Nonlinear Regression Analysis and Its Applications

Nonlinear Regression Analysis and Its Applications

Second edition

Douglas M. Bates and Donald G. Watts

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.

New York / Chichester / Weinheim / Brisbane / Singapore / Toronto

Preface

text...

This is a preface section

text...

Acknowledgments

To Mary Ellen and Valery

Contents

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* $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_P)^{\mathrm{T}}$ 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

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

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

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

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, $\boldsymbol{X}\boldsymbol{\beta}$, 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

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

or, equivalently,

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

We therefore call $\boldsymbol{X}\boldsymbol{\beta}$ the *expectation function* for the regression model. The matrix $\boldsymbol{X}$ 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

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

where $\boldsymbol{I}$ 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

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

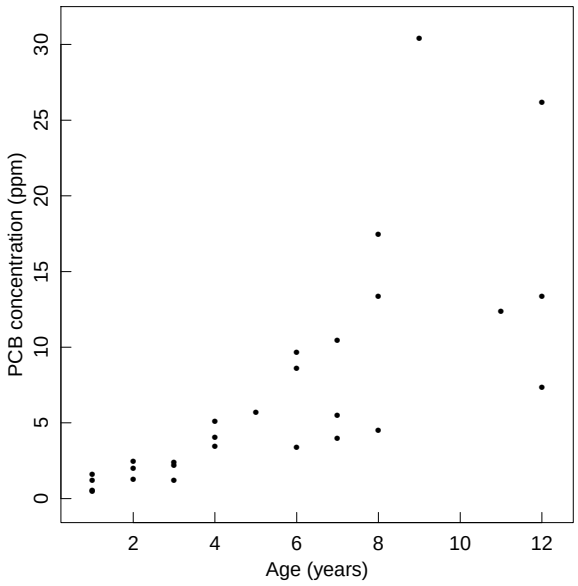

Fig. 1.1 Plot of PCB concentration versus age for lake trout.

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

Example:

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, Serum, Youngs and Lisk, 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

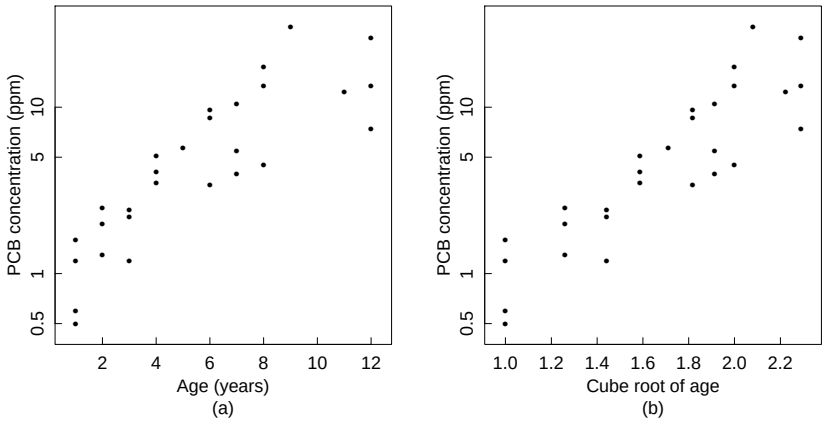

Fig. 1.2 Plot of PCB concentration versus age for lake trout. The concentration, on a logarithmic scale, is plotted versus age in part *a* and versus $\sqrt[3]{\text{age}}$ in part *b*.

function to be fitted is

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

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

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

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*, $l(\boldsymbol{\beta}, \sigma|\boldsymbol{y})$, for $\boldsymbol{\beta}$ and σ is identical in form to the joint probability density (1.5) except that $l(\boldsymbol{\beta}, \sigma|\boldsymbol{y})$ 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

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

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

$$S(\boldsymbol{\beta}) \quad = \quad \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 \tag{1.7}$$

$$= \sum_{n=1}^{N} \left[y_n - \left(\sum_{p=1}^{P} x_{np}\beta_p \right) \right]^2$$

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

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

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

4. A $1-\alpha$ *joint confidence region* for $\boldsymbol{\beta}$ is the ellipsoid

$$(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) \leq Ps^2F(P,N-P;\alpha) \tag{1.9}$$

where

$$s^2 = \frac{S(\hat{\boldsymbol{\beta}})}{N-P}$$

is the *residual mean square* or *variance estimate* based on $N-P$ *degrees of freedom*, and $F(P,N-P;\alpha)$ is the upper α quantile for Fisher's F distribution with P and $N-P$ degrees of freedom.

5. A $1-\alpha$ *marginal confidence interval* for the parameter β_p is

$$\hat{\beta}_p \pm \mathrm{se}(\hat{\beta}_p)t(N-P;\alpha/2) \tag{1.10}$$

where $t(N-P;\alpha/2)$ is the upper $\alpha/2$ quantile for Student's T distribution with $N-P$ degrees of freedom and the standard error of the parameter estimator is

$$\mathrm{se}(\hat{\beta}_p) = s\sqrt{\left\{(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\right\}_{pp}} \tag{1.11}$$

with $\left\{(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\right\}_{pp}$ equal to the pth diagonal term of the matrix $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$.

6. A $1-\alpha$ confidence interval for the expected response at $\boldsymbol{x}_0$ is

$$\boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}} \pm s\sqrt{\boldsymbol{x}_0^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_0}\; t(N-P;\alpha/2) \tag{1.12}$$

7. A $1-\alpha$ confidence interval for the expected response at $\boldsymbol{x}_0$ is

$$\boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}} \pm s\sqrt{\boldsymbol{x}_0^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_0}\; t(N-P;\alpha/2) \tag{1.13}$$

8. A $1-\alpha$ confidence band for the response function at any $\boldsymbol{x}$ is given by

$$\boldsymbol{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}} \pm s\sqrt{\boldsymbol{x}^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{x}}\sqrt{PF(P,N-P;\alpha)} \tag{1.14}$$

The expressions (1.13) and (1.14) differ because (1.13) concerns an interval at a single specific point, whereas (1.14) concerns the band produced by the intervals at all the values of $\boldsymbol{x}$ considered simultaneously.

1.1.3 Likelihood Inference Results

The likelihood $l(\boldsymbol{\beta}, \sigma \mid \boldsymbol{y})$, equation (1.6), depends on $\boldsymbol{\beta}$ only through $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|$, so likelihood contours are of the form

$$\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = c \tag{1.15}$$

where c is a constant. A likelihood region bounded by the contour for which

$$c = \ S(\hat{\boldsymbol{\beta}}) \left[1 + \frac{P}{N-P} F(P, N-P; \alpha)\right]$$

is identical to a $1-\alpha$ joint confidence region from the sampling theory approach. The interpretation of a likelihood region is quite different from that of a confidence region, however.

1.1.4 Bayesian Inference Results

As shown in Box and Tiao (1973), the Bayesian marginal posterior density for $\boldsymbol{\beta}$, assuming a noninformative prior density for $\boldsymbol{\beta}$ and σ of the form

$$p(\boldsymbol{\beta}, \sigma) \propto \sigma^{-1} \tag{1.16}$$

is

$$p(\boldsymbol{\beta}|\boldsymbol{y}) \propto \left\{1 + \frac{(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^{\mathrm{T}} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}{\nu s^2}\right\}^{-(\nu+P)/2} \tag{1.17}$$

which is in the form of a P-variate Student's T density with *location parameter* $\hat{\boldsymbol{\beta}}$, *scaling matrix* $s^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$, and $\nu = N - P$ degrees of freedom. Furthermore, the marginal posterior density for a single parameter β_p, say, is a univariate Student's T density with location parameter $\hat{\beta}_p$, scale parameter $s^2 \left\{(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\right\}_{pp}$, and degrees of freedom $N - P$. The marginal posterior density for the mean of y at $\boldsymbol{x}_0$ is a univariate Student's T density with location parameter $\boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}}$, scale parameter $s^2\boldsymbol{x}_0^{\mathrm{T}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{x}_0$, and degrees of freedom $N - P$.

A *highest posterior density* (HPD) region of content $1-\alpha$ is defined (Box and Tiao, 1973) as a region R in the parameter space such that $\Pr\{\boldsymbol{\beta} \in \mathrm{R}\} = 1 - \alpha$ and, for $\boldsymbol{\beta}_1 \in \mathrm{R}$ and $\boldsymbol{\beta}_2 \notin \mathrm{R}$, $p(\boldsymbol{\beta}_1|\boldsymbol{y}) \geq p(\boldsymbol{\beta}_2|\boldsymbol{y})$. For linear models with a noninformative prior, an HPD region is therefore given by the ellipsoid defined in (1.9). Similarly, the marginal HPD regions for β_p and $\boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{\beta}$ are numerically identical to the sampling theory regions (1.11, 1.12, and 1.13).

1.1.5 Comments

Although the three approaches to statistical inference differ considerably, they lead to essentially identical inferences. In particular, since the joint confidence,

likelihood, and Bayesian HPD regions are identical, we refer to them all as *inference regions*.

In addition, when referring to standard errors or correlations, we will use the Bayesian term "the standard error of β_p" when, for the sampling theory or likelihood methods, we should more properly say "the standard error of the estimate of β_p".

For linear least squares, any of the approaches can be used. For nonlinear least squares, however, the likelihood approach has the simplest and most direct geometrical interpretation, and so we emphasize it.

Example:

The PCB data can be used to determine parameter estimates and joint and marginal inference regions. In this linear situation, the regions can be summarized using $\hat{\boldsymbol{\beta}}$, s^2, $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$, and $\nu = N - P$. For the ln(PCB) data with $\sqrt[3]{\text{age}}$ as the regressor, we have $\hat{\boldsymbol{\beta}} = (-2.391, 2.300)^{\mathrm{T}}$, $s^2 = 0.246$ on $\nu = 26$ degrees of freedom, and

$$\begin{aligned} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} &= \begin{bmatrix} 28.000 & 46.941 \\ 46.941 & 83.367 \end{bmatrix} \\ \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1} &= \begin{bmatrix} 0.6374 & -0.3589 \\ -0.3589 & 0.2141 \end{bmatrix} \end{aligned}$$

The joint 95% inference region is then

$$\begin{aligned} 28.00(\beta_1 + 2.391)^2 + 93.88(\beta_1 + 2.391)(\beta_2 - 2.300) + 83.37(\beta_2 - 2.300)^2 &= 2(0.246)3.37 \\ &= 1.66 \end{aligned}$$

the marginal 95% inference interval for the parameter β_1 is

$$-2.391 \pm (0.496)\sqrt{0.6374}(2.056)$$

or

$$-3.21 \le \beta_1 \le -1.58$$

and the marginal 95% inference interval for the parameter β_2 is

$$2.300 \pm (0.496)\sqrt{0.2141}(2.056)$$

or

$$1.83 \le \beta_2 \le 2.77$$

The 95% inference band for the ln(PCB) value at any $\sqrt[3]{\text{age}} = x$, is

$$-2.391 + 2.300x \pm (0.496)\sqrt{0.637 - 0.718x + 0.214x^2}\sqrt{2(3.37)}$$

These regions are plotted in Figure 1.3. •

While it is possible to give formal expressions for the least squares estimators and the regression summary quantities in terms of the matrices $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ and $(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}$, the use of these matrices for computing the estimates is not recommended. Superior computing methods are presented in Section 1.2.2.

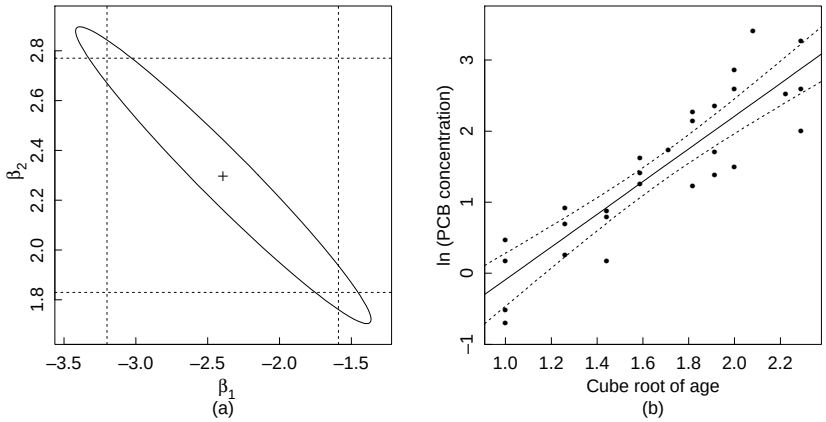

Fig. 1.3 Inference regions for the model $\ln(\text{PCB}) = \beta_1 + \beta_2 \sqrt[3]{\text{age}}$. Part *a* shows the least squares estimates (+), the parameter joint 95% inference region (solid line), and the marginal 95% inference intervals (dotted lines). Part *b* shows the fitted response (solid line) and the 95% inference band (dotted lines).

Finally, the assumptions which lead to the use of the least squares estimates should always be examined when using a regression model. Further discussion on assumptions and their implications is given in Section 1.3.

1.2 THE GEOMETRY OF LINEAR LEAST SQUARES

The model (1.2) and assumptions (1.3) and (1.4) lead to the use of the least squares estimate (1.8) which minimizes the residual sum of squares (1.7). As implied by (1.7), $S(\boldsymbol{\beta})$ can be regarded as the square of the distance from the data vector $\boldsymbol{y}$ to the expected response vector $\boldsymbol{X\beta}$. This links the subject of linear regression to Euclidean geometry and linear algebra. The assumption of a normally distributed disturbance term satisfying (1.3) and (1.4) indicates that the appropriate scale for measuring the distance between $\boldsymbol{y}$ and $\boldsymbol{X\beta}$ is the usual Euclidean distance between vectors. In this way the Euclidean geometry of the N-dimensional response space becomes statistically meaningful. This connection between geometry and statistics is exemplified by the use of the term *spherical normal* for the normal distribution with the assumptions (1.3) and (1.4), because then contours of constant probability are spheres.

Note that when we speak of the linear form of the expectation function $\boldsymbol{X\beta}$, we are regarding it as a function of the parameters $\boldsymbol{\beta}$, and that when determining parameter estimates we are only concerned with how the expected response depends on the *parameters*, not with how it depends on the *variables*. In the PCB example we fit the response to $\sqrt[3]{\text{age}}$ using linear least squares because the parameters $\boldsymbol{\beta}$ enter the model linearly.

1.2.1 The Expectation Surface

The process of calculating $S(\boldsymbol{\beta})$ involves two steps:

1. Using the P-dimensional parameter vector $\boldsymbol{\beta}$ and the $N \times P$ derivative matrix $\boldsymbol{X}$ to obtain the N-dimensional *expected response vector* $\boldsymbol{\eta}(\boldsymbol{\beta}) = \boldsymbol{X}\boldsymbol{\beta}$, and

2. Calculating the squared distance from $\boldsymbol{\eta}(\boldsymbol{\beta})$ to the observed response $\boldsymbol{y}$, $\|\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\beta})\|^2$.

The possible expected response vectors $\boldsymbol{\eta}(\boldsymbol{\beta})$ form a P-dimensional *expectation surface* in the N-dimensional response space. This surface is a linear subspace of the response space, so we call it the *expectation plane* when dealing with a linear model.

Example:

To illustrate the geometry of the expectation surface, consider just three cases from the ln(PCB) versus $\sqrt[3]{\text{age}}$ data,

$\sqrt[3]{\text{age}}$	ln(PCB)
1.26	0.92
1.82	2.15
2.22	2.52

The matrix $\boldsymbol{X}$ is then

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1.26 \\ 1 & 1.82 \\ 1 & 2.22 \end{bmatrix}$$

which consists of two column vectors $\boldsymbol{x}_1 = (1, 1, 1)^{\mathrm{T}}$ and $\boldsymbol{x}_2 = (1.26, 1.82, 2.22)^{\mathrm{T}}$. These two vectors in the 3-dimensional response space are shown in Figure 1.4*b*, and correspond to the points $\boldsymbol{\beta} = (1, 0)^{\mathrm{T}}$ and $\boldsymbol{\beta} = (0, 1)^{\mathrm{T}}$ in the parameter plane, shown in Figure 1.4*a*. The expectation function $\boldsymbol{\eta}(\boldsymbol{\beta}) = \boldsymbol{X}\boldsymbol{\beta}$ defines a 2-dimensional expectation plane in the 3-dimensional response space. This is shown in Figure 1.4*c*, where the parameter lines corresponding to the lines $\beta_1 = -3, \ldots, 5$ and $\beta_2 = -2, \ldots, 2$, shown in Figure 1.4*a*, are given. A parameter line is associated with the parameter which is varying so the lines corresponding to $\beta_1 = -3, \ldots, 5$ (dotdashed lines) are called β_2 lines.

Note that the parameter lines in the parameter plane are straight, parallel, and equispaced, and that their images on the expectation plane are also straight, parallel, and equispaced. Because the vector $\boldsymbol{x}_1$ is shorter than $\boldsymbol{x}_2$ ($\|\boldsymbol{x}_1\| = \sqrt{3}$ while $\|\boldsymbol{x}_2\| = \sqrt{9.83}$), the spacing between the lines of constant β_1 on the expectation plane is less than that between the lines of constant β_2. Also, the vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are not orthogonal. The angle ω between them can be calculated from

$$\cos\omega = \frac{\boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_2}{\|\boldsymbol{x}_1\|\|\boldsymbol{x}_2\|}$$

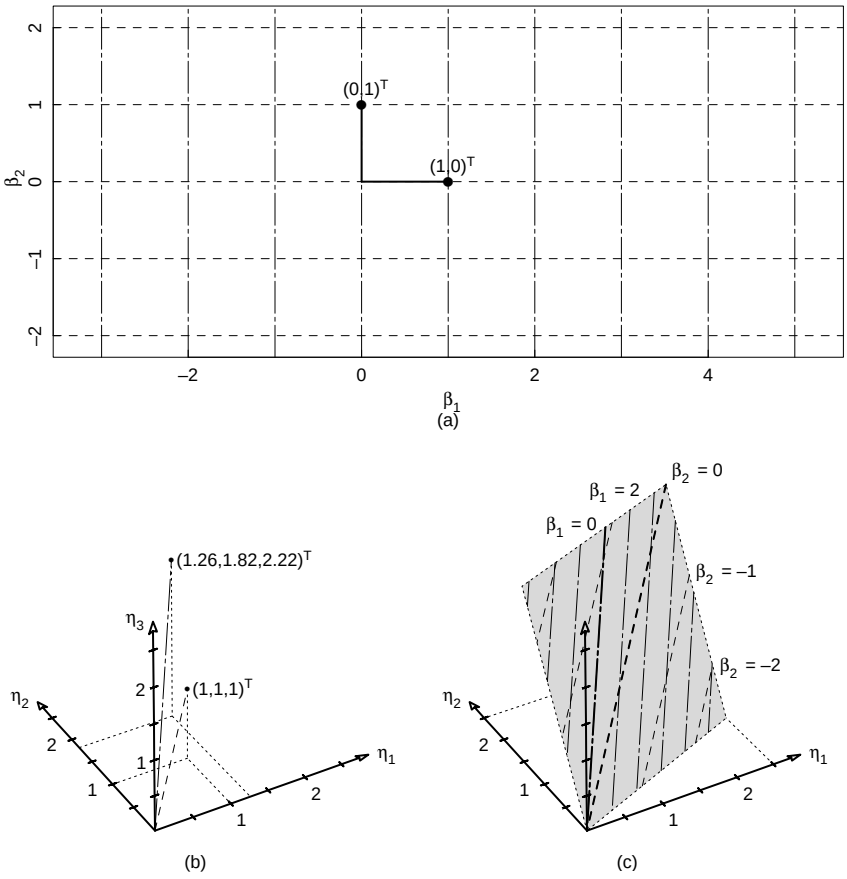

Fig. 1.4 Expectation surface for the 3-case PCB example. Part a shows the parameter plane with β_1 parameter lines (dashed) and β_2 parameter lines (dot–dashed). Part b shows the vectors $\boldsymbol{x}_1$ (dashed line) and $\boldsymbol{x}_2$ (dot–dashed line) in the response space. The end points of the vectors correspond to $\boldsymbol{\beta} = (1, 0)^{\mathrm{T}}$ and $\boldsymbol{\beta} = (0, 1)^{\mathrm{T}}$ respectively. Part c shows a portion of the expectation plane (shaded) in the response space, with β_1 parameter lines (dashed) and β_2 parameter lines (dot–dashed).

$$\begin{aligned} &= \frac{5.30}{\sqrt{(3)(9.83)}} \\ &= 0.98 \end{aligned}$$

to be about 11°, so the parameter lines on the expectation plane are not at right angles as they are on the parameter plane.

