses.

Chapter 8 aids the researcher in determining outlier values of the variables y and x. It also includes residual analysis schemas, such as standardized, Studentized, and jackknife residual analyses. Another important feature of this chapter is leverage value identification, or identifying values, *y*s and *x*s, that have undue influence.

Chapter 9 applies indicator or dummy variables to an assortment of analyses.

Chapter 10 presents forward and stepwise selections of x_i variables, as well as backward elimination, in terms of statistical software.

Chapter 11 introduces covariance analysis, which combines regression and analysis of variance into one model.

The concepts presented in this book have been used for the past 25 years, in the clinical trials and new product development and formulation areas at BioScience Laboratories, Inc. They have also been used in analyzing data supporting studies submitted to the Food and Drug Administration (FDA) and the Environmental Protection Agency (EPA), and in my work as a statistician for the Association of Analytical Chemists (AOAC) in projects related to EPA regulation and Homeland Security.

This book has been two years in the making, from my standpoint. Certainly, it has not been solely an individual process on my part. I thank my friend and colleague, John A. Mitchell, PhD, also known as doctor for his excellent and persistent editing of this book, in spite of his many other duties at BioScience Laboratories, Inc. I also thank Tammy Anderson, my assistant, for again managing the entire manuscript process of this book, which is her sixth one for me. I also want to thank Marsha Paulson, my wife, for stepping up to the plate and helping us with the grueling final edit.

Daryl S. Paulson, PhD

Author

Daryl S. Paulson is the president and chief executive officer of BioScience Laboratories, Inc., Bozeman, Montana. Previously, he was the manager of laboratory services at Skyland Scientific Services (1987–1991), Belgrade, Montana. A developer of statistical models for clinical trials of drugs and cosmetics, he is the author of more than 40 articles on clinical evaluations, software validations, solid dosage validations, and quantitative management science. In addition, he has also authored several books, including *Topical Antimicrobial Testing and Evaluation*, the *Handbook of Topical Antimicrobials, Applied Statistical Designs for the Researcher* (Marcel Dekker, Inc.), *Competitive Business, Caring Business: An Integral Business Perspective for the 21st Century* (Paraview Press), and *The Handbook of Regression Analysis* (Taylor & Francis Group). Currently, his books *Biostatistics and Microbiology: A Survival Manual* (Springer Group) and the *Handbook of Applied Biomedical Microbiology: A Biofilms Approach* (Taylor & Francis Group) are in progress. He is a member of the American Society for Microbiology, the American Society for Testing and Materials, the Association for Practitioners in Infection Control, the American Society for Quality Control, the American Psychological Association, the American College of Forensic Examiners, and the Association of Analytical Chemists.

Dr. Paulson received a BA (1972) in business administration and an MS (1981) in medical microbiology and biostatistics from the University of Montana, Missoula. He also received a PhD (1988) in psychology from Sierra University, Riverside, California; a PhD (1992) in psychoneuroimmunology from Saybrook Graduate School and Research Center, San Francisco, California; an MBA (2002) from the University of Montana, Missoula; and a PhD in art from Warnborough University, United Kingdom. He is currently working toward a PhD in both psychology and statistics and performs statistical services for the AOAC and the Department of Homeland Security.

Series Introduction

The primary objectives of the *Biostatistics Book Series* are to provide useful reference books for researchers and scientists in academia, industry, and government, and also to offer textbooks for undergraduate and graduate courses in the area of biostatistics. This book series will provide comprehensive and unified presentations of statistical designs and analyses of important applications in biostatistics, such as those in biopharmaceuticals. A well-balanced summary is given of current and recently developed statistical methods, and interpretations for both statisticians and researchers or scientists with minimal statistical knowledge and engaged in the field of applied biostatistics. The series is committed to providing easy-to-understand, state-of-the-art references and textbooks. In each volume, statistical concepts and methodologies are illustrated through real-world examples.

