DOUGLAS M. BATES
DONALD G. WATTS

NONLINEAR REGRESSION ANALYSIS AND ITS APPLICATIONS

WILEY SERIES IN PROBABILITY
AND MATHEMATICAL STATISTICS

WILEY

Nonlinear Regression Analysis and Its Applications

Douglas M. Bates

Department of Statistics
University of Wisconsin
Madison, Wisconsin

Donald G. Watts

Department of Mathematics
and Statistics
Queen's University
Kingston, Ontario, Canada

WILEY

JOHN WILEY & SONS

New York Chichester Brisbane Toronto Singapore

Library of Congress Cataloging in Publication Data:

Bates, Douglas M.
Nonlinear regression analysis and its applications / Douglas M. Bates, Donald G. Watts.

p. cm.—(Wiley series in probability and mathematical statistics. Applied probability and statistics)
Bibliography: p.
Includes index.

ISBN 0-471-81643-4
1. Regression analysis. 2. Linear models (Statistics)
3. Parameter estimation. I. Watts, Donald G. II. Title.
III. Series.
QA278.2.B375 1988 88-6065
519.5′36--dc19 CIP

Printed in the United States of America

10 9 8 7 6 5 4 3 2

To

Mary Ellen	Valery
Barbara	Lloyd
Michael	Megan

PREFACE

"Reading maketh a full man, conference; a ready man, and writing; an exact man."

– Francis Bacon

In this book we have tried to give a balanced presentation of the theory and practice of nonlinear regression.

We expect readers to have a working knowledge of linear regression at about the level of Draper and Smith (1981) or Montgomery and Peck (1982). Nevertheless, to provide background material and to establish notation, we give a summary review of linear least squares in Chapter 1, together with a geometrical development which is helpful in understanding both linear and nonlinear least squares. On the practical side, we discuss linear least squares in the context of modern computing methods and present useful material for checking the assumptions which are involved in regression and for modifying and improving fitted models. In Chapter 2 we discuss how nonlinear models can arise, and show how linear regression methods can be used iteratively to estimate the parameters. We also show how linear methods can be used to make approximate inferences about parameters and nonlinear model functions: again, the geometry is emphasized. The practical aspects of nonlinear estimation are discussed fully in Chapter 3, including such topics as getting starting values, transforming parameters, derivative-free methods, dealing with correlated residuals and with accumulated data, and comparing models.

In Chapter 4 we cover special methods for dealing with multiresponse data, and in Chapter 5, special techniques for compartment models in which the response function is specified as the solution to a set of linear differential equations.

In Chapter 6 we discuss improved methods for presenting the inferential results of a nonlinear analysis, using likelihood profile traces and profile t plots. Finally, in Chapter 7 we present material concerned with measuring how badly nonlinear a particular model–data set situation is. This chapter is helpful in understanding and appreciating the geometry of nonlinear least squares — and indeed, of linear least squares.

Extensive displays of geometrical constructs have been used to facilitate understanding. We have also used continuing examples so that readers can follow the development of ideas in manageable steps within familiar contexts.

All of the data sets used in this book are *real*, that is, the data were obtained from genuine physical, chemical, and biological experiments. We are grateful to the many authors, researchers, and publishers who gave permission to quote their data. In particular we would like to thank Don deBethizy, Rick Elliott, Steve Havriliak, Nico Linssen, Dave Pierson, Rob Stiratelli, Marg Treloar, and Eric Ziegel. We also acknowledge helpful comments made by participants in courses at Dalhousie University, Queen's University, and the University of Wisconsin, where the book was tested in class.

We are grateful to David Hamilton for stimulating discussions and collaborations, to Gunseog Kang for exemplary service in proofreading, and to Steve Czarniak, Mary Lindstrom and Dennis Wolf who also helped in proofreading.

The book was composed electronically using the *troff* text formatting language on the Statistics research computer at the University of Wisconsin–Madison. The figures were produced using the **S** language for statistics and graphics and both text and graphics were typeset on a Linotronic L300 using the PostScript language. We are appreciative of the good work done by Bea Shube and her colleagues at Wiley and by Bill Kasdorf of Impressions.