As a consequence of the unequal length and nonorthogonality of the vectors, unit squares on the parameter plane map to parallelograms on the expectation plane. The area of the parallelogram is

$$\begin{aligned} \|\boldsymbol{x}_1\|\|\boldsymbol{x}_2\|\sin\omega &= \|\boldsymbol{x}_1\|\|\boldsymbol{x}_2\|\sqrt{1-\cos^2\omega} \\ &= \sqrt{(\boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_1)(\boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{x}_2) - (\boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_2)^2} \\ &= \sqrt{\left|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right|} \end{aligned} \tag{1.18}$$

That is, the *Jacobian determinant* of the transformation from the parameter plane to the expectation plane is a constant equal to $\left|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right|^{1/2}$. Conversely, the ratio of areas in the parameter plane to those on the expectation plane is $\left|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right|^{-1/2}$. •

The simple linear mapping seen in the above example is true for all linear regression models. That is, for linear models, straight parallel equispaced lines in the parameter space map to straight parallel equispaced lines on the expectation plane in the response space. Consequently, rectangles in one plane map to parallelepipeds in the other plane, and circles or spheres in one plane map to ellipses or ellipsoids in the other plane. Furthermore, the Jacobian determinant, $\left|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right|^{1/2}$, is a constant for linear models, and so regions of fixed size in one plane map to regions of fixed size in the other, no matter where they are on the plane. These properties, which make linear least squares especially simple, are discussed further in Section 1.2.3.

1.2.2 Determining the Least Squares Estimates

The geometric representation of linear least squares allows us to formulate a very simple scheme for determining the parameters estimates $\hat{\boldsymbol{\beta}}$. Since the expectation surface is linear, all we must do to determine the point on the surface which is closest to the point $\boldsymbol{y}$, is to project $\boldsymbol{y}$ onto the expectation plane. This gives us $\hat{\boldsymbol{\eta}}$, and $\hat{\boldsymbol{\beta}}$ is then simply the value of $\boldsymbol{\beta}$ corresponding to $\hat{\boldsymbol{\eta}}$.

One approach to defining this projection is to observe that, after the projection, the residual vector $\boldsymbol{y}-\hat{\boldsymbol{\eta}}$ will be *orthogonal*, or *normal*, to the expectation plane. Equivalently, the residual vector must be orthogonal to all the columns of the $\boldsymbol{X}$ matrix, so

$$\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{0}$$

which is to say that the least squares estimate $\hat{\boldsymbol{\beta}}$ satisfies the *normal equations*

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{1.19}$$

Because of (1.19) the least squares estimates are often written $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$ as in (1.8). However, another way of expressing the estimate, and a more stable way of computing it, involves decomposing $\boldsymbol{X}$ into the product of an orthogonal matrix and an easily inverted matrix. Two such decompositions are the QR decomposition and the singular value decomposition (Dongarra, Bunch, Moler and Stewart, 1979, Chapters 9 and 11). We use the QR decomposition, where

$$\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R}$$

with the $N \times N$ matrix $\boldsymbol{Q}$ and the $N \times P$ matrix $\boldsymbol{R}$ constructed so that $\boldsymbol{Q}$ is orthogonal (that is, $\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}^{\mathrm{T}} = \boldsymbol{I}$) and $\boldsymbol{R}$ is zero below the main diagonal. Writing

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_1 \\ \boldsymbol{0} \end{bmatrix}$$

where $\boldsymbol{R}_1$ is $P \times P$ and upper triangular, and

$$\boldsymbol{Q} = [\boldsymbol{Q}_1 | \boldsymbol{Q}_2]$$

with $\boldsymbol{Q}_1$ the first P columns and $\boldsymbol{Q}_2$ the last $N - P$ columns of $\boldsymbol{Q}$, we have

$$\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R} = \boldsymbol{Q}_1\boldsymbol{R}_1 \tag{1.20}$$

Performing a QR decomposition is straightforward, as is shown in Appendix 2.

Geometrically, the columns of $\boldsymbol{Q}$ define an *orthonormal*, or *orthogonal*, basis for the response space with the property that the first P columns span the expectation plane. Projection onto the expectation plane is then very easy if we work in the coordinate system given by $\boldsymbol{Q}$. For example we transform the response vector to

$$\boldsymbol{w} = \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{y} \tag{1.21}$$

with components

$$\boldsymbol{w}_1 = \boldsymbol{Q}_1^{\mathrm{T}}\boldsymbol{y} \tag{1.22}$$

and

$$\boldsymbol{w}_2 = \boldsymbol{Q}_2^{\mathrm{T}}\boldsymbol{y} \tag{1.23}$$

The projection of $\boldsymbol{w}$ onto the expectation plane is then simply

$$\begin{bmatrix} \boldsymbol{w}_1 \\ \boldsymbol{0} \end{bmatrix}$$

in the $\boldsymbol{Q}$ coordinates and

$$\hat{\boldsymbol{\eta}} = \boldsymbol{Q}\begin{bmatrix} \boldsymbol{w}_1 \\ \boldsymbol{0} \end{bmatrix} = \boldsymbol{Q}_1\boldsymbol{w}_1 \tag{1.24}$$

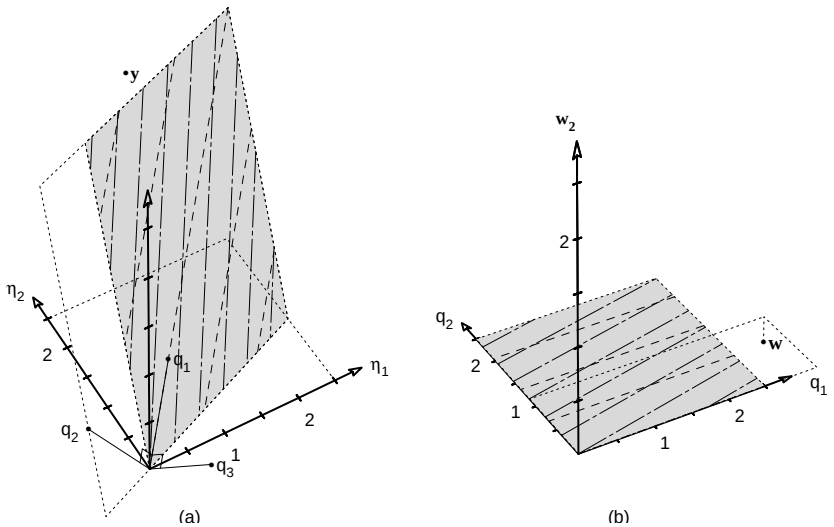

Fig. 1.5 Expectation surface for the 3-case PCB example. Part a shows a portion of the expectation plane (shaded) in the response space with β_1 parameter lines (dashed) and β_2 parameter lines (dot–dashed) together with the response vector $\boldsymbol{y}$. Also shown are the orthogonal unit vectors $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$ in the expectation plane, and $\boldsymbol{q}_3$ orthogonal to the plane. Part b shows the response vector $\boldsymbol{w}$, and a portion of the expectation plane (shaded) in the rotated coordinates given by $\boldsymbol{Q}$.

in the original coordinates.

Example:

As shown in Appendix 2, the QR decomposition (1.20) of the matrix

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1.26 \\ 1 & 1.82 \\ 1 & 2.22 \end{bmatrix}$$

for the 3-case PCB example is

$$\begin{bmatrix} 0.5774 & -0.7409 & 0.3432 \\ 0.5774 & 0.0732 & -0.8132 \\ 0.5774 & 0.6677 & 0.4700 \end{bmatrix} \begin{bmatrix} 1.7321 & 3.0600 \\ 0 & 0.6820 \\ 0 & 0 \end{bmatrix}$$

which gives [equation (1.21)]

$$\boldsymbol{w} = \begin{bmatrix} 3.23 \\ 1.16 \\ -0.24 \end{bmatrix}$$

In Figure 1.5a we show the expectation plane and observation vector in the original coordinate system. We also show the vectors $\boldsymbol{q}_1, \boldsymbol{q}_2, \boldsymbol{q}_3$,

which are the columns of $\boldsymbol{Q}$. It can be seen that $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$ lie in the expectation plane and $\boldsymbol{q}_3$ is orthogonal to it. In Figure 1.5*b* we show, in the transformed coordinates, the observation vector and the expectation plane, which is now horizontal. Note that projecting $\boldsymbol{w}$ onto the expectation plane is especially simple, since it merely requires replacing the last element in $\boldsymbol{w}$ by zero. •

To determine the least squares estimate we must find the value $\hat{\boldsymbol{\beta}}$ corresponding to $\hat{\boldsymbol{\eta}}$. Since

$$\hat{\boldsymbol{\eta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

using (1.24) and (1.20)

$$\boldsymbol{R}_1\hat{\boldsymbol{\beta}} = \boldsymbol{w}_1 \tag{1.25}$$

and we solve for $\hat{\boldsymbol{\beta}}$ by back-substitution (Stewart, 1973).

Example:

For the complete ln(PCB), $\sqrt[3]{\text{age}}$ data set,

$$\boldsymbol{R}_1 = \begin{bmatrix} 5.29150 & 8.87105 \\ 0 & 2.16134 \end{bmatrix}$$

and $\boldsymbol{w}_1 = (7.7570, 4.9721)^{\mathrm{T}}$, so $\hat{\boldsymbol{\beta}} = (-2.391, 2.300)^{\mathrm{T}}$. •

1.2.3 Parameter Inference Regions

Just as the least squares estimates have informative geometric interpretations, so do the parameter inference regions (1.9), (1.10), (1.15) and those derived from (1.17). Such interpretations are helpful for understanding linear regression, and are essential for understanding nonlinear regression. (The geometric interpretation is less helpful in the Bayesian approach, so we discuss only the sampling theory and likelihood approaches.)

The main difference between the likelihood and sampling theory geometric interpretations is that the likelihood approach centers on the point $\boldsymbol{y}$ and the length of the residual vector at $\boldsymbol{\eta}(\boldsymbol{\beta})$ compared to the shortest residual vector, while the sampling theory approach focuses on possible values of $\boldsymbol{\eta}(\boldsymbol{\beta})$ and the angle that the resulting residual vectors could make with the expectation plane.

1.2.3.1 The Geometry of Sampling Theory Results To develop the geometric basis of linear regression results from the sampling theory approach, we transform to the $\boldsymbol{Q}$ coordinate system. The model for the random variable $\boldsymbol{W} = \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Y}$ is

$$\boldsymbol{W} = \boldsymbol{R}\boldsymbol{\beta} + \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Z}$$

or

$$\boldsymbol{U} = \boldsymbol{W} - \boldsymbol{R}\boldsymbol{\beta} \tag{1.26}$$

where $\boldsymbol{U} = \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Z}$.

The spherical normal distribution of $\boldsymbol{Z}$ is not affected by the orthogonal transformation, so $\boldsymbol{U}$ also has a spherical normal distribution. This can be established on the basis of the geometry, since the spherical probability contours will not be changed by a rigid rotation or reflection, which is what an orthogonal transformation must be. Alternatively, this can be established analytically because $\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q} = \boldsymbol{I}$, so the determinant of $\boldsymbol{Q}$ is ± 1 and $\|\boldsymbol{Q}\boldsymbol{x}\| = \|\boldsymbol{x}\|$ for any N-vector $\boldsymbol{x}$. Now the joint density for the random variables $\boldsymbol{Z} = (Z_1, \ldots, Z_n)^{\mathrm{T}}$ is

$$p_{\boldsymbol{Z}}(\boldsymbol{z}) = (2\pi\sigma^2)^{-N/2} \exp\left(\frac{-\boldsymbol{z}^{\mathrm{T}}\boldsymbol{z}}{2\sigma^2}\right)$$

and, after transformation, the joint density for $\boldsymbol{U} = \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Z}$ is

$$p_{\boldsymbol{U}}(\boldsymbol{u}) = (2\pi\sigma^2)^{-N/2}|\boldsymbol{Q}| \exp\left(\frac{-\boldsymbol{u}^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{u}}{2\sigma^2}\right)$$

$$= (2\pi\sigma^2)^{-N/2} \exp\left(\frac{-\boldsymbol{u}^{\mathrm{T}}\boldsymbol{u}}{2\sigma^2}\right)$$

From (1.26), the form of $\boldsymbol{R}$ leads us to partition $\boldsymbol{U}$ into two components:

$$\boldsymbol{U} = \begin{bmatrix} \boldsymbol{U}_1 \\ \boldsymbol{U}_2 \end{bmatrix}$$

where $\boldsymbol{U}_1$ consists of the first P elements of $\boldsymbol{U}$, and $\boldsymbol{U}_2$ the remaining $N - P$ elements. Each of these components has a spherical normal distribution of the appropriate dimension. Furthermore, independence of elements in the original disturbance vector $\boldsymbol{Z}$ leads to independence of the elements of $\boldsymbol{U}$, so the components $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ are independent.

The dimensions ν_i of the components $\boldsymbol{U}_i$, called the *degrees of freedom*, are $\nu_1 = P$ and $\nu_2 = N - P$. The sum of squares of the coordinates of a ν-dimensional spherical normal vector has a $\sigma^2\chi^2$ distribution on ν degrees of freedom, so

$$\begin{aligned} \|\boldsymbol{U}_1\|^2 &\sim \sigma^2\chi^2_P \\ \|\boldsymbol{U}_2\|^2 &\sim \sigma^2\chi^2_{N-P} \end{aligned}$$

where the symbol $\sim$ is read "is distributed as." Using the independence of $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$, we have

$$\frac{\|\boldsymbol{U}_1\|^2/P}{\|\boldsymbol{U}_2\|^2/(N-P)} \sim F(P, N-P) \tag{1.27}$$

since the scaled ratio of two independent χ^2 random variables is distributed as Fisher's F distribution.

The distribution (1.27) gives a reference distribution for the ratio of the squared component lengths or, equivalently, for the angle that the disturbance

vector makes with the horizontal plane. We may therefore use (1.26) and (1.27) to test the hypothesis that $\boldsymbol{\beta}$ equals some specific value, say $\boldsymbol{\beta}^0$, by calculating the residual vector $\boldsymbol{u}^0 = \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{y} - \boldsymbol{R}\boldsymbol{\beta}^0$ and comparing the lengths of the components $\boldsymbol{u}_1^0$ and $\boldsymbol{u}_2^0$ as in (1.27). The reasoning here is that a large $\|\boldsymbol{u}_1^0\|$ compared to $\|\boldsymbol{u}_2^0\|$ suggests that the vector $\boldsymbol{y}$ is not very likely to have been generated by the model (1.2) with $\boldsymbol{\beta} = \boldsymbol{\beta}^0$, since $\boldsymbol{u}^0$ has a suspiciously large component in the $\boldsymbol{Q}_1$ plane.

Note that

$$\frac{\|\boldsymbol{u}_2^0\|^2}{N-P} = \frac{S(\hat{\boldsymbol{\beta}})}{N-P} = s^2$$

and

$$\|\boldsymbol{u}_1^0\|^2 = \|\boldsymbol{R}_1\boldsymbol{\beta}^0 - \boldsymbol{w}_1\|^2 \tag{1.28}$$

and so the ratio (1.27) becomes

$$\frac{\|\boldsymbol{R}_1\boldsymbol{\beta}^0 - \boldsymbol{w}_1\|^2}{Ps^2} \tag{1.29}$$

Example:

We illustrate the decomposition of the residual $\boldsymbol{u}$ for testing the null hypothesis

$$\mathrm{H}_0 : \boldsymbol{\beta} = (-2.0, 2.0)^{\mathrm{T}}$$

versus the alternative

$$\mathrm{H}_{\mathrm{A}} : \boldsymbol{\beta} \neq (-2.0, 2.0)^{\mathrm{T}}$$

for the full PCB data set in Figure 1.6. Even though the rotated data vector $\boldsymbol{w}$ and the expectation surface for this example are in a 28-dimensional space, the relevant distances can be pictured in the 3-dimensional space spanned by the expectation surface (vectors $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$) and the residual vector. The scaled lengths of the components $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ are compared to determine if the point $\boldsymbol{\beta}^0 = (-2.0, 2.0)^{\mathrm{T}}$ is reasonable.

The numerator in (1.29) is

$$\left\| \begin{bmatrix} 5.29150 & 8.87105 \\ 0 & 2.16134 \end{bmatrix} \begin{bmatrix} -2.0 \\ 2.0 \end{bmatrix} - \begin{bmatrix} 7.7570 \\ 4.9721 \end{bmatrix} \right\|^2 = 0.882$$

The ratio is then $0.882/(2 \times 0.246) = 1.79$, which corresponds to a tail probability (or p value) of 0.19 for an F distribution with 2 and 26 degrees of freedom. Since the probability of obtaining a ratio at least as large as 1.79 is 19%, we do not reject the null hypothesis. •

A $1-\alpha$ joint confidence region for the parameters $\boldsymbol{\beta}$ consists of all those values for which the above hypothesis test is not rejected at level α. Thus, a value $\boldsymbol{\beta}^0$ is within a $1-\alpha$ confidence region if

$$\frac{\|\boldsymbol{u}_1^0\|^2/P}{\|\boldsymbol{u}_2^0\|^2/(N-P)} \leq F(P, N-P; \alpha)$$

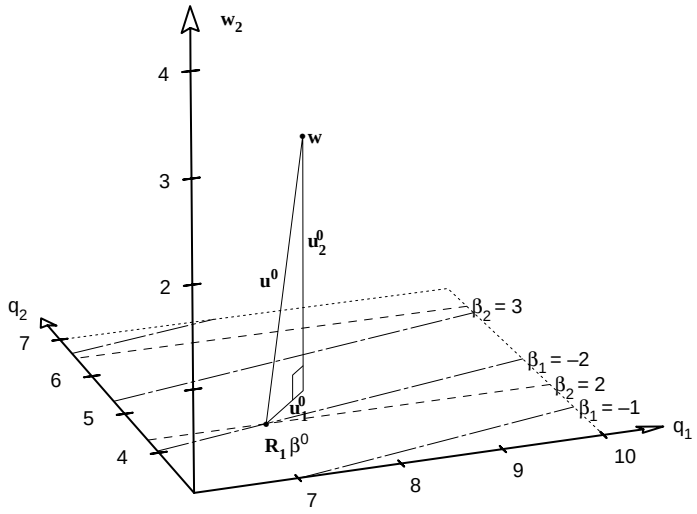

Fig. 1.6 A geometric interpretation of the test $\mathrm{H}_0 : \boldsymbol{\beta} = (-2.0, 2.0)^{\mathrm{T}}$ for the full PCB data set. We show the projections of the response vector $\boldsymbol{w}$ and a portion of the expectation plane projected into the 3-dimensional space given by the tangent vectors $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$, and the orthogonal component of the response vector, $\boldsymbol{w}_2$. For the test point $\boldsymbol{\beta}^0$, the residual vector $\boldsymbol{u}^0$ is decomposed into a tangential component $\boldsymbol{u}_1^0$ and an orthogonal component $\boldsymbol{u}_2^0$.

Since s^2 does not depend on $\boldsymbol{\beta}^0$, the points inside the confidence region form a disk on the expectation plane defined by

$$\|\boldsymbol{u}_1\|^2 \leq Ps^2F(P, N-P; \alpha)$$

Furthermore, from (1.25) and (1.28) we have

$$\|\boldsymbol{u}_1\|^2 = \|\boldsymbol{R}_1(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\|^2$$

so a point on the boundary of the confidence region in the parameter space satisfies

$$\boldsymbol{R}_1(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) = \sqrt{Ps^2F(P, N-P; \alpha)}\ \boldsymbol{d}$$

where $\|\boldsymbol{d}\| = 1$. That is, the confidence region is given by

$$\left\{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}} + \sqrt{Ps^2F(P, N-P; \alpha)}\boldsymbol{R}_1^{-1}\boldsymbol{d} \mid \|\boldsymbol{d}\| = 1\right\} \tag{1.30}$$

Thus the region of "reasonable" parameter values is a disk centered at $\boldsymbol{R}_1\hat{\boldsymbol{\beta}}$ on the expectation plane and is an ellipse centered at $\hat{\boldsymbol{\beta}}$ in the parameter space.