Regression and modeling are commonly employed in pharmaceutical research and development. The purpose is not only to provide a valid and fair assessment of the pharmaceutical entity under investigation before regulatory approval, but also to assure that the pharmaceutical entity possesses good characteristics with the desired accuracy and reliability. In addition, it is to establish a predictive model for identifying patients who are most likely to respond to the test treatment under investigation. This volume is a condensation of various useful statistical methods that are commonly employed in pharmaceutical research and development. It covers important topics in pharmaceutical research and development such as multiple linear regression, model building or model selection, and analysis of covariance. This handbook provides useful approaches to pharmaceutical research and development. It would be beneficial to biostatisticians, medical researchers, and pharmaceutical scientists who are engaged in the areas of pharmaceutical research and development.

Shein-Chung Chow

Table of Contents

1 Basic Statistical Concepts

The use of statistics in clinical and pharmaceutical settings is extremely common. Because the data are generally collected under experimental conditions that result in measurements containing a certain amount of error,* statistical analyses, though not perfect, are the most effective way of making sense of the data. The situation is often portrayed as

$$T = t + e.$$

Here, the true but unknown value of a measurement, T, consists of a sample measurement, t, and random error or variation, e. Statistical error is considered to be the random variability inherent in any system, not a mistake. For example, the incubation temperature of bacteria in an incubator might have a normal random fluctuation of $\pm 1°\mathrm{C}$, which is considered a statistical error. A timer might have an inherent fluctuation of ± 0.01 sec for each minute of actual time. Statistical analysis enables the researcher to account for this random error.

Fundamental to statistical measurement are two basic parameters: the population mean, μ, and the population standard deviation, σ. The population parameters are generally unknown and are estimated by the sample mean, $\bar{x}$, and sample standard deviation, s. The sample mean is simply the central tendency of a sample set of data that is an unbiased estimate of the population mean, μ. The central tendency is the sum of values in a set, or population, of numbers divided by the number of values in that set or population. For example, for the sample set of values 10, 13, 19, 9, 11, and 17, the sum is 79. When 79 is divided by the number of values in the set, 6, the average is $79 \div 6 = 13.17$. The statistical formula for average is

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n},$$

*Statistical error is not a wrong measurement or a mistaken measurement. It is, instead, a representation of uncertainty concerning random fluctuations.

where the operator, $\sum_{i=1}^{n} x_i$, means to sum (add) the values beginning with $i = 1$ and ending with the value n; where n is the sample size.

The standard deviation for the population is written as σ, and for a sample as s.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{N}},$$

where $\sum_{i=1}^{n} (x_i - \mu)^2$ is the sum of the actual x_i values minus the population mean, the quantities squared; and N the total population size.

The sample standard deviation is given by

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}},$$

where $\sum_{i=1}^{n} (x_i - \bar{x})^2$ is the sum of the actual sample values minus the sample mean, the quantities squared; and $n - 1$ is the sample size minus 1, to account for the loss of one degree of freedom from estimating μ by $\bar{x}$. Note that the standard deviation σ or s is the square root of the variance σ^2 or s^2.

MEANING OF STANDARD DEVIATION

The standard deviation provides a measure of variability about the mean or average value. If two data sets have the same mean, but their data range differ,* so will their standard deviations. The larger the range, the larger the standard deviation.

For instance, using our previous example, the six data points—10, 13, 19, 9, 11, and 17—have a range of $19 - 9 = 10$. The standard deviation is calculated as

$$\sqrt{\frac{(10-13.1667)^2+(13-13.1667)^2+(19-13.1667)^2+(9-13.1667)^2+(11-13.1667)^2+(17-13.1667)^2}{6-1}}$$
$$= 4.0208.$$

Suppose the values were 1, 7, 11, 3, 28, and 29,

$$\bar{x} = \frac{1 + 7 + 11 + 3 + 28 + 29}{6} = 13.1667.$$

*Range = maximum value − minimum value.