Considerable research was involved in developing the material in this book, and we thank the Natural Sciences and Engineering Research Council of Canada and the United States National Science Foundation for support.

Finally, we thank our wives for their continued love and encouragement.

June, 1988

Douglas M. Bates
Donald G. Watts

CONTENTS

CHAPTER 1.

Review of Linear Regression

"Non sunt multiplicanda entia praeter necessitatem."
(Entities are not to be multiplied beyond necessity.)

– *William of Ockham*

We begin with a brief review of linear regression, because a thorough grounding in linear regression is fundamental to understanding nonlinear regression. For a more complete presentation of linear regression see, for example, Draper and Smith (1981), Montgomery and Peck (1982), or Seber (1977). Detailed discussion of regression diagnostics is given in Belsley, Kuh, and Welsch (1980) and Cook and Weisberg (1982), and the Bayesian approach is discussed in Box and Tiao (1973).

Two topics which we emphasize are modern numerical methods and the geometry of linear least squares. As will be seen, attention to efficient computing methods increases understanding of linear regression, while the geometric approach provides insight into the methods of linear least squares and the analysis of variance, and subsequently into nonlinear regression.

1.1 The Linear Regression Model

Linear regression provides estimates and other inferential results for the *parameters* $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_P)^T$ in the model

$$\begin{aligned} Y_n &= \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_P x_{nP} + Z_n \\ &= (x_{n1}, \ldots, x_{nP})\boldsymbol{\beta} + Z_n \end{aligned} \tag{1.1}$$

In this model, the random variable Y_n, which represents the *response* for *case* n, $n = 1, 2, \ldots, N$, has a *deterministic* part and a *stochastic* part. The deterministic part, $(x_{n1}, \ldots, x_{nP})\boldsymbol{\beta}$, depends upon the parameters $\boldsymbol{\beta}$ and upon the *predictor* or *regressor variables* x_{np}, $p = 1, 2, \ldots, P$. The stochastic part, represented by the random variable Z_n, is a *disturbance* which perturbs the response for that case. The superscript $^{\mathrm{T}}$ denotes the transpose of a matrix.

The model for N cases can be written

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{Z} \tag{1.2}$$

where $\boldsymbol{Y}$ is the vector of random variables representing the data we may get, $\mathbf{X}$ is the $N \times P$ matrix of regressor variables,

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1P} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2P} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ x_{N1} & x_{N2} & x_{N3} & \cdots & x_{NP} \end{bmatrix}$$

and $\boldsymbol{Z}$ is the vector of random variables representing the disturbances. (We will use bold face italic letters for vectors of random variables.)

The deterministic part, $\mathbf{X}\boldsymbol{\beta}$, a function of the parameters and the regressor variables, gives the mathematical model or the model function for the responses. Since a nonzero mean for Z_n can be incorporated into the model function, we assume that

$$\mathrm{E}[\boldsymbol{Z}] = \mathbf{0} \tag{1.3}$$

or, equivalently,

$$\mathrm{E}[\boldsymbol{Y}] = \mathbf{X}\boldsymbol{\beta}$$

We therefore call $\mathbf{X}\boldsymbol{\beta}$ the *expectation function* for the regression model. The matrix $\mathbf{X}$ is called the *derivative matrix*, since the (n,p)th term is the derivative of the nth row of the expectation function with respect to the pth parameter.

Note that for linear models, *derivatives with respect to any of the parameters are independent of all the parameters.*

If we further assume that $\boldsymbol{Z}$ is normally distributed with

$$\mathrm{Var}[\boldsymbol{Z}] = \mathrm{E}[\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}] = \sigma^2\mathbf{I} \tag{1.4}$$

where $\mathbf{I}$ is an $N \times N$ identity matrix, then the joint probability density function for $\boldsymbol{Y}$, given $\boldsymbol{\beta}$ and the *variance* σ^2, is

$$\begin{aligned} p(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2) &= (2\pi\sigma^2)^{-N/2} \exp\left(\frac{-(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-N/2} \exp\left(\frac{-\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2}{2\sigma^2}\right) \end{aligned} \tag{1.5}$$

where the double vertical bars denote the length of a vector. When provided with a derivative matrix $\mathbf{X}$ and a vector of observed data $\mathbf{y}$, we wish to make inferences about σ^2 and the P parameters $\boldsymbol{\beta}$.