Example:

For the ln(PCB) versus $\sqrt[3]{\text{age}}$ data, $\hat{\boldsymbol{\beta}} = (-2.391, 2.300)^{\mathrm{T}}$ and $s^2 = 0.246$ based on 26 degrees of freedom, so the 95% confidence disk on the transformed expectation surface is

$$\boldsymbol{R}_1\boldsymbol{\beta} = \begin{bmatrix} 7.7570 \\ 4.9721 \end{bmatrix} + 1.288 \begin{bmatrix} \cos\omega \\ \sin\omega \end{bmatrix}$$

where $0 \leq \omega \leq 2\pi$. The disk is shown in the expectation plane in Figure 1.7*a*, and the corresponding ellipse

$$\boldsymbol{\beta} = \begin{bmatrix} -2.391 \\ 2.300 \end{bmatrix} + 1.288 \begin{bmatrix} 0.18898 & -0.77566 \\ 0 & 0.46268 \end{bmatrix} \begin{bmatrix} \cos\omega \\ \sin\omega \end{bmatrix}$$

is shown in the parameter plane in Figure 1.7*b*. •

1.2.4 Marginal Confidence Intervals

We can create a marginal confidence interval for a single parameter, say β_1, by "inverting" a hypothesis test of the form

$$\mathrm{H}_0 : \beta_1 = \beta_1^0$$

versus

$$\mathrm{H}_\mathrm{A} : \beta_1 \neq \beta_1^0$$

Any β_1^0 for which H_0 is not rejected at level α is included in the $1-\alpha$ confidence interval. To perform the hypothesis test, we choose any parameter vector with $\beta_1 = \beta_1^0$, say $\left(\beta_1^0, \mathbf{0}^{\mathrm{T}}\right)^{\mathrm{T}}$, calculate the transformed residual vector $\boldsymbol{u}^0$, and divide it into three components: the first component $\boldsymbol{u}_1^0$ of dimension $P-1$

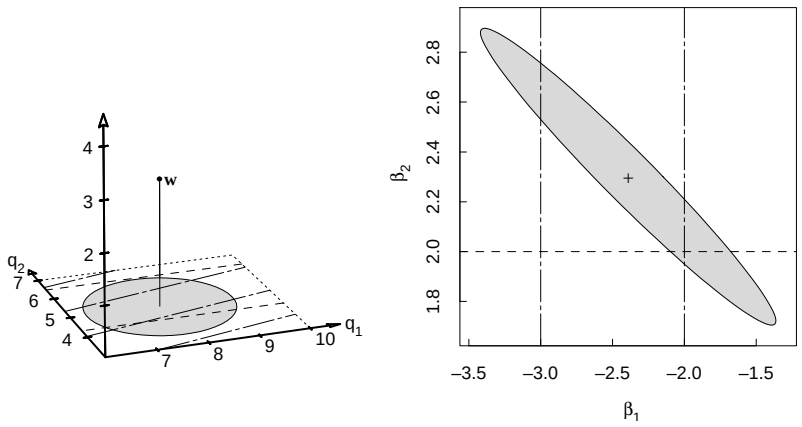

Fig. 1.7 The 95% confidence disk and parameter confidence region for the PCB data. Part *a* shows the response vector $\boldsymbol{w}$ and a portion of the expectation plane projected into the 3-dimensional space given by the tangent vectors $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$, and the orthogonal component of the response vector, $\boldsymbol{w}_2$. The 95% confidence disk (shaded) in the expectation plane (part *a*) maps to the elliptical confidence region (shaded) in the parameter plane (part *b*).

and parallel to the hyperplane defined by $\beta_1 = \beta_1^0$; the second component u_2^0 of dimension 1 and in the expectation plane but orthogonal to the β_1^0 hyperplane; and the third component $\boldsymbol{u}_3^0$ of length $(N-P)s^2$ and orthogonal to the expectation plane. The component u_2^0 is the same for any parameter $\boldsymbol{\beta}$ with $\beta_1 = \beta_1^0$, and, assuming that the true β_1 is β_1^0, the scaled ratio of the corresponding random variables U_2 and $\boldsymbol{U}_3$ has the distribution

$$\frac{U_2^2/1}{\|\boldsymbol{U}_3\|^2/(N-P)} \sim F(1, N-P)$$

Thus we reject H_0 at level α if

$$\left(u_2^0\right)^2 s^2 F(1, N-P; \alpha)$$

Example:

To test the null hypothesis

$$\mathrm{H}_0 : \beta_1 = -2.0$$

versus the alternative

$$\mathrm{H}_\mathrm{A} : \beta_1 \neq -2.0$$

for the complete PCB data set, we decompose the transformed residual vector at $\boldsymbol{\beta}^0 = (-2.0, 2.2)^\mathrm{T}$ into three components as shown in Figure 1.8 and calculate the ratio

$$\frac{\left(u_2^0\right)^2}{s^2} = \frac{0.240}{0.246}$$

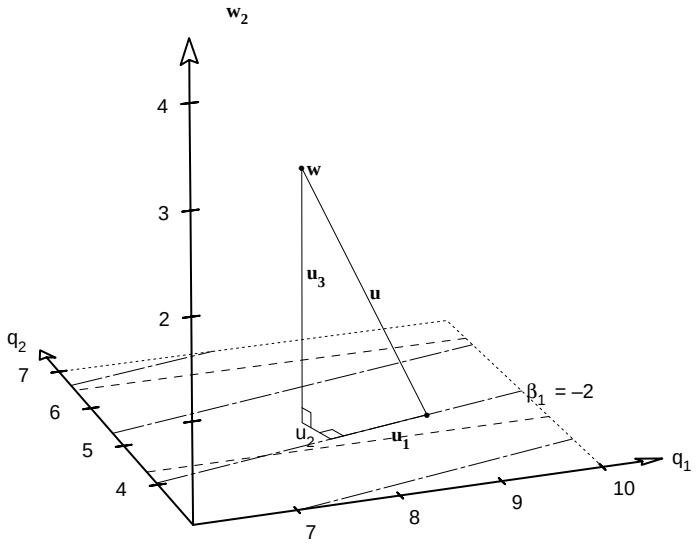

Fig. 1.8 A geometric interpretation of the test $H_{0:}\beta_1 = -2.0$ for the full PCB data set. We show the response vector $\boldsymbol{w}$, and a portion of the expectation plane projected into the 3-dimensional space given by the tangent vectors $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$, and the orthogonal component of the response vector, $\boldsymbol{w}_2$. For a representative point on the line $\beta_1 = -2$ the residual vector $\boldsymbol{u}$ is decomposed into a tangential component $\boldsymbol{u}_1^0$ along the line, a tangential component u_2^0 perpendicular to the line, and an orthogonal component $\boldsymbol{u}_3^0$.

$$= \quad 0.97$$

This corresponds to a p value of 0.33, and so we do not reject the null hypothesis. •

We can create a $1-\alpha$ marginal confidence interval for β_1 as all values for which

$$\left(u_2^0\right)^2 \leq s^2 F(1, N-P; \alpha)$$

or, equivalently,

$$|u_2^0| \leq s \cdot t(N-P; \alpha/2) \tag{1.31}$$

Since $|u_2^0|$ is the distance from the point $\boldsymbol{R}_1\hat{\boldsymbol{\beta}}$ to the line corresponding to $\beta_1 = \beta_1^0$ on the transformed parameter plane, the confidence interval will include all values β_1^0 for which the corresponding parameter line intersects the disk

$$\left\{\boldsymbol{R}_1\hat{\boldsymbol{\beta}} + st(N-P;\alpha/2)\boldsymbol{d} \mid \|\boldsymbol{d}\| = 1\right\} \tag{1.32}$$

Instead of determining the value of $|u_2^0|$ for each β_1^0, we take the disk (1.32) and determine the minimum and maximum values of β_1 for points on the disk. Writing $\boldsymbol{r}^1$ for the first row of $\boldsymbol{R}_1^{-1}$, the values of β_1 corresponding to points on the expectation plane disk are

$$\boldsymbol{r}^1(\boldsymbol{R}_1\hat{\boldsymbol{\beta}} + s \cdot t(N-P;\alpha/2)\boldsymbol{d}) = \hat{\beta}_1 + s \cdot t(N-P;\alpha/2)\boldsymbol{r}^1\boldsymbol{d}$$

and the minimum and maximum occur for the unit vectors in the direction of $\boldsymbol{r}^1$; that is, $\boldsymbol{d} = \pm\left(\boldsymbol{r}^1\right)^{\mathrm{T}}/\|\boldsymbol{r}^1\|$. This gives the confidence interval

$$\hat{\beta}_1 \pm s\|\boldsymbol{r}^1\|t(N-P;\alpha/2)$$

In general, a marginal confidence interval for parameter β_p is

$$\hat{\beta}_p \pm s\|\boldsymbol{r}^p\|t(N-P;\alpha/2) \tag{1.33}$$

where $\boldsymbol{r}^p$ is the pth row of $\boldsymbol{R}_1^{-1}$. The quantity

$$\mathrm{se}(\hat{\beta}_p) = s\|\boldsymbol{r}^p\| \tag{1.34}$$

is called the *standard error* for the pth parameter. Since

$$\begin{aligned}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1} &= (\boldsymbol{R}_1^{\mathrm{T}}\boldsymbol{R}_1)^{-1} \\ &= \boldsymbol{R}_1^{-1}\boldsymbol{R}_1^{-\mathrm{T}}\end{aligned}$$

$\|\boldsymbol{r}^p\|^2 = \left\{(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\right\}_{pp}$, so the standard error can be written as in equation (1.11).

A convenient summary of the variability of the parameter estimates can be obtained by factoring $\boldsymbol{R}_1^{-1}$ as

$$\boldsymbol{R}_1^{-1} = \mathrm{diag}(\|\boldsymbol{r}^1\|, \|\boldsymbol{r}^2\|, \ldots, \|\boldsymbol{r}^P\|)\boldsymbol{L} \tag{1.35}$$

where $\boldsymbol{L}$ has unit length rows. The diagonal matrix provides the parameter standard errors, while the *correlation matrix*

$$\boldsymbol{C} = \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}} \tag{1.36}$$

gives the correlations between the parameter estimates.

Example:

For the ln(PCB) data, $\hat{\boldsymbol{\beta}} = (-2.391, 2.300)^{\mathrm{T}}$, $s^2 = 0.246$ with 26 degrees of freedom, and

$$\begin{aligned}
\boldsymbol{R}_1^{-1} &= \begin{bmatrix} 5.29150 & 8.87105 \\ 0 & 2.16134 \end{bmatrix}^{-1} \\
&= \begin{bmatrix} 0.18898 & -0.77566 \\ 0 & 0.46268 \end{bmatrix} \\
&= \begin{bmatrix} 0.798 & 0 \\ 0 & 0.463 \end{bmatrix} \begin{bmatrix} 0.237 & -0.972 \\ 0 & 1 \end{bmatrix}
\end{aligned}$$

which gives standard errors of $0.798\sqrt{0.246} = 0.396$ for β_1 and $0.463\sqrt{0.246} = 0.230$ for β_2. Also

$$\boldsymbol{C} = \begin{bmatrix} 1 & -0.972 \\ -0.972 & 1 \end{bmatrix}$$

so the correlation between β_1 and β_2 is -0.97. The 95% confidence intervals for the parameters are given by $-2.391 \pm 2.056(0.396)$ and $2.300 \pm 2.056(0.230)$, which are plotted in Figure 1.3*a*. •

Marginal confidence intervals for the expected response at a design point $\boldsymbol{x}_0$ can be created by determining which hyperplanes formed by constant $\boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{\beta}$ intersect the disk (1.32). Using the same argument as was used to derive (1.33), we obtain a standard error for the expected response at $\boldsymbol{x}_0$ as $s\|\boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{R}_1^{-1}\|$, so the confidence interval is

$$\boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}} \pm s\|\boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{R}_1^{-1}\| t(N-P;\alpha/2) \tag{1.37}$$

Similarly, a confidence band for the response function is

$$\boldsymbol{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}} \pm s\|\boldsymbol{x}^{\mathrm{T}}\boldsymbol{R}_1^{-1}\| \sqrt{PF(P,N-P;\alpha)} \tag{1.38}$$

Example:

A plot of the fitted expectation function and the 95% confidence bands for the PCB example was given in Figure 1.3*b*. •

Ansley (1985) gives derivations of other sampling theory results in linear regression using the QR decomposition, which, as we have seen, is closely related to the geometric approach to regression.

1.2.5 The Geometry of Likelihood Results

The likelihood function indicates the plausibility of values of $\boldsymbol{\eta}$ relative to $\boldsymbol{y}$, and consequently has a simple geometrical interpretation. If we allow $\boldsymbol{\eta}$ to

take on any value in the N-dimensional response space, the likelihood contours are spheres centered on $\boldsymbol{y}$. Values of $\boldsymbol{\eta}$ of the form $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$ generate a P-dimensional expectation plane, and so the intersection of the plane with the likelihood spheres produces disks.

Analytically, the likelihood function (1.6) depends on $\boldsymbol{\eta}$ through

$$\begin{aligned} \|\boldsymbol{\eta} - \boldsymbol{y}\|^2 &= \|\boldsymbol{Q}^{\mathrm{T}}(\boldsymbol{\eta} - \boldsymbol{y})\|^2 \\ &= \|\boldsymbol{Q}_1^{\mathrm{T}}(\boldsymbol{\eta} - \boldsymbol{y})\|^2 + \|\boldsymbol{Q}_2^{\mathrm{T}}(\boldsymbol{\eta} - \boldsymbol{y})\|^2 \\ &= \|\boldsymbol{w}(\boldsymbol{\beta}) - \boldsymbol{w}_1\|^2 + \|\boldsymbol{w}_2\|^2 \end{aligned} \tag{1.39}$$

where $\boldsymbol{w}(\boldsymbol{\beta}) = \boldsymbol{Q}_1^{\mathrm{T}}\boldsymbol{\eta}$ and $\boldsymbol{Q}_2^{\mathrm{T}}\boldsymbol{\eta} = \boldsymbol{0}$. A constant value of the total sum of squares specifies a disk of the form

$$\|\boldsymbol{w}(\boldsymbol{\beta}) - \boldsymbol{w}_1\|^2 = c$$

on the expectation plane. Choosing

$$c = Ps^2F(P, N-P; \alpha)$$

produces the disk corresponding to a $1-\alpha$ confidence region. In terms of the total sum of squares, the contour is

$$S(\boldsymbol{\beta}) = S(\hat{\boldsymbol{\beta}})\left\{1 + \frac{P}{N-P}F(P, N-P; \alpha)\right\} \tag{1.40}$$

As shown previously, and illustrated in Figure 1.7, this disk transforms to an ellipsoid in the parameter space.

1.3 ASSUMPTIONS AND MODEL ASSESSMENT

The statistical assumptions which lead to the use of the least squares estimates encompass several different aspects of the regression model. As with any statistical analysis, if the assumptions on the model and data are not appropriate, the results of the analysis will not be valid.

Since we cannot guarantee *a priori* that the different assumptions are all valid, we must proceed in an iterative fashion as described, for example, in Box, Hunter and Hunter (1978). We entertain a plausible statistical model for the data, analyze the data using that model, then go back and use *diagnostics* such as plots of the residuals to assess the assumptions. If the diagnostics indicate failure of assumptions in either the deterministic or stochastic components of the model, we must modify the model or the analysis and repeat the cycle.

It is important to recognize that the design of the experiment and the method of data collection can affect the chances of assumptions being valid in a particular experiment. In particular *randomization* can be of great help in ensuring the appropriateness of all the assumptions, and *replication* allows greater ability to check the appropriateness of specific assumptions.

1.3.1 Assumptions and Their Implications

The assumptions, as listed in Section 1.1.1, are:

1. *The expectation function is correct.* Ensuring the validity of this assumption is, to some extent, the goal of all science. We wish to build a model with which we can predict natural phenomena. It is in building the mathematical model for the expectation function that we frequently find ourselves in an iterative loop. We proceed as though the expectation function were correct, but we should be prepared to modify it as the data and the analyses dictate. In almost all linear, and in many nonlinear, regression situations we do not know the "true" model, but we choose a plausible one by examining the situation, looking at data plots and cross-correlations, and so on. As the analysis proceeds we can modify the expectation function and the assumptions about the disturbance term to obtain a more sensible and useful answer. Models should be treated as just models, and it must be recognized that some will be more appropriate or adequate than others. Nevertheless, assumption (1) is a strong one, since it implies that the expectation function includes all the important predictor variables in precisely the correct form, and that it does *not* include any unimportant predictor variables. A useful technique to enable checking the adequacy of a model function is to include replications in the experiment. It is also important to actually manipulate the predictor variables and randomize the order in which the experiments are done, to ensure that *causation*, not *correlation*, is being determined (Box, 1960).

2. *The response is expectation function plus disturbance.* This assumption is important theoretically, since it allows the probability density function for the random variable $\boldsymbol{Y}$ describing the responses to be simply calculated from the probability density function for the random variable $\boldsymbol{Z}$ describing the disturbances. Thus,

$$p_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) = p_{\boldsymbol{Z}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}|\sigma^2)$$

   In practice, this assumption is closely tied to the assumption of constant variance of the disturbances. It may be the case that the disturbances can be considered as having constant variance, but as entering the model multiplicatively, since in many phenomena, as the level of the "signal" increases, the level of the "noise" increases. This lack of additivity of the disturbance will manifest itself as a nonconstant variance in the diagnostic plots. In both cases, the corrective action is the same—either use weighted least squares or take a transformation of the response as was done in Example PCB 1.

3. *The disturbance is independent of the expectation function.* This assumption is closely related to assumption (2), since they both relate to

appropriateness of the additive model. One of the implications of this assumption is that the control or predictor variables are measured perfectly. Also, as a converse to the implication in assumption (1) that all important variables are included in the model, this assumption implies that *any important variables which are not included are not systematically related* to the response. An important technique to improve the chances that this is true is to randomize the order in which the experiments are done, as suggested by Fisher (1935). In this way, if an important variable has been omitted, its effect may be manifested as a disturbance (and hence simply inflate the variability of the observations) rather than being confounded with one of the predictor effects (and hence bias the parameter estimates). And, of course, it is important to actually manipulate the predictor variables not merely record their values.

4. *Each disturbance has a normal distribution.* The assumption of normality of the disturbances is important, since this dictates the form of the sampling distribution of the random variables describing the responses, and through this, the likelihood function for the parameters. This leads to the criterion of least squares, which is enormously powerful because of its mathematical tractability. For example, given a linear model, it is possible to write down the analytic solution for the parameter estimators and to show [Gauss's theorem (Seber, 1977)] that the least squares estimates are the best *both individually and in any linear combination*, in the sense that they have the smallest mean square error of any linear estimators. The normality assumption can be justified by appealing to the central limit theorem, which states that the resultant of many disturbances, no one of which is dominant, will tend to be normally distributed. Since most experiments involve many operations to set up and measure the results, it is reasonable to assume, at least tentatively, that the disturbances will be normally distributed. Again, the assumption of normality will be more likely to be appropriate if the order of the experiments is randomized. The assumption of normality may be checked by examining the residuals.

5. *Each disturbance has zero mean.* This assumption is primarily a simplifying one, which reduces the number of unknown parameters to a manageable level. Any nonzero mean common to all observations can be accommodated by introducing a constant term in the expectation function, so this assumption is unimportant in linear regression. It can be important in nonlinear regression, however, where many expectation functions occur which do not include a constant. The main implication of this assumption is that there is no systematic bias in the disturbances such as could be caused by an unsuspected influential variable. Hence, we see again the value of randomization.

6. *The disturbances have equal variances.* This assumption is more important practically than theoretically, since a solution exists for the least squares estimation problem for the case of unequal variances [see, e.g., Draper and Smith (1981) concerning weighted least squares]. Practically, however, one must describe how the variances vary, which can only be done by making further assumptions, or by using information from replications and incorporating this into the analysis through generalized least squares, or by transforming the data. When the variance is constant, the likelihood function is especially simple, since the parameters can be estimated independently of the nuisance parameter σ^2. The main implication of this assumption is that all data values are *equally unreliable*, and so the simple least squares criterion can be used. The appropriateness of this assumption can sometimes be checked after a model has been fitted by plotting the residuals versus the fitted values, but it is much better to have replications. With replications, we can check the assumption before even fitting a model, and can in fact use the replication averages and variances to determine a suitable *variance-stabilizing* transformation; see Section 1.3.2. Transforming to constant variance often has the additional effect of making the disturbances behave more normally. This is because a constant variance is necessarily independent of the mean (and anything else, for that matter), and this independence property is fundamental to the normal density.

7. *The disturbances are distributed independently.* The final assumption is that the disturbances in different experiments are independent of one another. This is an enormously simplifying assumption, because then the joint probability density function for the vector $\boldsymbol{Y}$ is just the product of the probability densities for the individual random variables $Y_n, n = 1, 2, \ldots, N$. The implication of this assumption is that the disturbances on separate runs are not systematically related, an assumption which can usually be made to be more appropriate by randomization. Nonindependent disturbances can be treated by generalized least squares, but, as in the case where there is nonconstant variance, modifications to the model must be made either through information gained from the data, or by additional assumptions as to the nature of the interdependence.

1.3.2 Model Assessment

In this subsection we present some simple methods for verifying the appropriateness of assumptions, especially through plots of residuals. Further discussion on regression diagnostics for linear models is given in Hocking (1983), and in the books by Belsley et al. (1980), Cook and Weisberg (1982), and Draper and Smith (1981). In Chapter 3 we discuss model assessment for nonlinear models.

1.3.3 Plotting Residuals

A simple, effective method for checking the adequacy of a model is to plot the *studentized residuals*, $\hat{z}_n/s\sqrt{1-h_{nn}}$, versus the predictor variables and any other possibly important "lurking" variables (Box, 1960; Joiner, 1981). The term h_{nn} is the nth diagonal term of the "hat" matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}} = \boldsymbol{Q}_1\boldsymbol{Q}_1^{\mathrm{T}}$, and $\hat{z}_n$ is the residual for the nth case,

$$\hat{z}_n = y_n - \hat{y}_n$$

A relationship between the residuals and any variable then suggests that there is an effect due to that variable which has not been accounted for. Features to look for include systematic linear or curvilinear behavior of the residuals with respect to a variable. Important common "lurking" variables include time or the order number of the experiment; if a plot of residuals versus time shows suspicious behavior, such as runs of residuals of the same sign, then the assumption of independence of the disturbances may be inappropriate.

Plotting residuals versus the fitted values $\hat{y}_n$ is also useful, since such plots can reveal outliers or general inadequacy in the form of the expectation function. It is also a very effective plot for revealing whether the assumption of constant variance is appropriate. The most common form of nonconstant variance is an increase in the variability in the responses when the level of the response changes. This behavior was noticed in the original PCB data. If a regression model is fitted to such data, the plot of the studentized residuals versus the fitted values tends to have a wedge-shaped pattern.

When residual plots or the data themselves give an indication of nonconstant variance, the estimation procedure should be modified. Possible changes include transforming the data as was done with the PCB data or using weighted least squares.

A quantile–quantile plot (Chambers, Cleveland, Kleiner and Tukey, 1983) of the studentized residuals versus a normal distribution gives a direct check on the assumption of normality. If the expectation function is correct and the assumption of normality is appropriate, such a *normal probability plot* of the residuals should be a fairly straight line. Departures from a straight line therefore suggest inappropriateness of the normality assumption, although, as demonstrated in Daniel and Wood (1980), considerable variability can be expected in normal plots. Normal probability plots are also good for revealing outliers.

Example:

> Plots of residuals are given in Figure 1.9 for the fit of ln(PCB) to $\sqrt[3]{\text{age}}$. Since the fitted values are a linear function of the regressor variable $\sqrt[3]{\text{age}}$, the form of the plot of the studentized residuals versus $\hat{y}$ will be the same as that versus $\sqrt[3]{\text{age}}$, so we only display the former. The plot versus $\hat{y}$ and the quantile–quantile plot are well behaved. Neither plot reveals outliers. •

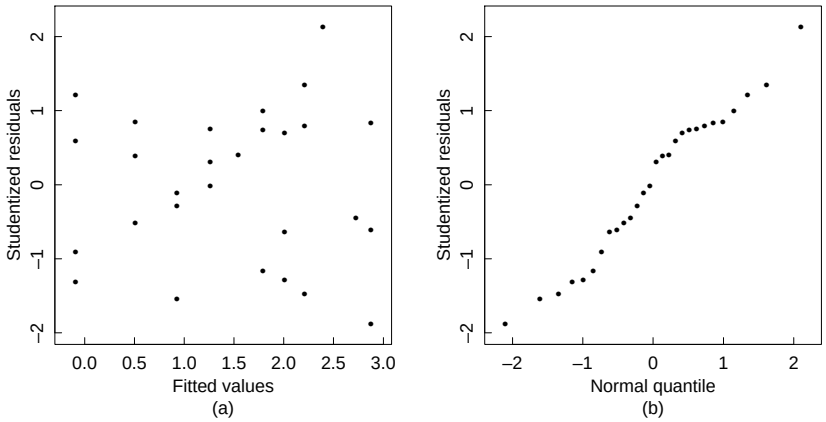

Fig. 1.9 Studentized residuals for the PCB data plotted versus fitted values in part *a* and versus normal quantiles in part *b*.

1.3.4 Stabilizing Variance

An experiment which includes replications allows further tests to be made on the appropriateness of assumptions. For example, even before an expectation function has been proposed, it is possible to check the assumption of constant variance by using an analysis of variance to get averages and variances for each set of replications and plotting the variances and standard deviations versus the averages. If the plots show systematic relationships, then one can use a variance-stabilizing procedure to transform to constant variance.

One procedure is to try a range of power transformations in the form (Box and Cox, 1964)

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln y & \lambda = 0 \end{cases}$$

We calculate and plot variances versus averages for $y^{(\lambda)}, \lambda = 0, \pm 0.5, \pm 1, \ldots$ and select that value of λ for which the variance appears to be most stable. Alternatively, for a random variable $\boldsymbol{Y}$, if there is a power relationship between the standard deviation σ and the mean μ such that $\sigma \propto \mu^\alpha$, it can be shown (Draper and Smith, 1981; Montgomery and Peck, 1982; Box et al., 1978) that the variance of the transformed random variable $\boldsymbol{Y}^{1-\alpha}$ will be approximately constant.

Variance-stabilizing transformations usually have the additional benefit of making the distribution of the disturbances appear more nearly normal, as discussed in Section 1.3.1. Alternatively, one can use the replication information to assist in choosing a form of weighting for weighted least squares.

Example:

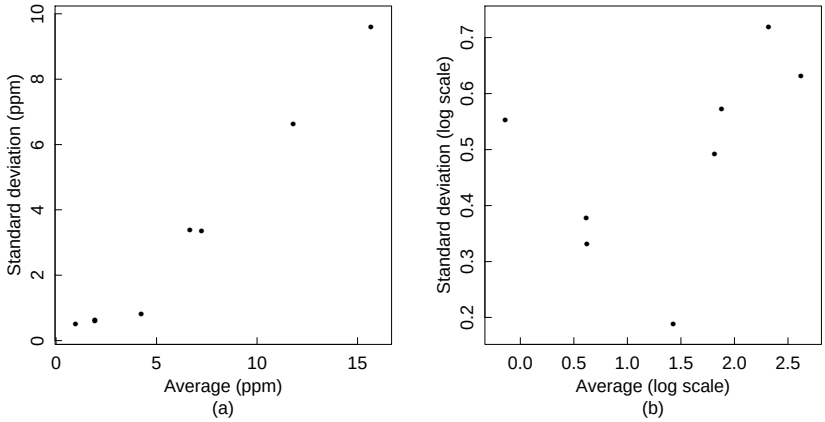

Fig. 1.10 Replication standard deviations plotted versus replication averages for the PCB data in part a and for the ln(PCB) data in part b.

A plot of the standard deviations versus the averages for the original PCB data is given in Figure 1.10a. It can be seen that there is a good straight line relationship between s and $\bar{y}$, and so the variance-stabilizing technique leads to the logarithmic transformation. In Figure 1.10b we plot the standard deviations versus the averages for the ln(PCB) data. This plot shows no systematic relationship, and hence substantiates the effectiveness of the logarithmic transformation in stabilizing the variance. •

1.3.5 Lack of Fit

When the data set includes replications, it is also possible to perform tests for *lack of fit* of the expectation function. Such analyses are based on an analysis of variance in which the residual sum of squares $S(\hat{\boldsymbol{\beta}})$ with $N - P$ degrees of freedom is decomposed into the *replication* sum of squares S_r (equal to the total sum of squares of deviations of the replication values about their averages) with, say, ν_r degrees of freedom, and the *lack of fit* sum of squares $S_l = S(\hat{\boldsymbol{\beta}}) - S_r$, with $\nu_l = fN - P - \nu_r$ degrees of freedom. We then compare the ratio of the lack of fit mean square over the replication mean square with the appropriate value in the F table. That is, we compare

$$\frac{S_l/\nu_l}{S_r/\nu_r} \text{ with } F(\nu_l, \nu_r; \alpha)$$

to determine whether there is significant lack of fit at level α. The geometric justification for this analysis is that the replication subspace is always orthogonal to the subspace containing the averages and the expectation function.

Table 1.1 Lack of fit analysis of the model fitted to the PCB data

Source	Degrees of Freedom	Sum of Squares	Mean Square	F Ratio	p Value
Lack of fit	9	1.923	0.214	0.812	0.61
Replication	17	4.475	0.263		
Residuals	26	6.398	0.246		

If no lack of fit is found, then the lack of fit analysis of variance has served its purpose, and the estimate of σ^2 should be based on the residual mean square. That is, the replication and lack of fit sums of squares and degrees of freedom should be recombined to give an estimate with the largest number of degrees of freedom, so as to provide the most reliable parameter and expected value confidence regions. If lack of fit is found, the analyst should attempt to discover why, and modify the expectation function accordingly. Further discussion on assessing the fit of a model and on modifying and comparing models is given in Sections 3.7 and 3.10.

Example:

For the ln(PCB) versus $\sqrt[3]{\text{age}}$ data, the lack of fit analysis is presented in Table 1.3.5. Because the p value suggests no lack of fit, we combine the lack of fit and replication sums of squares and degrees of freedom and take as our estimate of σ^2, the residual mean square of 0.246 based on 26 degrees of freedom. If there had been lack of fit, we would have had to modify the model: in either situation, we do not simply use the replication mean square as an estimate of the variance. •

Problems

1.1 Write a computer routine in a language of your choice to perform a QR decomposition of a matrix using Householder transformations.

1.2 Draw a picture to show the Householder transformation of a vector $\boldsymbol{y} = (y_1, y_2)^{\mathrm{T}}$ to the x axis. Use both forms of the vector $\boldsymbol{u}$ corresponding to equations (A2.1) and (A2.2). Hint: Draw a circle of radius $\|\boldsymbol{y}\|$.

1.3 Perform a QR decomposition of the matrix $\boldsymbol{X}$ from Example PCB 3,

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1.26 \\ 1 & 1.82 \\ 1 & 2.22 \end{bmatrix}$$

using $\boldsymbol{u}$ as in equation (A2.2). Compare the result with that in Appendix 2.

1.4

1.4.1. Perform a QR decomposition of the matrix

$$\boldsymbol{D} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}$$

and obtain the matrix $\boldsymbol{Q}_1$. This matrix is used in Example α-pinene 6, Section 4.3.4.

1.4.2. Calculate $\boldsymbol{Q}_2^{\mathrm{T}}\boldsymbol{y}$, where $\boldsymbol{y} = (50.4, 32.9, 6.0, 1.5, 9.3)^{\mathrm{T}}$, without explicitly solving for $\boldsymbol{Q}_2$.

1.5

1.5.1. Fit the model $\ln(\mathrm{PCB}) = \beta_1 + \beta_2 \mathrm{age}$ to the PCB data and perform a lack of fit analysis of the model. What do you conclude about the adequacy of this model?

1.5.2. Plot the residuals versus age, and assess the adequacy of the model. Now what do you conclude about the adequacy of the model?

1.5.3. Fit the model $\ln(\mathrm{PCB}) = \beta_1 + \beta_2 \mathrm{age} + \beta_3 age^2$ to the PCB data and perform a lack of fit analysis of the model. What do you conclude about the adequacy of this model?

1.5.4. Perform an extra sum of squares analysis to determine whether the quadratic term is a useful addition.

1.5.5. Explain the difference between your answers in (a), (b), and (d).

2

Nonlinear Regression: Iterative Estimation and Linear Approximations

> Although this may seem a paradox, all exact science is dominated by the idea of approximation.
>
> —Bertrand Russell

Linear regression is a powerful method for analyzing data described by models which are linear in the parameters. Often, however, a researcher has a mathematical expression which relates the response to the predictor variables, and these models are usually nonlinear in the parameters. In such cases, linear regression techniques must be extended, which introduces considerable complexity.

2.1 THE NONLINEAR REGRESSION MODEL

A nonlinear regression model can be written

$$Y_n = f(\boldsymbol{x}_n, \boldsymbol{\theta}) + Z_n \tag{2.1}$$

where f is the expectation function and $\boldsymbol{x}_n$ is a vector of associated regressor variables or independent variables for the nth case. This model is of exactly the same form as (1.1) except that the expected responses are nonlinear functions of the parameters. That is, for nonlinear models, *at least one of the derivatives of the expectation function with respect to the parameters depends on at least one of the parameters.*

To emphasize the distinction between linear and nonlinear models, we use $\boldsymbol{\theta}$ for the parameters in a nonlinear model. As before, we use P for the number of parameters.

When analyzing a particular set of data we consider the vectors $\boldsymbol{x}_n$, $n = 1, 2, \ldots, N$, as fixed and concentrate on the dependence of the expected responses on $\boldsymbol{\theta}$. We create the N-vector $\boldsymbol{\eta}(\boldsymbol{\theta})$ with nth element

$$\eta_n(\boldsymbol{\theta}) = f(\boldsymbol{x}_n, \boldsymbol{\theta}) n = 1, \ldots, N$$

and write the nonlinear regression model as

$$\boldsymbol{Y} = \boldsymbol{\eta}(\boldsymbol{\theta}) + \boldsymbol{Z} \tag{2.2}$$

with $\boldsymbol{Z}$ assumed to have a spherical normal distribution. That is,

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

$$\mathrm{Var}(\boldsymbol{Z}) = \mathrm{E}[\boldsymbol{Z}\boldsymbol{Z}^{\mathrm{T}}] = \sigma^2 \boldsymbol{I}$$

as in the linear model.

Example:

Count Rumford of Bavaria was one of the early experimenters on the physics of heat. In 1798 he performed an experiment in which a cannon barrel was heated by grinding it with a blunt bore. When the cannon had reached a steady temperature of 130°F, it was allowed to cool and temperature readings were taken at various times. The ambient temperature during the experiment was 60°F, so [under Newton's law of cooling, which states that $df/dt = -\theta(f - T_0)$, where T_0 is the ambient temperature] the temperature at time t should be

$$f(t, \theta) = 60 + 70e^{-\theta t}$$

Since $\partial f/\partial\theta = -70te^{-\theta t}$ depends on the parameter θ, this model is nonlinear. Rumford's data are presented in Appendix A, Section A.2.

•

Example:

The Michaelis–Menten model for enzyme kinetics relates the initial "velocity" of an enzymatic reaction to the substrate concentration x through the equation

$$f(x, \boldsymbol{\theta}) = \frac{\theta_1 x}{\theta_{2+x}} \tag{2.3}$$

In Appendix A, Section A.3 we present data from Treloar (1974) on the initial rate of a reaction for which the Michaelis–Menten model is believed to be appropriate. The data, for an enzyme treated with Puromycin, are plotted in Figure 2.1.

Differentiating f with respect to θ_1 and θ_2 gives

$$\begin{aligned} \frac{\partial f}{\partial \theta_1} &= \frac{x}{\theta_2 + x} \\ \frac{\partial f}{\partial \theta_2} &= \frac{-\theta_1 x}{(\theta_2 + x)^2} \end{aligned} \tag{2.4}$$

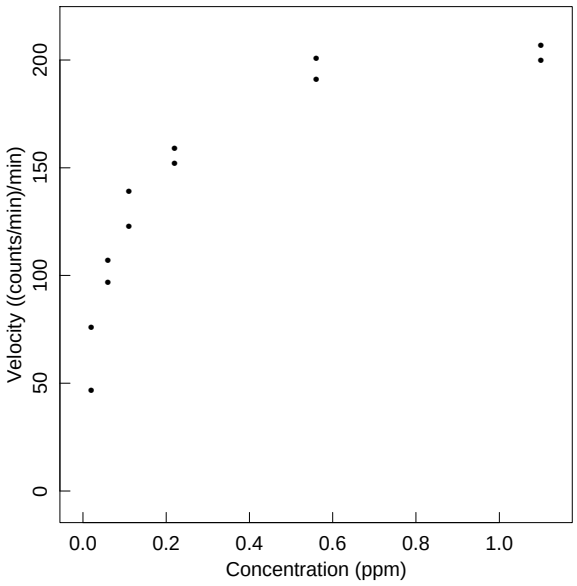

Fig. 2.1 Plot of reaction velocity versus substrate concentration for the Puromycin data.

and since both these derivatives involve at least one of the parameters, the model is recognized as nonlinear. •

2.1.1 Transformably Linear Models

The Michaelis–Menten model (2.3) can be transformed to a linear model by expressing the reciprocal of the velocity as a function of the reciprocal substrate concentration,

$$\begin{aligned} \frac{1}{f} &= \frac{1}{\theta_1} + \frac{\theta_2}{\theta_1}\frac{1}{x} \\ &= \beta_1 + \beta_2 u \end{aligned} \tag{2.5}$$

We call such models *transformably linear*. Some authors use the term "intrinsically linear", but we reserve the term "intrinsic" for a special geometric property of nonlinear models, as discussed in Chapter 7. As will be seen in Chapter 3, transformably linear models have some advantages in nonlinear regression because it is easy to get starting values for some of the parameters.

It is important to understand, however, that a transformation of the data involves a transformation of the disturbance term too, which affects the assumptions on it. Thus, if we assume the model function (2.2) with an additive, spherical normal disturbance term is an appropriate representation of the experimental situation, then these same assumptions will not be appropriate for

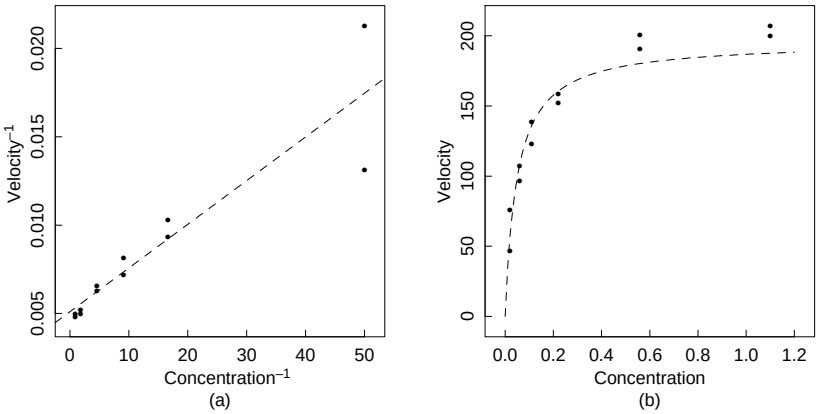

Fig. 2.2 Plot of inverse velocity versus inverse substrate concentration for the Puromycin experiment with the linear regression line (dashed) in part a, and the corresponding fitted curve (dashed) in the original scale in part b.

the transformed data. Hence we should use nonlinear regression on the original data, or else weighted least squares on the transformed data. Sometimes, of course, transforming a data set to induce constant variance also produces a linear expectation function in which case linear regression can be used on the transformed data.

Example:

Because there are replications in the Puromycin data set, it is easy to see from Figure 2.1 that the variance of the original data is constant, and hence that nonlinear regression should be used to estimate the parameters. However, the reciprocal data, plotted in Figure 2.2a, while showing a simple straight line relationship, also show decidely nonconstant variance.

If we use linear regression to fit the model (2.5) to these data, we obtain the estimates

$$\hat{\boldsymbol{\beta}} = (0.005107, 0.0002472)^{\mathrm{T}}$$

corresponding to

$$\hat{\boldsymbol{\theta}} = (195.8, 0.04841)^{\mathrm{T}}$$

The fitted curve is overlaid with the data in the original scale in Figure 2.2b, where we see that the predicted asymptote is too small. Because the variance of the replicates has been distorted by the transformation, the cases with low concentration (high reciprocal concentration) dominate the determination of the parameters and the curve does not fit the data well at high concentrations. •

This example demonstrates two important features. First, it emphasizes the value of replications, because without replications it may not be possible to detect either the constant variance in the original data or the nonconstant

variance in the transformed data; and second, it shows that while transforming can produce simple linear behavior, it also affects the disturbances.

2.1.2 Conditionally Linear Parameters

The Michaelis–Menten model is also an example of a model in which there is a conditionally linear parameter, θ_1. It is *conditionally linear* because the derivative of the expectation function with respect to θ_1 does not involve θ_1. We can therefore estimate θ_1, conditional on θ_2, by a linear regression of velocity $x/(\theta_2 + x)$. Models with conditionally linear parameters enjoy some advantageous properties, which can be exploited in nonlinear regression.

2.1.3 The Geometry of the Expectation Surface

The assumption of a spherical normal distribution for the disturbance term $\boldsymbol{Z}$ leads us to consider the Euclidean geometry of the N-dimensional response space, because again we will be interested in the least squares estimates $\hat{\boldsymbol{\theta}}$ of the parameters. The N-vectors $\boldsymbol{\eta}(\boldsymbol{\theta})$ define a P-dimensional surface called the *expectation surface* in the response space, and the least squares estimates correspond to the point on the expectation surface,

$$\hat{\boldsymbol{\eta}} = \boldsymbol{\eta}(\hat{\boldsymbol{\theta}})$$

which is closest to $\boldsymbol{y}$. That is, $\hat{\boldsymbol{\theta}}$ minimizes the residual sum of squares

$$S(\boldsymbol{\theta}) = \|\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta})\|^2$$

Example:

To illustrate the geometry of nonlinear models, consider the two cases $t = 4$ and $t = 41$ for the Rumford data. Under the assumption that Newton's law of cooling holds for these data, the expected responses are

$$\boldsymbol{\eta}(\theta) = \begin{bmatrix} 60 + 70e^{-4\theta} \\ 60 + 70e^{-41\theta} \end{bmatrix} \theta \geq 0$$

Substituting values for θ in these equations and plotting the points in a 2-dimensional response space gives the 1-dimensional expectation surface (curve) shown in Figure 2.3.