Example: PCB 1

As a simple example of a linear regression model, we consider the concentration of polychlorinated biphenyls (PCBs) in Lake Cayuga trout as a function of age (Bache et al., 1972). The data set is described in Appendix 1, Section A1.1. A plot of the PCB concentration versus age, Figure 1.1, reveals a curved relationship between PCB concentration and age. Furthermore, there is increasing variance in the PCB concentration as the concentration increases. Since the assumption (1.4) requires that the variance of the disturbances be constant, we seek a transformation of the PCB concentration which will stabilize the variance (see Section 1.3.2). Plotting the PCB concentration on a logarithmic scale, as in Figure 1.2*a*, nicely stabilizes the variance and produces a more nearly linear relationship. Thus, a linear expectation function of the form

$$\ln(\text{PCB}) = \beta_1 + \beta_2 \text{ age}$$

could be considered appropriate, where ln denotes the natural logarithm (logarithm to the base e). Transforming the regressor variable (Box and Tidwell, 1962) can produce an even straighter plot, as shown in Figure 1.2*b*, where we use the cube root of age. Thus a simple expectation function to be fitted is

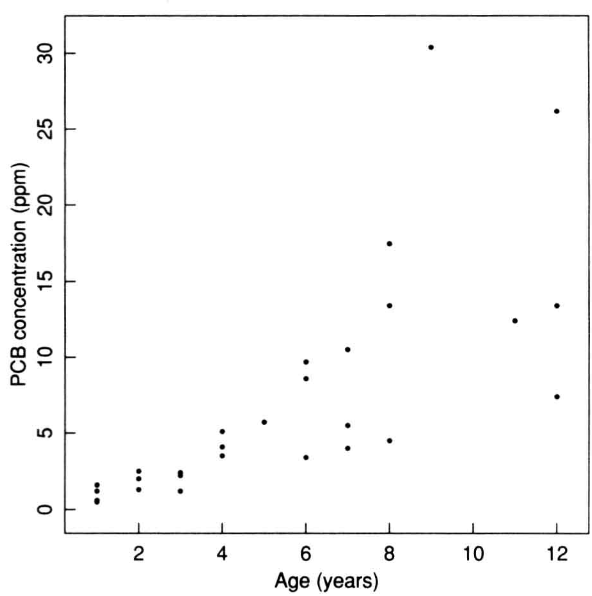

Figure 1.1 Plot of PCB concentration versus age for lake trout.

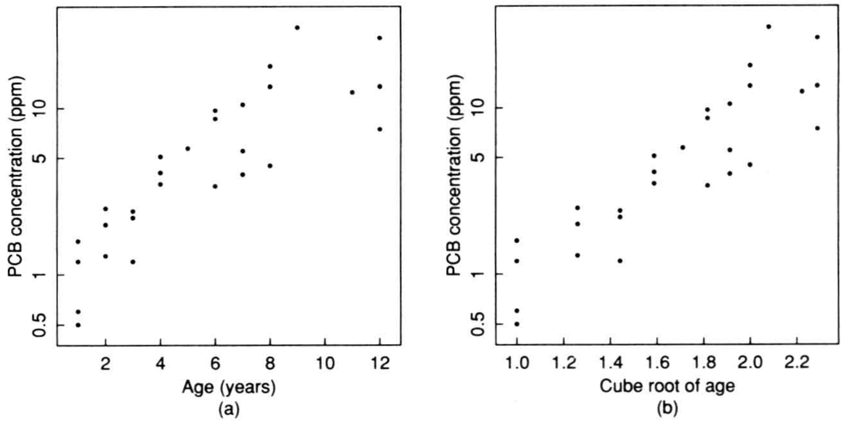

Figure 1.2 Plot of PCB concentration versus age for lake trout. The concentration, on a logarithmic scale, is plotted versus age in part *a* and versus $\sqrt[3]{\text{age}}$ in part *b*.