Note that the expectation surface is *curved* and of *finite extent*, which is in contrast to the linear model in which the expectation surface is a plane of infinite extent. Note, too, that points with equal spacing on the parameter line (θ) map to points with unequal spacing on the expectation surface. •

Example:

As another example, consider the three cases from Example Puromycin 2.1: $x = 1.10$, $x = 0.56$, and $x = 0.22$. Under the assumption that the

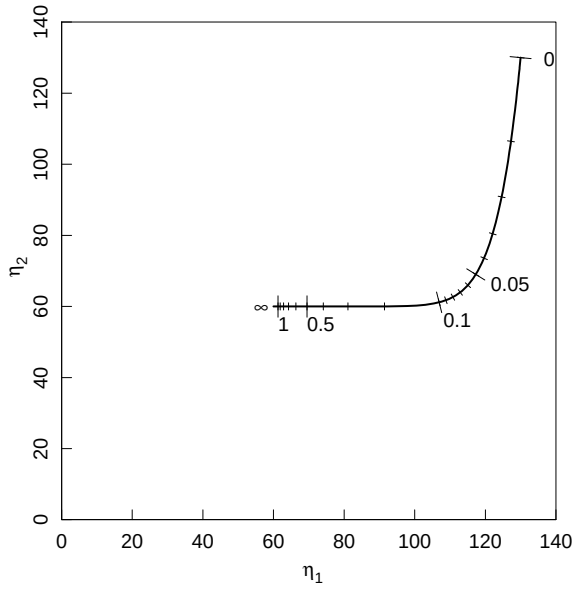

Fig. 2.3 Plot of the expectation surface (solid line) in the response space for the 2-case Rumford data. The points corresponding to $\theta = 0, 0.01, 0.02, \ldots, 0.1, 0.2, \ldots, 1, \infty$ are marked.

expectation function (2.3) is the correct one, the expected responses for these substrate values are

$$\boldsymbol{\eta}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\theta_1(1.10)}{\theta_2+1.10} \\ \frac{\theta_1(0.56)}{\theta_2+0.56} \\ \frac{\theta_1(0.22)}{\theta_2+0.22} \end{bmatrix} \quad \theta_1, \theta_2 \geq 0$$

and so we can plot the expectation surface by substituting values for $\boldsymbol{\theta}$ in these equations. A portion of the 2-dimensional expectation surface for these x values is shown in Figure 2.4. Again, in contrast to the linear model, this expectation surface is not an infinite plane, and in general, straight lines in the parameter plane do not map to straight lines on the expectation surface. It is also seen that unit squares in the parameter plane map to irregularly shaped areas on the expectation surface and that the sizes of these areas vary. Thus, the Jacobian determinant is not constant, which can be seen analytically, of course, because the derivatives (2.4) depend on $\boldsymbol{\theta}$.

For this model, there are straight lines on the expectation surface in Figure 2.4 corresponding to the θ_1 parameter lines (lines with θ_2 held constant), reflecting the fact that θ_1 is conditionally linear. However, the θ_1 parameter lines are neither parallel nor equispaced. The θ_2 lines are not straight, parallel, or equispaced. •

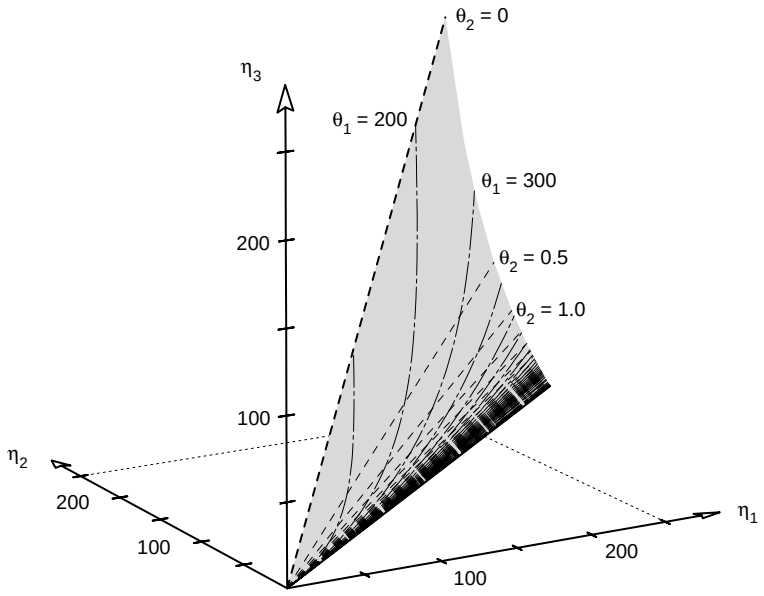

Fig. 2.4 Expectation surface for the 3-case Puromycin example. We show a portion of the expectation surface (shaded) in the expectation space with θ_1 parameter lines (dashed) and θ_2 parameter lines (dot–dashed).

As can be seen from these examples, for nonlinear models with P parameters, it is generally true that:

1. the expectation surface, $\boldsymbol{\eta}(\boldsymbol{\theta})$, is a P-dimensional *curved surface* in the N-dimensional response space;

2. parameter *lines* in the parameter space map to *curves* on the curved expectation surface;

3. the *Jacobian determinant*, which measures how large unit areas in $\boldsymbol{\theta}$ become in $\boldsymbol{\eta}(\boldsymbol{\theta})$, is *not constant*.

We explore these interesting and important aspects of the expectation surface later, but first we discuss how to obtain the least squares estimates $\hat{\boldsymbol{\theta}}$ for the parameters $\boldsymbol{\theta}$. Nonlinear least squares estimation from the point of view of sum of squares contours is given in Section 2.4.

2.2 DETERMINING THE LEAST SQUARES ESTIMATES

The problem of finding the least squares estimates can be stated very simply geometrically—given a data vector $\boldsymbol{y}$, an expectation function $f(\boldsymbol{x}_n, \boldsymbol{\theta})$, and a set of design vectors $\boldsymbol{x}_n$, $n = 1, \ldots, N$

(1)find the point $\hat{\boldsymbol{\eta}}$ on the expectation surface which is closest to $\boldsymbol{y}$, and then (2)determine the parameter vector $\hat{\boldsymbol{\theta}}$ which corresponds to the point $\hat{\boldsymbol{\eta}}$.

For a linear model, step (1) is straightforward because the expectation surface is a plane of infinite extent, and we may write down an explicit expression for the point on that plane which is closest to $\boldsymbol{y}$,

$$\hat{\boldsymbol{\eta}} = \boldsymbol{Q}_1 \boldsymbol{Q}_1^{\mathrm{T}} \boldsymbol{y}$$

For a linear model, step (2) is also straightforward because the P-dimensional parameter plane maps linearly and invertibly to the expectation plane, so once we know where we are on one plane we can easily find the corresponding point on the other. Thus

$$\hat{\boldsymbol{\beta}} = \boldsymbol{R}_1^{-1} \boldsymbol{Q}_1^{\mathrm{T}} \hat{\boldsymbol{\eta}}$$

In the nonlinear case, however, the two steps are very difficult: the first because the expectation surface is curved and often of finite extent (or, at least, has edges) so that it is difficult even to find $\hat{\boldsymbol{\eta}}$, and the second because we can map points easily only in one direction—from the parameter plane to the expectation surface. That is, even if we know $\hat{\boldsymbol{\eta}}$, it is extremely difficult to determine the parameter plane coordinates $\hat{\boldsymbol{\theta}}$ corresponding to that point. To overcome these difficulties, we use iterative methods to determine the least squares estimates.

2.2.1 The Gauss–Newton Method 2 2 1

An approach suggested by Gauss is to use a linear approximation to the expectation function to iteratively improve an initial guess $\boldsymbol{\theta}^0$ for $\boldsymbol{\theta}$ and keep improving the estimates until there is no change. That is, we expand the expectation function $f(\boldsymbol{x}_n, \boldsymbol{\theta})$ in a first order Taylor series about $\boldsymbol{\theta}^0$ as

$$f(\boldsymbol{x}_n, \boldsymbol{\theta}) \approx f(\boldsymbol{x}_n, \boldsymbol{\theta}^0) + \mathrm{v}_{n1}(\theta_1 - \theta_1^0) + \mathrm{v}_{n2}(\theta_2 - \theta_2^0) + \ldots + \mathrm{v}_{nP}(\theta_P - \theta_P^0)$$

where

$$\mathrm{v}_{np} = \left.\frac{\partial f(\boldsymbol{x}_n, \boldsymbol{\theta})}{\partial \theta_p}\right|_{\theta^0} p = 1, 2, \ldots, P$$

Incorporating all N cases, we write

$$\boldsymbol{\eta}(\boldsymbol{\theta}) \approx \boldsymbol{\eta}(\boldsymbol{\theta}^0) + \boldsymbol{V}^0(\boldsymbol{\theta} - \boldsymbol{\theta}^0) \tag{2.6}$$

where $\boldsymbol{V}^0$ is the $N \times P$ derivative matrix with elements v_{np}. This is equivalent to approximating the residuals, $\boldsymbol{z}(\boldsymbol{\theta}) = \boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta})$, by

$$\boldsymbol{z}(\boldsymbol{\theta}) \approx \boldsymbol{y} - [\boldsymbol{\eta}(\boldsymbol{\theta}^0) + \boldsymbol{V}^0\boldsymbol{\delta}] = \boldsymbol{z}^0 - \boldsymbol{V}^0\boldsymbol{\delta} \tag{2.7}$$

where $\boldsymbol{z}^0 = \boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta}^0)$ and $\boldsymbol{\delta} = \boldsymbol{\theta} - \boldsymbol{\theta}^0$.

We then calculate the *Gauss increment* $\boldsymbol{\delta}^0$ to minimize the approximate residual sum of squares $\|\boldsymbol{z}^0 - \boldsymbol{V}^0\boldsymbol{\delta}\|^2$, using

$$\begin{aligned} \boldsymbol{V}^0 &= \boldsymbol{Q}\boldsymbol{R} = \boldsymbol{Q}_1\boldsymbol{R}_1 [\text{cf.}(1.19)] \\ \boldsymbol{w}_1 &= \boldsymbol{Q}_1^{\mathrm{T}}\boldsymbol{z}^0 [\text{cf.}(1.21)] \\ \hat{\boldsymbol{\eta}}^1 &= \boldsymbol{Q}_1\boldsymbol{w}_1 [\text{cf.}(1.23)] \end{aligned}$$

and so

$$\boldsymbol{R}_1\boldsymbol{\delta}^0 = \boldsymbol{w}_1 [\text{cf.}(1.24)]$$

The point

$$\hat{\boldsymbol{\eta}}^1 = \boldsymbol{\eta}(\boldsymbol{\theta}^1) = \boldsymbol{\eta}(\boldsymbol{\theta}^0 + \boldsymbol{\delta}^0)$$

should now be closer to $\boldsymbol{y}$ than $\boldsymbol{\eta}(\boldsymbol{\theta}^0)$, and so we move to this better parameter value $\boldsymbol{\theta}^1 = \boldsymbol{\theta}^0 + \boldsymbol{\delta}^0$ and perform another iteration by calculating new residuals $\boldsymbol{z}^1 = \boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta}^1)$, a new derivative matrix $\boldsymbol{V}^1$, and a new increment. This process is repeated until convergence is obtained, that is, until the increment is so small that there is no useful change in the elements of the parameter vector.

Example:

> To illustrate these calculations, consider the data from Example Puromycin 2.1, with the starting estimates $\boldsymbol{\theta}^0 = (205, 0.08)^{\mathrm{T}}$. The data, along with the fitted values, residuals, and derivatives evaluated at $\boldsymbol{\theta}^0$, are shown in Table 2.1.
>
> Collecting these derivatives into the derivative matrix $\boldsymbol{V}^0$, we then perform a QR decomposition, from which we generate $\boldsymbol{w}_1 = \boldsymbol{Q}_1^{\mathrm{T}}\boldsymbol{z}^0$ and

Table 2.1 Residuals and derivatives for Puromycin data at $\theta = (205, 0.08)^{\mathrm{T}}$.

n	x_n	y_n	η_n^0	z_n^0	v_{n1}^0	v_{n2}^0
1	0.02	76	41.00	35.00	0.2000	–410.00
2	0.02	47	41.00	6.00	0.2000	–410.00
3	0.06	97	87.86	9.14	0.4286	–627.55
4	0.06	107	87.86	19.14	0.4286	–627.55
5	0.11	123	118.68	4.32	0.5789	–624.65
6	0.11	139	118.68	20.32	0.5789	–624.65
7	0.22	159	150.33	8.67	0.7333	–501.11
8	0.22	152	150.33	1.67	0.7333	–501.11
9	0.56	191	179.38	11.62	0.8750	–280.27
10	0.56	201	179.38	21.62	0.8750	–280.27
11	1.10	207	191.10	15.90	0.9322	–161.95
12	1.10	200	191.10	8.90	0.9322	–161.95

then solve for $\boldsymbol{\delta}^0$ using $\boldsymbol{R}_1\boldsymbol{\delta}^0 = \boldsymbol{w}_1$. In this case, $\boldsymbol{\delta}^0 = (8.03, -0.017)^{\mathrm{T}}$ and the sum of squares at $\boldsymbol{\theta}^1 = \boldsymbol{\theta}^0 + \boldsymbol{\delta}^0$ is $S(\boldsymbol{\theta}^1) = 1206$, which is much smaller than $S(\boldsymbol{\theta}^0) = 3155$. We therefore move to $\boldsymbol{\theta}^1 = (213.03, 0.063)^{\mathrm{T}}$ and perform another iteration. •

Example:

As a second example, we consider data on biochemical oxygen demand (BOD) from Marske (1967), reproduced in Appendix A.4. The data are plotted in Figure 2.5. For these data, the model

$$f(x, \boldsymbol{\theta}) = \theta_1(1 - e^{\theta_2 x}) \tag{2.8}$$

is considered appropriate.

Using the starting estimates $\boldsymbol{\theta}^0 = (20, 0.24)^{\mathrm{T}}$, for which $S(\boldsymbol{\theta}^0) = 128.2$, produces an increment to $\boldsymbol{\theta}^1 = (13.61, 0.52)^{\mathrm{T}}$ with $S(\boldsymbol{\theta}^1) = 145.2$. In this case, the sum of squares has increased and so we must modify the increment as discussed below. •

Step Factor

As seen in the last example, the Gauss–Newton increment can produce an *increase* in the sum of squares when the requested increment extends beyond the region where the linear approximation is valid. Even in these circumstances, however, the linear approximation will be a close approximation to the actual surface for a sufficiently small region around $\boldsymbol{\eta}(\boldsymbol{\theta}^0)$. Thus a small step in the direction $\boldsymbol{\delta}^0$ should produce a decrease in the sum of squares. We therefore introduce a *step factor* λ, and calculate

$$\boldsymbol{\theta}^1 = \boldsymbol{\theta}^0 + \lambda\boldsymbol{\delta}^0$$

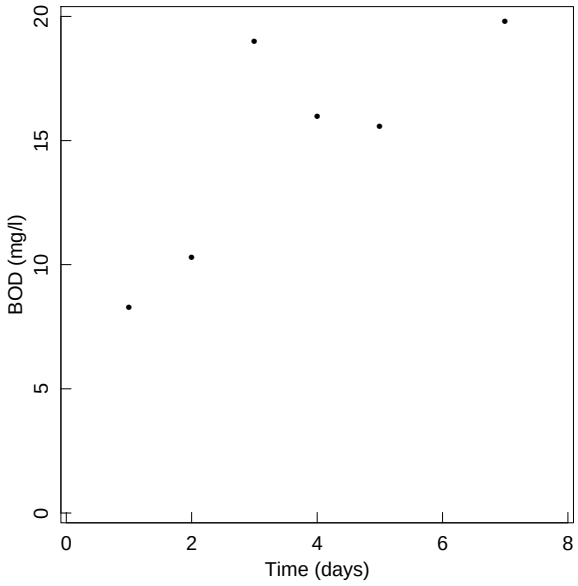

Fig. 2.5 Plot of BOD versus time

where λ is chosen to ensure that

$$S(\boldsymbol{\theta}^1) < S(\boldsymbol{\theta}^0) \tag{2.9}$$

A common method of selecting λ is to start with $\lambda = 1$ and halve it until (2.9) is satisfied. This modification to the Gauss–Newton algorithm was suggested in Box (1960)Hartley (1961)

Example:

> For the data and starting estimates in Example BOD 2.2.1, the value $\lambda = 0.5$ gave a reduced sum of squares, 94.2, at $\boldsymbol{\theta} = (16.80, 0.38)^{\mathrm{T}}$.
>
> •

Pseudocode for the Gauss–Newton algorithm for nonlinear least squares is given in Appendix 3, Section A3.1, together with implementations in GAUSS, S, and SAS/IML.

2.2.2 The Geometry of Nonlinear Least Squares

Geometrically a Gauss–Newton iteration consists of:

1. approximating $\boldsymbol{\eta}(\boldsymbol{\theta})$ by a Taylor series expansion at $\boldsymbol{\eta}^0 = \boldsymbol{\eta}(\boldsymbol{\theta}^0)$,
2. generating the residual vector $\boldsymbol{z}^0 = \boldsymbol{y} - \boldsymbol{\eta}^0$,
3. projecting the residual $\boldsymbol{z}^0$ onto the tangent plane to give $\hat{\boldsymbol{\eta}}^1$,

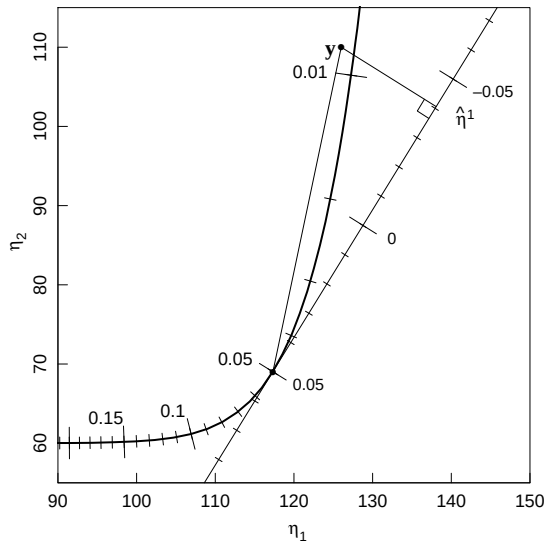

Fig. 2.6 A geometric interpretation of calculation of the Gauss–Newton increment using the 2-case Rumford data. A portion of the expectation surface (heavy solid line) is shown in the response space together with the observed response $\boldsymbol{y}$. Also shown is the projection $\hat{\boldsymbol{\eta}}^1$ of $\boldsymbol{y} - \boldsymbol{\eta}(0.05)$ onto the tangent plane at $\boldsymbol{\eta}(0.05)$ (solid line). The tick marks indicate true positions on the expectation surface and linear approximation positions on the tangent plane.

4. mapping $\hat{\boldsymbol{\eta}}^1$ through the linear coordinate system to produce the increment $\boldsymbol{\delta}^0$, and finally

5. moving to $\boldsymbol{\eta}(\boldsymbol{\theta}^0 + \lambda\boldsymbol{\delta}^0)$.

The first step actually involves two distinct approximations:

1. the *planar* assumption, in which we approximate the expectation surface $\boldsymbol{\eta}(\boldsymbol{\theta})$ near $\boldsymbol{\eta}(\boldsymbol{\theta}^0)$ by its tangent plane at $\boldsymbol{\eta}(\boldsymbol{\theta}^0)$, and

2. the *uniform coordinate* assumption, in which we impose a linear coordinate system $\boldsymbol{V}(\boldsymbol{\theta} - \boldsymbol{\theta}^0)$ on the approximating tangent plane.

We give geometrical interpretations of these steps and assumptions in the following examples.

Example:

For the 2-case Rumford data set of Example Rumford 2.1.3, we plot $\boldsymbol{y}$ and a portion of the expectation surface in Figure 2.6. The expectation surface is a curved line, and the points corresponding to $\theta = 0.01, 0.02, \ldots, 0.2$ are unevenly spaced.

For the initial estimate $\theta^0 = 0.05$, a Gauss–Newton iteration involves the linear approximation

$$\boldsymbol{\eta}(\theta) \approx \boldsymbol{\eta}^0 + \boldsymbol{v}\delta$$

where $\delta = (\theta - 0.05)$, $\boldsymbol{\eta}^0$ is the expectation vector at $\theta = 0.05$,

$$\begin{bmatrix} 60 + 70e^{-4\theta} \\ 60 + 70e^{-41\theta} \end{bmatrix} = \begin{bmatrix} 117.31 \\ 69.01 \end{bmatrix}$$

and $\boldsymbol{v}$ is the derivative vector at $\theta = 0.05$,

$$\boldsymbol{v} = \begin{bmatrix} -70(4)e^{-4\theta} \\ -70(41)e^{-41\theta} \end{bmatrix} = \begin{bmatrix} -229.25 \\ -369.47 \end{bmatrix}$$

The Taylor series approximation, consisting of the tangent plane and the linear coordinate system, is shown as a solid line in Figure 2.6. This replaces the curved expectation surface with the nonlinear parameter coordinates by a linear surface with a uniform coordinate system on it.

Next we use linear least squares to obtain the point $\hat{\boldsymbol{\eta}}^1$ on the tangent line which is closest to $\boldsymbol{y}$. We then calculate the *apparent* parameter increment δ^0 corresponding to $\hat{\boldsymbol{\eta}}^1$ and from this obtain $\theta^1 = \theta^0 + \delta^0$. For this example,

$$\boldsymbol{z}^0 = \begin{bmatrix} 126 \\ 110 \end{bmatrix} - \begin{bmatrix} 117.31 \\ 69.01 \end{bmatrix} = \begin{bmatrix} 8.69 \\ 40.99 \end{bmatrix}$$

so $\hat{\boldsymbol{\eta}}^1 = (138.1, 102.5)^{\mathrm{T}}$, $\delta^0 = -0.091$, and $\theta^1 = 0.05 - 0.091 = -0.041$.

It is clear that the linear approximation increment is too large, since $\theta^1 = -0.041$, whereas we can see from the points on the expectation surface that $\hat{\theta}$ is near 0.01. We must therefore use a step factor to reduce the increment before proceeding. •

Example:

For a two parameter example, we consider the data and the starting values from Example Puromycin 2.2.1. Since the response space is 12-dimensional, we cannot picture it directly, but we can represent the salient features in the 3-dimensional space spanned by the tangent plane and the residual vector. We do this in Figure 2.7, where we show a portion of the curved expectation surface, the residual vector, and the approximating tangent plane. It can be seen that the expectation surface is only slightly curved, and so is well approximated by the tangent plane.

In Figure 2.8*a* we show the parameter curves for $\theta_1 = 200, 210, 220, 230$ and $\theta_2 = 0.06, 0.07, \ldots, 0.1$ projected onto the tangent plane, and in Figure 2.8*b* the corresponding linear approximation lines on the tangent plane. It can be seen that the linear approximation lines match the true parameter curves very well. Also shown on the tangent planes are the points $\boldsymbol{\eta}^0$ and $\hat{\boldsymbol{\eta}}^1$ and in Figure 2.8*a* the projection of the curve $\boldsymbol{\eta}(\boldsymbol{\theta}^0 + \lambda\boldsymbol{\delta}^0)$ for $0 \le \lambda \le 1$. The points corresponding to $\lambda = 0.25, 0.5$, and 1 ($\boldsymbol{\eta}^1$) are marked.

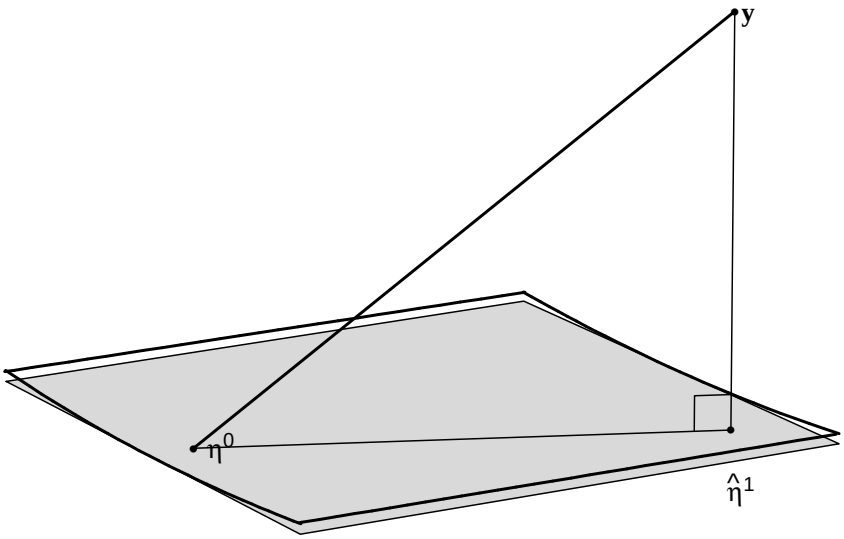

Fig. 2.7 A geometric interpretation of calculation of the Gauss–Newton increment using the full Puromycin data set. We show the projection of a portion of the expectation surface into the subspace spanned by the tangent plane at $\boldsymbol{\eta}^0$ (shaded) and the residual vector $\boldsymbol{y} - \boldsymbol{\eta}^0$. The region on the expectation surface is bordered by the heavy solid lines. Also shown is the projection $\hat{\boldsymbol{\eta}}^1$ of the residual vector onto the tangent plane.

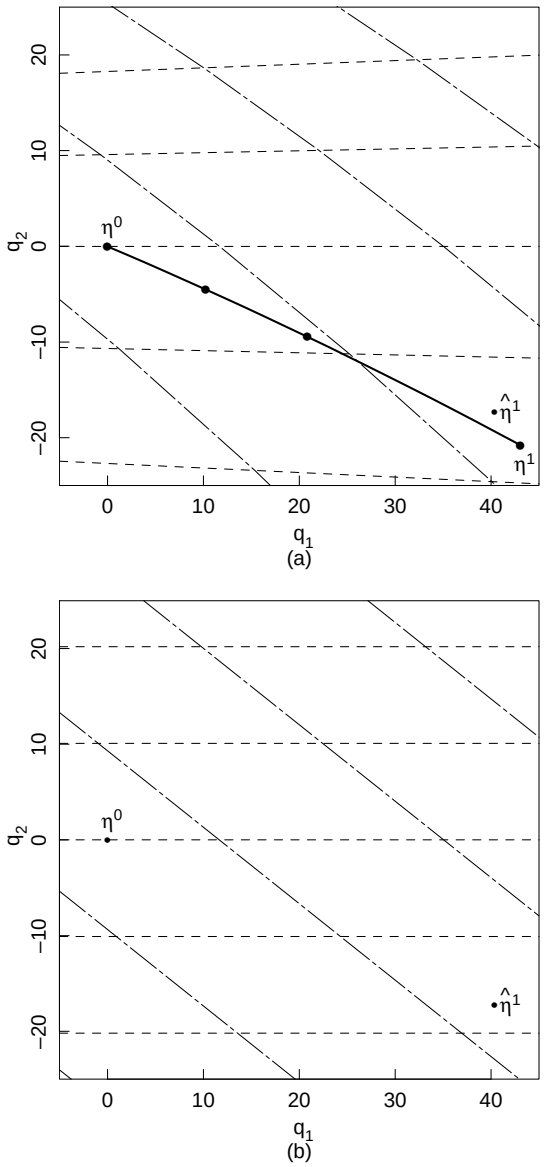

Fig. 2.8 A geometric interpretation of calculation of the Gauss–Newton increment using the full Puromycin data set (continued). The points $\boldsymbol{\eta}^0$ and $\hat{\boldsymbol{\eta}}^1$ are shown in the tangent planes together with the parameter curves in part a and the linear approximation parameter lines in part b. In part a we also show the projection $\boldsymbol{\eta}^1$ of the point $\boldsymbol{\eta}(\boldsymbol{\theta}^0 + \boldsymbol{\delta}^0)$. The curve (heavy solid line) joining $\boldsymbol{\eta}^0$ to $\boldsymbol{\eta}^1$ is the projection of $\boldsymbol{\eta}(\boldsymbol{\theta}^0 + \lambda\boldsymbol{\delta}^0)$ for $0 \leq \lambda \leq 1$. The points corresponding to $\lambda = 0.25$ and 0.5 are marked.

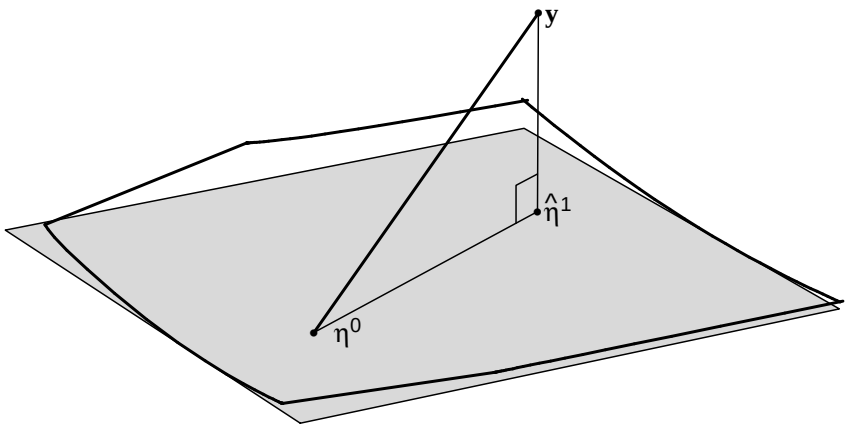

Fig. 2.9 A geometric interpretation of calculation of the Gauss–Newton increment using the BOD data set. We show the projection of a portion of the expectation surface into the subspace spanned by the tangent plane at $\boldsymbol{\eta}^0$ (shaded) and the residual vector $\boldsymbol{y} - \boldsymbol{\eta}^0$. The region on the expectation surface is bordered by the heavy solid lines. Also shown is the projection $\hat{\boldsymbol{\eta}}^1$ of the residual vector onto the tangent plane.

Because the planar and uniform coordinate assumptions are both valid, the points $\hat{\boldsymbol{\eta}}^1$ and $\boldsymbol{\eta}^1$ are close together and are much closer to $\boldsymbol{y}$ than $\boldsymbol{\eta}^0$. In this case, a full step ($\lambda = 1$) can be taken resulting in a decrease in the sum of squares as shown in Example Puromycin 2.2.1. •

Example:

As a second two-parameter example, we consider the data and starting values from Example BOD 2.2.1. In Figure 2.9 we show a portion of the curved expectation surface, the residual vector, and the approximating tangent plane in the space spanned by the tangent plane and the residual vector. It can be seen that the expectation surface is moderately curved, but is still apparently well approximated by the tangent plane. In this example, the edge of the finite expectation surface is shown as the angled solid line along the top edge of the surface.

In Figure 2.10*a* we show the parameter curves for $\theta_1 = 20, 30, \ldots$ and $\theta_2 = 0.2, 0.4, \ldots$ projected onto the tangent plane. In Figure 2.10*b* we show the corresponding linear approximation lines on the tangent plane. In this case, the linear approximation lines do not match the true parameter curves well at all. Also shown on the tangent planes are the points $\boldsymbol{\eta}^0$ and $\hat{\boldsymbol{\eta}}^1$, and in Figure 2.10*a* the projection of the curve

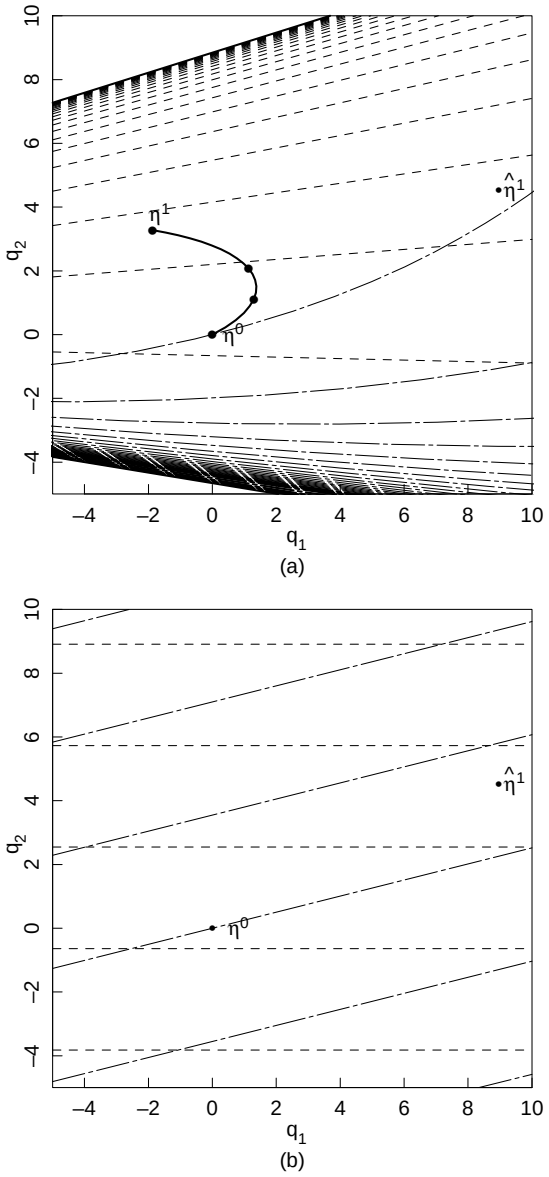

Fig. 2.10 A geometric interpretation of calculation of the Gauss–Newton increment using the BOD data set (continued). The points $\boldsymbol{\eta}^0$and $\hat{\boldsymbol{\eta}}^1$ are shown in the tangent planes together with the parameter curves in part *a* and the linear approximation parameter lines in part *b*. In part *a* we also show the projection $\boldsymbol{\eta}^1$ of the point $\boldsymbol{\eta}(\boldsymbol{\theta}^0+\boldsymbol{\delta}^0)$. The curve (heavy solid line) joining $\boldsymbol{\eta}^0$ to $\boldsymbol{\eta}^1$ is the projection of $\boldsymbol{\eta}(\boldsymbol{\theta}^0+\lambda\boldsymbol{\delta}^0)$ for $0 \leq \lambda \leq 1$. The points corresponding to $\lambda = 0.25$ and 0.5 are marked.

$\boldsymbol{\eta}(\boldsymbol{\theta}^0 + \lambda\boldsymbol{\delta}^0)$ for $0 \leq \lambda \leq 1$. The points corresponding to $\lambda = 0.25, 0.5$, and 1 ($\boldsymbol{\eta}^1$) are marked.

Because the uniform coordinate assumption is not valid this far from $\boldsymbol{\theta}^0$, the points $\hat{\boldsymbol{\eta}}^1$ and $\boldsymbol{\eta}^1$ are widely separated, and in fact $\boldsymbol{\eta}^1$ is farther from $\hat{\boldsymbol{\eta}}^1$ than is $\boldsymbol{\eta}^0$. In this case, the reduced step, $\lambda = 0.5$, is successful, as was shown in Example BOD 2.2.1. •

To summarize, geometrically we are using local information to generate a tangent plane with a linear coordinate system dictated by the derivative vectors, projecting the residual vector onto that tangent plane, and then mapping the tangent plane coordinates to the parameter plane using the linear mapping.

2.2.3 Convergence

We have indicated that the Gauss–Newton iterative method is continued until the values of $\boldsymbol{\theta}$ on successive iterations stabilize. This can be measured by the size of each parameter increment relative to the previous parameter value, which is the basis for one of the common criteria used to declare convergence (Bard, 1974; Draper and Smith, 1981; Jennrich and Sampson, 1968; Ralston and Jennrich, 1978; Kennedy, Jr. and Gentle, 1980). Another criterion for convergence used, for example, in SAS (SAS, 1985), is that the relative change in the sum of squares on successive iterations be small. Himmelblau (1972) recommends that both these criteria be used, since compliance with one does not imply compliance with the other. However, compliance even with both relative change criteria does not guarantee convergence, as discussed in Bates and Watts (1981). Kennedy, Jr. and Gentle (1980) mention a relative step size criterion as well as relative change in the sum of squares and gradient size criteria. Chambers (1977) quotes several other criteria, including the size of the gradient, the size of the Gauss–Newton step, and the fact that the residual vector should be orthogonal to the derivative vectors; but no scale is suggested.

The main criticism of these criteria is that they indicate lack of progress rather than convergence. In most cases, of course, lack of progress occurs because a minimum is encountered: nevertheless, situations can occur where the parameter increment and sum of squares convergence criteria indicate lack of progress and yet a minimum has not been reached.

Examination of the geometry of nonlinear least squares provides a better procedure for determining convergence (Bates and Watts, 1981b). We know that a critical point is reached whenever the residual vector $\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta})$ is orthogonal to the expectation surface and therefore to the tangent plane to the expectation surface at $\boldsymbol{\eta}(\boldsymbol{\theta})$. We can thus adopt orthogonality of the residual vector to the tangent plane as a convergence criterion.

In practice, it would be unusual to obtain exact orthogonality in the presence of numerical roundoff, and we do not want to waste effort calculating small changes in the parameter vector while trying to achieve perfect orthog-

onality. We therefore need to establish a *tolerance level* which we can use to declare the residual vector to be "sufficiently orthogonal." One way to do this is to consider the statistical variability in the least squares estimates.

If we assume that the tangent plane forms a good approximation to the expectation surface near $\hat{\boldsymbol{\theta}}$, so a likelihood region for $\boldsymbol{\theta}$ roughly corresponds to a disk on the tangent plane with a radius proportional to $\sqrt{S(\hat{\boldsymbol{\theta}})}$, then we can measure the relative offset of the current parameter values from the exact least squares estimates by calculating the ratio of the length of the component of the residual vector in the tangent plane to $\sqrt{S(\hat{\boldsymbol{\theta}})}$. When this ratio is small, the numerical uncertainty of the least squares estimates is negligible compared to the statistical uncertainty of the parameters.

Unfortunately, this criterion involves the unknown least squares vector $\hat{\boldsymbol{\theta}}$. We therefore modify the criterion by substituting the current estimate, $\boldsymbol{\theta}^i$, for $\hat{\boldsymbol{\theta}}$, and measure the scaled length of the tangent plane component of the residual vector relative to the scaled length of the orthogonal component of the residual vector at $\boldsymbol{\theta}^i$. This leads to a *relative offset convergence criterion*

$$\frac{\|\boldsymbol{Q}_1^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{\eta}(\boldsymbol{\theta}^i))\|/\sqrt{P}}{\|\boldsymbol{Q}_2^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{\eta}(\boldsymbol{\theta}^i))\|/\sqrt{N-P}} \tag{2.10}$$

where $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$ are the first P and last $N-P$ columns respectively of the matrix $\boldsymbol{Q}$ from a QR decomposition of $\boldsymbol{V}$. The criterion is related to the cotangent of the angle that the residual vector makes with the tangent plane, so that a small relative offset corresponds to an angle near 90°.

To declare convergence, we require the relative offset to be less than 0.001, reasoning that any inferences will not be affected materially by the fact that the current parameter vector is less than 0.1% of the radius of the confidence region disk from the least squares point.

Example:

> We illustrate the convergence criterion and its development with the 2-observation Rumford example. We wish to test whether the parameter value $\theta = 0.01$ could be considered a point of convergence. Figure 2.11 shows a portion of the expectation surface, the observation point $\boldsymbol{y}$, and the tangent plane at $\boldsymbol{\eta}(0.01)$. Also shown is the component of the residual vector in the tangent plane, $\boldsymbol{Q}_1^{\mathrm{T}}\boldsymbol{z}$, and the component orthogonal to the tangent plane, $\boldsymbol{Q}_2^{\mathrm{T}}\boldsymbol{z}$. The tangent plane component is large relative to the orthogonal component, having a relative offset of 1.92, and so we conclude that the residual vector at $\theta = 0.01$ is not sufficiently orthogonal for us to accept $\theta = 0.01$ as the converged value.
>
> •

Convergence implies that the best estimates of the parameters have been obtained, under the assumption that the model is adequate. Before characterizing the precision of the estimates using inference intervals or regions, therefore, we should check the residuals for signs of model inadequacy. A complete discussion of the practical aspects of nonlinear regression is given in

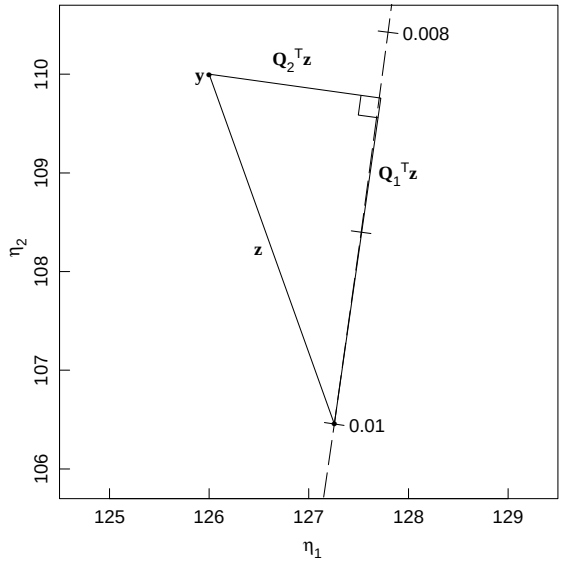

Fig. 2.11 A geometric interpretation of relative offset using the 2-case Rumford data. A portion of the expectation surface (dashed line) is shown in the expectation space together with the residual vector $\boldsymbol{z}$ and its projections into the tangent plane ($\boldsymbol{Q}_1^{\mathrm{T}}\boldsymbol{z}$) and orthogonal to the tangent plane ($\boldsymbol{Q}_2^{\mathrm{T}}\boldsymbol{z}$).