$$\ln(\text{PCB}) = \beta_1 + \beta_2 \sqrt[3]{\text{age}}$$

(Note that the methods of Chapter 2 can be used to fit models of the form

$$f(\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\alpha}) = \beta_0 + \beta_1 x_1^{\alpha_1} + \beta_2 x_2^{\alpha_2} + \cdots + \beta_P x_P^{\alpha_P}$$

by simultaneously estimating the conditionally linear parameters $\boldsymbol{\beta}$ and the transformation parameters $\boldsymbol{\alpha}$. The powers $\alpha_1, \ldots, \alpha_P$ are used to transform the factors so that a simple linear model in $x_1^{\alpha_1}, \ldots, x_P^{\alpha_P}$ is appropriate. In this book we use the power $\alpha = 0.33$ for the age variable even though, for the PCB data, the optimal value is 0.20.) ■

1.1.1 The Least Squares Estimates

The *likelihood function*, or more simply, the *likelihood*, $l(\boldsymbol{\beta}, \sigma \mid \mathbf{y})$, for $\boldsymbol{\beta}$ and σ is identical in form to the joint probability density (1.5) except that $l(\boldsymbol{\beta}, \sigma \mid \mathbf{y})$ is regarded as a function of the parameters conditional on the observed data, rather than as a function of the responses conditional on the values of the parameters. Suppressing the constant $(2\pi)^{-N/2}$, we write

$$l(\boldsymbol{\beta}, \sigma \mid \mathbf{y}) \propto \sigma^{-N} \exp\left(\frac{-\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \|^2}{2\sigma^2} \right) \tag{1.6}$$

The likelihood is maximized with respect to $\boldsymbol{\beta}$ when the *residual sum of squares*

$$\begin{aligned} S(\boldsymbol{\beta}) &= \| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \|^2 \\ &= \sum_{n=1}^{N} \left[y_n - \left(\sum_{p=1}^{P} x_{np} \beta_p \right) \right]^2 \end{aligned} \tag{1.7}$$

is a minimum. Thus the *maximum likelihood estimate* $\hat{\boldsymbol{\beta}}$ is the value of $\boldsymbol{\beta}$ which minimizes $S(\boldsymbol{\beta})$. This $\hat{\boldsymbol{\beta}}$ is called the *least squares* estimate and can be written

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{1.8}$$

Least squares estimates can also be derived by using sampling theory, since the least squares estimator is the minimum variance unbiased estimator for $\boldsymbol{\beta}$, or by using a Bayesian approach with a noninformative prior density on $\boldsymbol{\beta}$ and σ. In the Bayesian approach, $\hat{\boldsymbol{\beta}}$ is the mode of the marginal posterior density function for $\boldsymbol{\beta}$.

All three of these methods of inference, the likelihood approach, the sampling theory approach, and the Bayesian approach, produce the same point estimates for $\boldsymbol{\beta}$. As we will see shortly, they also produce similar regions of "reasonable" parameter values. First, however, it is important to realize that the least squares estimates are only appropriate when the model (1.2) and the assumptions on the disturbance term, (1.3) and (1.4), are valid. Expressed in another way, in using the least squares estimates we assume:

(1) The expectation function is correct.
(2) The response is expectation function plus disturbance.
(3) The disturbance is independent of the expectation function.
(4) Each disturbance has a normal distribution.
(5) Each disturbance has zero mean.
(6) The disturbances have equal variances.
(7) The disturbances are independently distributed.

When these assumptions appear reasonable and have been checked using diagnostic plots such as those described in Section 1.3.2, we can go on to make further inferences about the regression model.

Looking in detail at each of the three methods of statistical inference, we can characterize some of the properties of the least squares estimates.

1.1.2 Sampling Theory Inference Results

The least squares estimator has a number of desirable properties as shown, for example, in Seber (1977):

(1) The least squares estimator $\hat{\boldsymbol{\beta}}$ is normally distributed. This follows because the estimator is a linear function of $\boldsymbol{Y}$, which in turn is a linear function of $\boldsymbol{Z}$. Since $\boldsymbol{Z}$ is assumed to be normally distributed, $\hat{\boldsymbol{\beta}}$ is normally distributed.

(2) $\mathrm{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$: the least squares estimator is unbiased.

(3) $\mathrm{Var}[\hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$: the covariance matrix of the least squares estimator depends on the variance of the disturbances and on the derivative matrix $\mathbf{X}$.