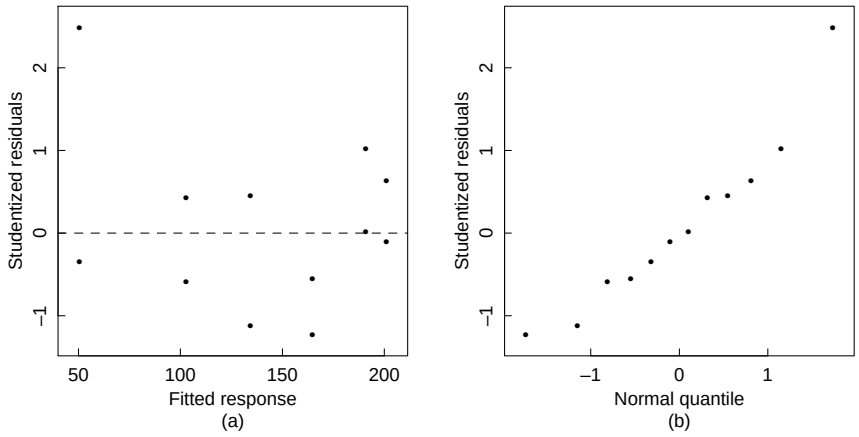

Fig. 2.12 Studentized residuals for the Puromycin data plotted versus fitted values in part *a* and versus normal quantiles in part *b*.

Chapter 3, but in the interests of completeness in analyzing the Puromycin and BOD data, we simply plot the residuals versus the fitted values and using probability plots before continuing.

Example:

Convergence for the Puromycin data was declared at $\hat{\boldsymbol{\theta}} = (212.7, 0.0641)^{\mathrm{T}}$, with $s^2 = 119.5$ on 10 degrees of freedom. Studentized residuals from the least squares fit are plotted in Figure 2.12 versus fitted values in part *a* and as a normal probability plot in part *b*. Although there is one relatively large residual, the overall fit appears adequate, and so we proceed to develop parameter inference regions. •

Example:

Convergence for the BOD data was declared at $\hat{\boldsymbol{\theta}} = (19.143, 0.5311)^{\mathrm{T}}$, with $s^2 = 6.498$ on 4 degrees of freedom. Studentized residuals from the least squares fit are plotted in Figure 2.13 versus fitted values in part *a* and as a normal probability plot in part *b*. Since the residuals are well behaved, we proceed to develop parameter inference regions.

•

2.3 NONLINEAR REGRESSION INFERENCE USING THE LINEAR APPROXIMATION

In the Gauss–Newton algorithm for calculating $\hat{\boldsymbol{\theta}}$, the derivative matrix $\boldsymbol{V}$ is evaluated at each iteration and used to calculate the increment and the convergence criterion. It is natural, then, to apply the linear approximation to *inference* for nonlinear models with the derivative matrix evaluated at the least squares parameter estimates. This yields approximate likelihood,

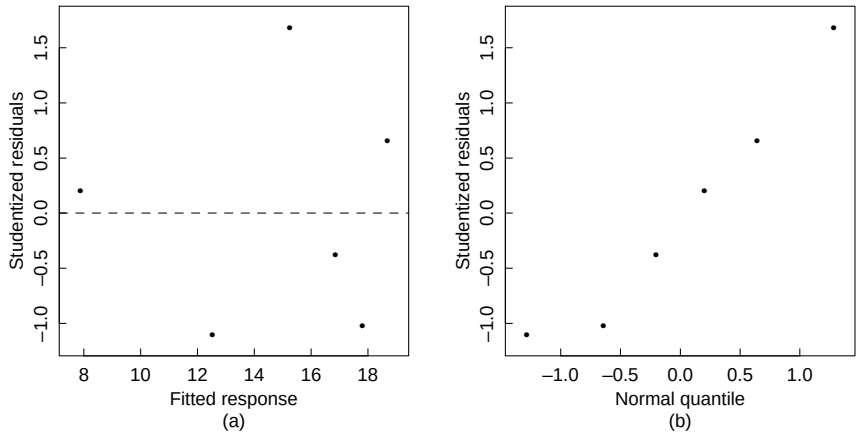

Fig. 2.13 Studentized residuals for the BOD data plotted versus fitted values in part *a* and versus normal quantiles in part *b*.

confidence, or Bayesian HPD regions, based on

$$\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\eta}(\hat{\boldsymbol{\theta}}) + \hat{\boldsymbol{V}}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \tag{2.11}$$

2.3.1 Approximate Inference Regions for Parameters 2 3 1

Recall that in the linear case, a $1 - \alpha$ parameter inference region can be expressed as [cf. (1.9)]

$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^{\mathrm{T}} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \leq P s^2 F(P, N - P; \alpha) \tag{2.12}$$

Geometrically this region results because the expectation surface is a plane and the residual vector is orthogonal to that plane, so the region of plausible values on the expectation plane is a disk. Taking the disk through the linear mapping relating points on the expectation plane to points on the parameter plane, then maps the disk to an ellipsoid on the parameter plane.

Approximate inference regions for a nonlinear model are defined, by analogy with equation (2.12), as

$$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^{\mathrm{T}} \hat{\boldsymbol{V}}^{\mathrm{T}} \hat{\boldsymbol{V}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \leq P s^2 F(P, N - P; \alpha) \tag{2.13}$$

or equivalently

$$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^{\mathrm{T}} \hat{\boldsymbol{R}}_1^{\mathrm{T}} \hat{\boldsymbol{R}}_1 (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \leq P s^2 F(P, N - P; \alpha) \tag{2.14}$$

where the derivative matrix $\hat{\boldsymbol{V}} = \hat{\boldsymbol{Q}}_1 \hat{\boldsymbol{R}}_1$ is evaluated at $\hat{\boldsymbol{\theta}}$. The boundary of this inference region (2.14) is [cf. (1.28)]

$$\left\{ \boldsymbol{\theta} = \hat{\boldsymbol{\theta}} + \sqrt{P s^2 F(P, N - P; \alpha)} \hat{\boldsymbol{R}}_1^{-} 1 \boldsymbol{d} \| \boldsymbol{d} \| = 1 \right\} \tag{2.15}$$

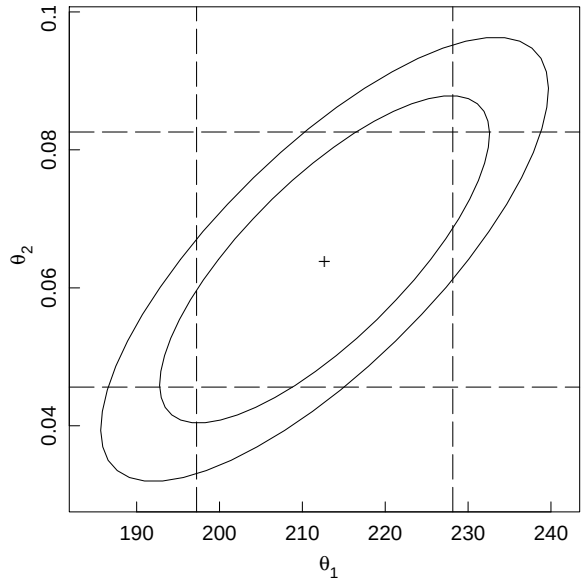

Fig. 2.14 Parameter approximate inference regions for the Puromycin data. We show the least squares estimates (+), the parameter joint 95 and 99% inference regions (solid lines), and the marginal 95% inference intervals (dashed lines).

Similarly, the approximate standard error for θ_p is s times the length of the pth row of $\hat{\boldsymbol{R}}_1^{-}1$'-1p' [cf. (1.33)]. Approximate correlations and standard errors for the parameters are easily calculated by factoring $\hat{\boldsymbol{R}}_{-1}$ into a diagonal matrix [cf. (1.34)] giving the lengths of the rows of $\hat{\boldsymbol{R}}^{-}1$ and a matrix with unit length rows as described in Section 1.2.3. The parameter approximate correlation matrix is calculated as in (1.35).

Example:

Convergence for the Puromycin data was declared at $\hat{\boldsymbol{\theta}} = (212.7, 0.0641)^{\mathrm{T}}$, with $s^2 = 119.5$ on 10 degrees of freedom and

$$\hat{\boldsymbol{R}}_1 = \begin{bmatrix} -2.4441 & 1568.7 \\ 0 & 1320.3 \end{bmatrix}$$

The 95 and 99% approximate joint inference regions were obtained by evaluating (2.15) with $\boldsymbol{d} = (\cos\omega, \sin\omega)^{\mathrm{T}}$ and are plotted in Figure 2.14. To calculate approximate marginal inference intervals, we factor

$$\begin{aligned} \hat{\boldsymbol{R}}_1^{-1} &= \begin{bmatrix} -0.4092 & 0.4861 \\ 0 & 0.0007574 \end{bmatrix} \\ &= \begin{bmatrix} 0.6354 & 0 \\ 0 & 0.0007574 \end{bmatrix} \begin{bmatrix} -0.6439 & 0.7651 \\ 0 & 1.0000 \end{bmatrix} \end{aligned}$$

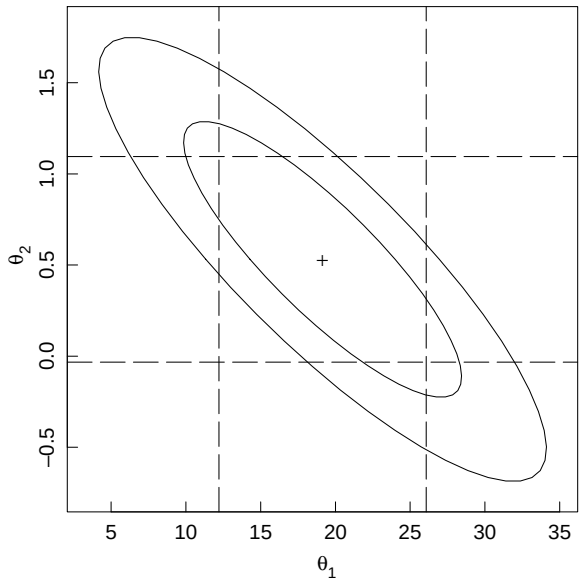

Fig. 2.15 Parameter approximate inference regions for the BOD data. We show the least squares estimates (+), the parameter joint 95 and 99% inference regions (solid lines), and the marginal 95% inference intervals (dashed lines).

so the approximate standard errors are 6.95 and $8.28 \times 10^{-}3$ and the approximate correlation between θ_1 and θ_2 is 0.77. A 95% approximate marginal inference interval for θ_2, for example, is

$$0.0641 \pm \sqrt{1}19.5(0.0007574)t(10; 0.025)$$

or 0.0641 ± 0.0185. The 95% marginal inference intervals for both parameters are shown as dashed lines in Figure 2.14. •

Example:

Convergence for the BOD data was declared at $\hat{\boldsymbol{\theta}} = (19.143, 0.5311)^{\mathrm{T}}$, with $s^2 = 6.498$ on 4 degrees of freedom and

$$\hat{\boldsymbol{R}}_1 = \begin{bmatrix} -1.9556 & -20.4986 \\ 0 & -12.5523 \end{bmatrix}$$

giving approximate standard errors of 2.50 and 0.203.

The 95 and 99% approximate joint inference regions are plotted in Figure 2.15 together with the 95% approximate marginal intervals. Note that the regions include negative values for θ_2, and such values are not physically meaningful. The approximate correlation between θ_1 and θ_2 is -0.85. •

When there are more than two parameters, it is not possible to plot the joint approximate inference region, and so it is common to summarize the inferential

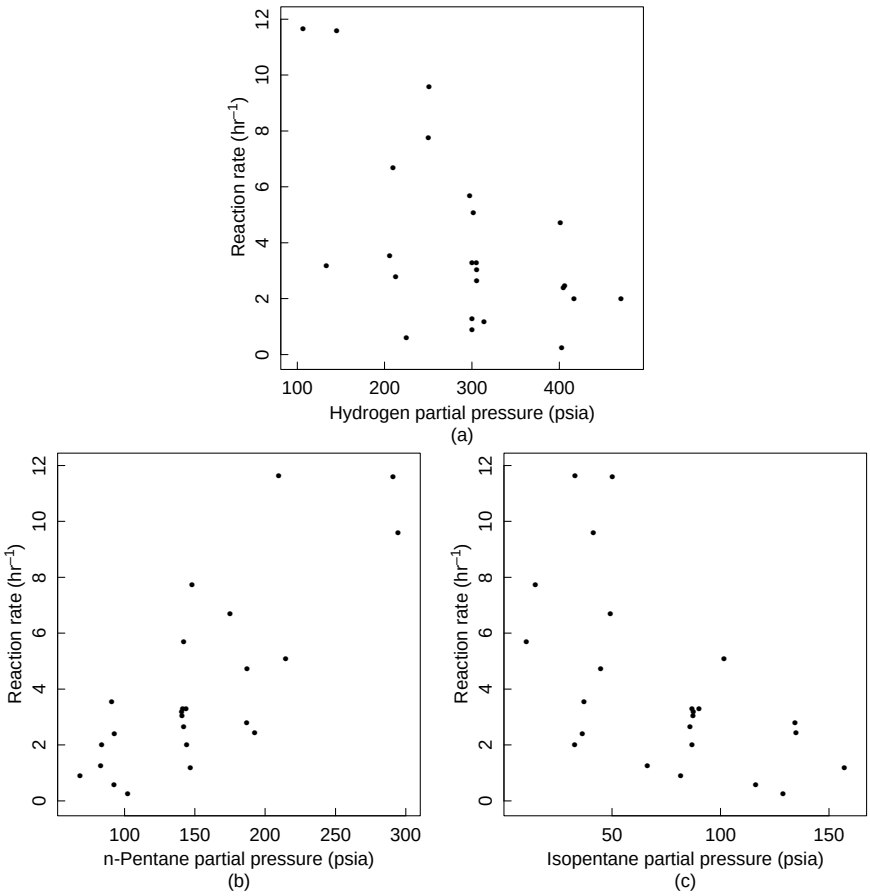

Fig. 2.16 Plots of reaction rate of the isomerization of n-pentane to isopentane versus the partial pressures of hydrogen in part a, n-pentane in part b, and isopentane in part c.

situation by quoting the approximate marginal inference intervals and the parameter correlation matrix and by making pairwise plots of the inference region. More exact methods for summarizing the inferential situtation are presented in Chapter 6.

Example:

Data on the reaction rate of the catalytic isomerization of n-pentane to isopentane versus the partial pressures of hydrogen, n-pentane, and isopentane as given in Carr (1960)Appendix A, Section A.5, and plotted in Figure 2.16. A proposed model function for these data is

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{\theta_1 \theta_3 (x_2 - x_3/1.632)}{1 + \theta_2 x_1 + \theta_3 x_2 + \theta_4 x_3} \tag{2.16}$$

Table 2.2 Parameter summary for the isomerization data

Parameter estimates and summary statistics are given in Table 2.2, and residual plots versus the partial pressures and the fitted values in Figure 2.17. The plots show the residuals are generally well behaved. The summary statistics suggest potential difficulties, since some of the correlations are extremely high and some of the standard errors produce approximate 95% intervals which include negative values, but the parameters must be positive to be physically meaningful. The pairwise plots of the parameter approximate 95% inference region, given in Figure 2.18, clearly extend into negative parameter regions. •

2.3.2 Approximate Inference Bands for the Expected Response

Linear approximation inference intervals and bands for the expected response in nonlinear regression can be generated using the analogs of the equation for linear regression, (1.11) and (1.12). In those equations, we simply replace the estimated value $\boldsymbol{x}_0^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ by $f(\boldsymbol{x}_0, \hat{\boldsymbol{\theta}})$, the matrix $\boldsymbol{X}$ by $\hat{\boldsymbol{V}}$, and the derivative vector $\boldsymbol{x}_0$ by

$$\boldsymbol{v}_0 = \left.\frac{\partial f(\boldsymbol{x}_0, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\right|_{\hat{\theta}}$$

The $1-\alpha$ approximate inference interval is then

$$f(\boldsymbol{x}_0, \hat{\boldsymbol{\theta}}) \pm s\|\boldsymbol{v}_0^{\mathrm{T}}\hat{\boldsymbol{R}}_1^{-1}\|t(N-P;\alpha/2) \quad \text{[cf.(1.36)]}$$

and the $1-\alpha$ approximate inference band is

$$f(\boldsymbol{x}, \hat{\boldsymbol{\theta}}) \pm s\|\boldsymbol{v}^{\mathrm{T}}\hat{\boldsymbol{R}}_1^{-1}\|\sqrt{PF(P, N-P;\alpha)} \quad \text{[cf.(1.37)]}$$

Example:

For the Puromycin data, the estimated response at $x = 0.4$ is 183.3 and the derivative vector is $\boldsymbol{v} = (0.8618, -394.9)^{\mathrm{T}}$, so that, using $\hat{\boldsymbol{R}}_1^{-1}$ from Example Puromycin 6, $\boldsymbol{v}^{\mathrm{T}}\hat{\boldsymbol{R}}_1^{-1} = (-0.3526, 0.1198)$. The inference band at $x = 0.4$ is then $(171.6, 195.0)$. A plot of the approximate 95% inference band is given in Figure 2.19. The band gradually widens from zero width at $x = 0$ to a constant width as $x \to \infty$. •

Example:

The estimated response function for the BOD data and the approximate 95% inference band is plotted in Figure 2.20. The band widens from zero

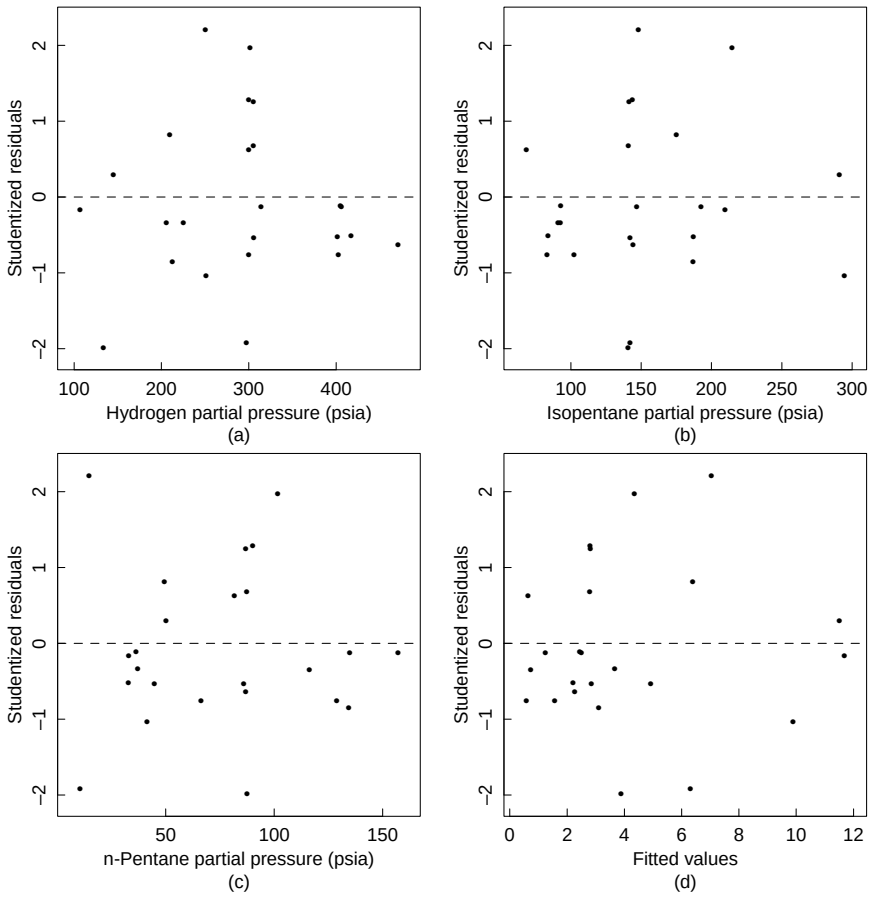

Fig. 2.17 Studentized residuals for the isomerization data are plotted versus the partial pressures of hydrogen in part a, isopentane in part b, and n-pentane in part c, and versus the fitted values in part d.

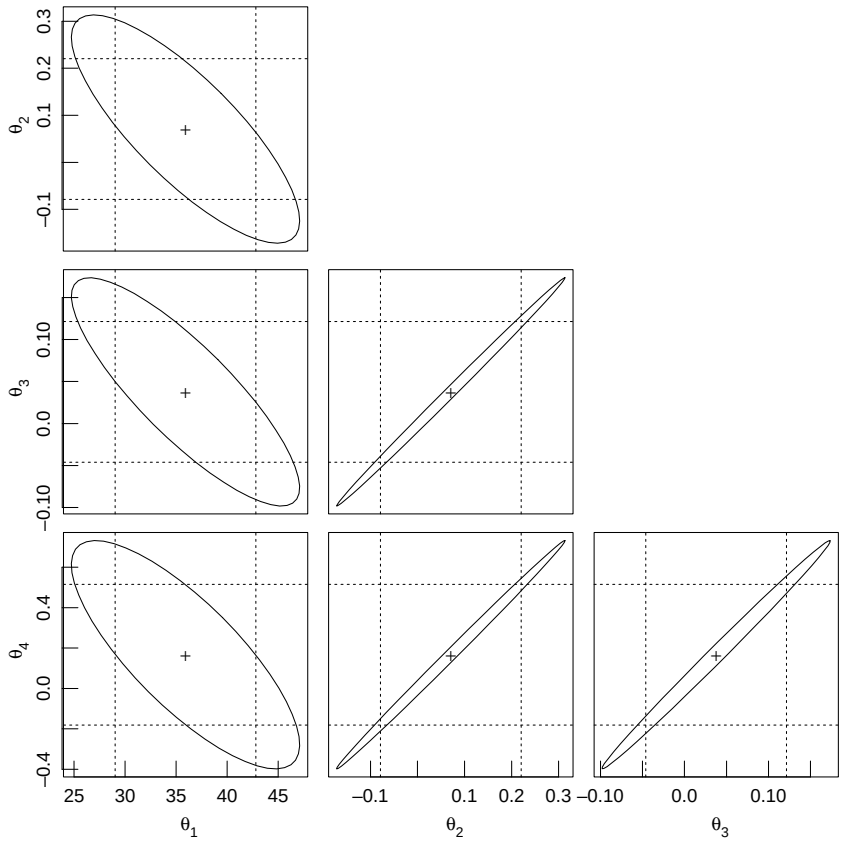

Fig. 2.18 Pairwise plots of the parameter approximate 95% inference region for the isomerization data. For each pair of parameters we show the least squares estimates (+), the parameter approximate joint 95% inference region (solid line), and the approximate marginal 95% inference intervals (dotted lines).

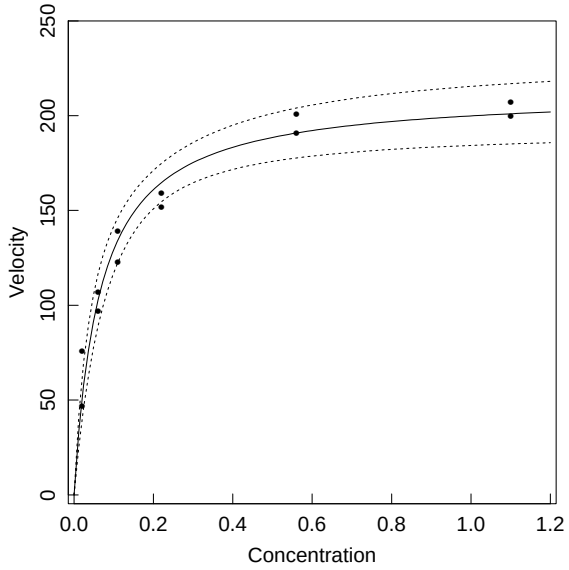

Fig. 2.19 Approximate 95% inference band for the Puromycin data. The fitted expectation function is shown as a solid line, and the 95% inference band is shown as a pair of dotted lines.

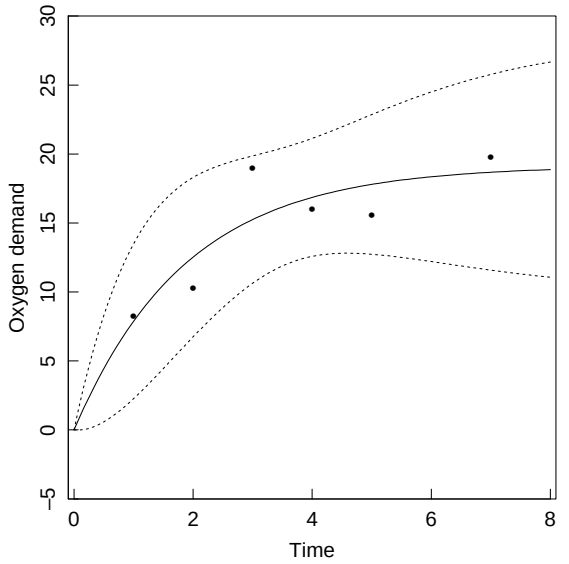

Fig. 2.20 Approximate 95% inference band for the BOD data. The fitted expectation function is shown as a solid line, and the 95% inference band is shown as a pair of dotted lines.

width at $x = 0$, narrows around $x = 4$ and then gradually approaches a constant width as $x \to \infty$. •

Inference bands for nonlinear models behave quite differently from those for linear models. In the above examples, because the functions are constrained to go through the origin, the bands reduce to 0 there. Also, because the model functions approach horizontal asymptotes, the inference bands approach asymptotes. These characteristics differ from those of the inference bands for linear models as exemplified in Figure 1.3. There it is seen that the bands are narrowest near the middle of the data, and expand without limit.

2.4 NONLINEAR LEAST SQUARES VIA SUMS OF SQUARES

Sums of squares occur explicitly in linear and nonlinear least squares because of the assumptions of normality, independence, and constant variance of the disturbances. It is therefore natural to view linear and nonlinear regression via sums of squares, which can help in understanding these two topics. The likelihood approach is especially closely linked to sum of squares contours, because the loglikelihood function is directly proportional to the sum of squares function $S(\boldsymbol{\theta})$.

An important characteristic of linear models is that the sum of squares function $S(\boldsymbol{\beta})$ is quadratic. Because of this, contours of constant sums of squares are well-behaved regular curves or surfaces, such as ellipses and ellipsoids, and so the loglikelihood function can be completely summarized by:

- the minimum value of the sum of squares function, $S(\hat{\boldsymbol{\beta}})$,
- the location of the minimum of the sum of squares function, $\hat{\boldsymbol{\beta}}$, and
- the second derivative (Hessian) of the sum of squares function,

$$\frac{\partial^2 S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathrm{T}}} = \boldsymbol{X}^{\mathrm{T}} \boldsymbol{X}$$

Furthermore, all these quantities can be determined analytically. For nonlinear models, however, the sum of squares function is not regular or well behaved, and so it is difficult to summarize the loglikelihood function.

2.4.1 The Linear Approximation

Linear approximations of the expectation function are used to determine increments while seeking the least squares estimates, and to determine approximate inference regions when convergence has been achieved. The linear approximation to $\boldsymbol{\eta}(\boldsymbol{\theta})$ based at $\boldsymbol{\theta}^0$, (2.6), produces a linear approximation to the residual vector $\boldsymbol{z}(\boldsymbol{\theta})$, (2.7), and hence a *quadratic* approximation $\tilde{S}(\boldsymbol{\theta})$ to the

sum of squares function $S(\boldsymbol{\theta})$, since

$$\begin{aligned} S(\boldsymbol{\theta}) &= \|\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta})\|^2 \\ &= \boldsymbol{z}(\boldsymbol{\theta})^{\mathrm{T}}\boldsymbol{z}(\boldsymbol{\theta}) \approx \tilde{S}(\boldsymbol{\theta}) \\ &= [\boldsymbol{z}^0 - \boldsymbol{V}^0(\boldsymbol{\theta} - \boldsymbol{\theta}^0)]^{\mathrm{T}}[\boldsymbol{z}^0 - \boldsymbol{V}^0(\boldsymbol{\theta} - \boldsymbol{\theta}^0)] \\ &= \boldsymbol{z}^{0\mathrm{T}}\boldsymbol{z}^0 - 2\boldsymbol{z}^{0\mathrm{T}}\boldsymbol{V}^0(\boldsymbol{\theta} - \boldsymbol{\theta}^0) + (\boldsymbol{\theta} - \boldsymbol{\theta}^0)^{\mathrm{T}}\boldsymbol{V}^{0\mathrm{T}}\boldsymbol{V}^0(\boldsymbol{\theta} - \boldsymbol{\theta}^0) \\ &= S(\boldsymbol{\theta}^0) - 2[\boldsymbol{y} - \boldsymbol{\eta}(\boldsymbol{\theta}^0)]^{\mathrm{T}}\boldsymbol{V}^0(\boldsymbol{\theta} - \boldsymbol{\theta}^0) + (\boldsymbol{\theta} - \boldsymbol{\theta}^0)^{\mathrm{T}}\boldsymbol{V}^{0\mathrm{T}}\boldsymbol{V}^0(\boldsymbol{\theta} - \boldsymbol{\theta}^0) \end{aligned} \tag{2.17}$$

The location of the minimum of $\tilde{S}(\boldsymbol{\theta})$ is

$$\boldsymbol{\theta}^1 = \boldsymbol{\theta}^0 + (\boldsymbol{V}^{0\mathrm{T}}\boldsymbol{V}^0)^{-}1\boldsymbol{V}^{0\mathrm{T}}\boldsymbol{z}^0$$

which gives the Gauss–Newton increment.

Note that the quadratic approximation (2.17) is not the second order Taylor series approximation to $S(\boldsymbol{\theta})$ based at $\boldsymbol{\theta}^0$. The Hessian in the Taylor series approximation includes a term involving the second order partial derivatives of the model function with respect to the parameters (see Section 3.5.1).

Contours of the approximate sum of squares function (2.17) are ellipsoids centered at $\boldsymbol{\theta}^1$ and of the form

$$(\boldsymbol{\theta} - \boldsymbol{\theta}^1)^{\mathrm{T}}\boldsymbol{V}^{0\mathrm{T}}\boldsymbol{V}^0(\boldsymbol{\theta} - \boldsymbol{\theta}^1) = c$$

Of particular interest is the approximating contour

$$(\boldsymbol{\theta} - \boldsymbol{\theta}^1)^{\mathrm{T}}\boldsymbol{V}^{0\mathrm{T}}\boldsymbol{V}^0(\boldsymbol{\theta} - \boldsymbol{\theta}^1) = \boldsymbol{z}^{0\mathrm{T}}\boldsymbol{V}^0(\boldsymbol{V}^{0\mathrm{T}}\boldsymbol{V}^0)^{-1}\boldsymbol{V}^{0\mathrm{T}}\boldsymbol{z}^0$$

which passes through $\boldsymbol{\theta}^0$. If this contour is close to the actual sum of squares contour which passes through $\boldsymbol{\theta}^0$, then we can expect that $\boldsymbol{\theta}^1$ will be close to the optimal value of $\boldsymbol{\theta}$.

Example:

In Figure 2.21 we plot the sum of squares function, $S(\theta)$, for the Rumford data as a solid line. Superimposed on the plot is the approximating quadratic, $\tilde{S}(\theta)$, obtained by taking a linear Taylor series approximation to the expectation function at $\theta^0 = 0.02$, shown as a dashed line.

A careful examination of $S(\theta)$ shows that it is not a parabola but is asymmetric, with a steeper rise to the left of the minimum than to the right. The closeness of $S(\theta)$ to a parabola indicates the small degree of nonlinearity of this model–data set combination. The minimum of the approximating parabola is at 0.008, and so the Gauss–Newton increment is $0.008 - 0.02 = -0.012$. •

Example:

In Figure 2.22 we plot sum of squares contours, $S(\boldsymbol{\theta})$, for the Puromycin data, shown as solid lines, and the location of the minimum, shown as +. Also shown, as a dashed line, is the ellipse derived from the linear approximation to the expectation function at $\boldsymbol{\theta}^0 = (205, 0.08)^{\mathrm{T}}$. The

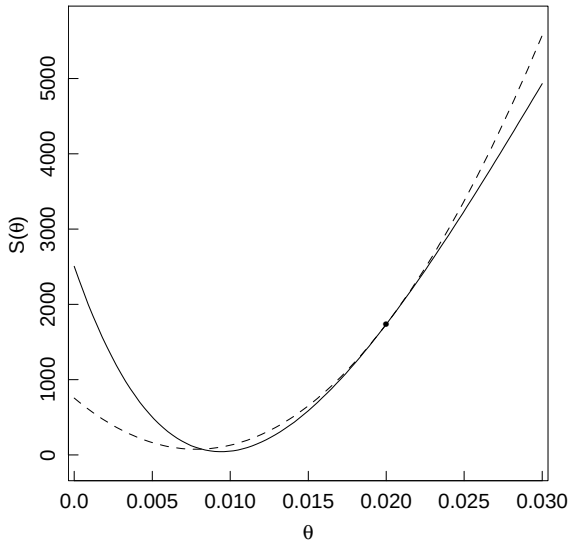

Fig. 2.21 Sum of squares function for the Rumford data. The true sum of squares curve is shown as a solid line, and the parabola from the linear approximation at $\theta^0 = 0.02$ is shown as a dashed line.

approximating paraboloid has the same value and curvature at $\boldsymbol{\theta}^0$ as the true sum of squares surface, and so the location of the minimum of the paraboloid, denoted by $*$, is used as the apparent minimum of the true sum of squares surface. The Gauss increment is therefore the vector joining the starting point $\boldsymbol{\theta}^0$ to the point indicated by $*$.

Because the model–data set combination is not badly nonlinear, the sums of squares contours are quite elliptical, and the minimum of the approximating paraboloid is near the minimum of the true sum of squares surface. •

Example:

In Figure 2.23 we plot sum of squares contours, $S(\boldsymbol{\theta})$, for the BOD data, shown as solid lines, and location of the minimum, shown as $+$. Also shown, as a dashed line, is a portion of the ellipse derived from the linear approximation to the expectation function at $\boldsymbol{\theta}^0 = (20, 0.24)^{\mathrm{T}}$. The center of the ellipse is indicated by $*$.

In this example, the ellipse is a poor approximation to the true contour. The center of the ellipse is not close to the minimum of the true sum of squares surface and furthermore has a true sum of squares greater than that at $\boldsymbol{\theta}^0$. •

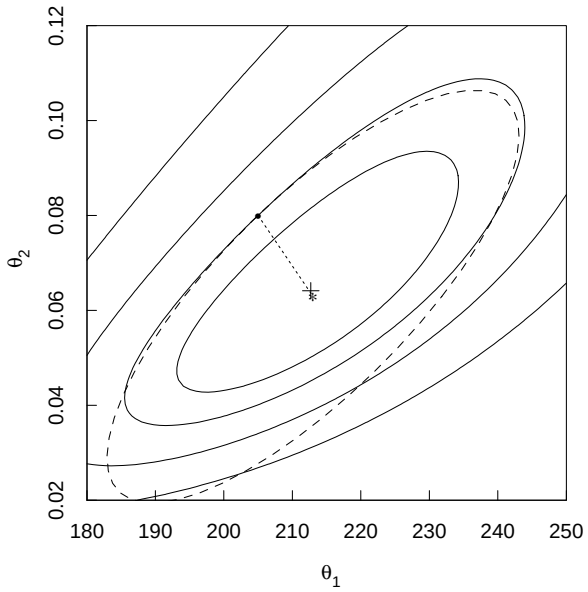

Fig. 2.22 Sum of squares contours for the Puromycin data. True sum of squares contours are shown as solid lines, and the elliptical approximate contour from the linear approximation at $\boldsymbol{\theta}^0 = (205, 0.08)^{\mathrm{T}}$ is shown as a dashed line. The location of the minimum sum of squares ($+$) and the center of the ellipse ($*$) are also shown. The dotted line is the Gauss–Newton increment.

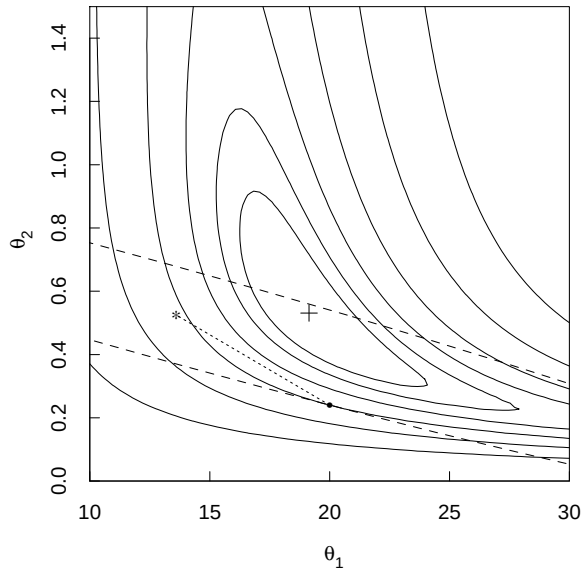

Fig. 2.23 Sum of squares contours for the BOD data. True sum of squares contours are shown as solid lines, and a portion of the elliptical approximate contour from the linear approximation at $\boldsymbol{\theta}^0 = (20, 0.24)^{\mathrm{T}}$ is shown as a dashed line. The location of the minimum sum of squares ($+$) and the center of the ellipse ($*$) are also shown. The dotted line is the Gauss–Newton increment.

2.4.2 Overshoot

The next iteration is carried out from the location of the apparent minimum of the sum of squares surface—provided, of course, that $S(\boldsymbol{\theta}^1)$ is less than $S(\boldsymbol{\theta}^0)$. In the Rumford example and in the Puromycin example, because the nonlinearity is moderate, the sum of squares at $\boldsymbol{\theta}^1$ is less than that at $\boldsymbol{\theta}^0$, and so we can proceed to iterate from $\boldsymbol{\theta}^1$. For the BOD example, however, the sum of squares at $\boldsymbol{\theta}^1$ is greater than that at $\boldsymbol{\theta}^0$, so we have overshot the minimum. By incorporating a step factor, so that only a fraction of the increment is used, we can find a point with a smaller sum of squares, as described in Section 2.2.1.

Problems

2.1 Write a computer routine in a language of your choice to perform nonlinear least squares using the Gauss–Newton approach. Take the function, its derivatives with respect to the parameters, and starting values as input to the routine. If necessary, use the pseudocode in Appendix 3, Section A3.1 for guidance.

2.2 Use a nonlinear least squares routine to fit a model of the form $\beta_1 + \beta_2(\text{age})^\alpha$ to the ln(PCB) data. Use starting values of $(-2.4, 2.3, 0.33)^{\mathrm{T}}$ (the least squares estimates for β_1, β_2 for $\alpha = 0.33$ from Example PCB 2).

2.3

2.3.1. Plot the expectation surface for the Rumford model, using the design $\boldsymbol{x} = (7, 28)^{\mathrm{T}}$. Mark the points on the expectation surface corresponding to the values $\theta = 0, 0.01, \ldots, 0.1, 0.2, \ldots, 1.0, \infty$. Compare this expectation surface with the one based on the design $\boldsymbol{x} = (4, 41)^{\mathrm{T}}$ plotted in Figure 2.3. Which design has smaller overall intrinsic nonlinearity? Which design has smaller overall parameter effects nonlinearity?

2.3.2. Plot the expectation surface for the Rumford model, using the design $\boldsymbol{x} = (12, 14)^{\mathrm{T}}$. Mark the points on the expectation surface corresponding to the values $\theta = 0, 0.01, \ldots, 0.1, 0.2, \ldots, 1.0, \infty$. Compare this expectation surface with the one based on the design $\boldsymbol{x} = (4, 41)^{\mathrm{T}}$ plotted in Figure 2.3 and with that from part (a). Which design has smallest overall intrinsic nonlinearity? Which design has smallest overall parameter effects nonlinearity?

2.3.3. What kind of design would have zero intrinsic nonlinearity everywhere? Why?

2.3.4. Would the design in part (c) have zero parameter effects nonlinearity? Why?

2.4

2.4.1. Plot the expectation surface for the linear model $\ln(\text{PCB}) = \beta \ln(\text{age})$ for the design age $= 5, 10$. Mark the points on the surface corresponding to $\beta = 0, 1, 2, 3$.

2.4.2. Compare this expectation surface and its properties with those of the nonlinear Rumford model shown in Figure 2.3.

2.4.3. Compare this expectation surface and its properties with those of the nonlinear Rumford model plotted in Problem 2.3.

2.5

2.5.1. Generate the expectation vector, the residual vector, the sum of squares $S(\boldsymbol{\theta}^0)$, and the derivative matrix $\boldsymbol{V}^0$ for the data and model from Appendix 4, Section A4.1, at the starting values $\boldsymbol{\theta}^0 = (2.20, 0.26)^{\mathrm{T}}$.

2.5.2. Calculate the increment $\boldsymbol{\delta}^0$ and $S(\boldsymbol{\theta}^1)$, where $\boldsymbol{\theta}^1 = \boldsymbol{\theta}^0 + \lambda\boldsymbol{\delta}^0$, for $\lambda = 0.25, 0.50$, and 1.0. Is a step factor less than 1 necessary in this case?

2.6

2.6.1. Use the fact that, for the model in Problem 2.5, θ_1 is conditionally linear, and generate and plot exact sum of squares contours for the data in Appendix 4, Section A4.1. (That is, for any specified value of θ_2, it is possible to use linear least squares to obtain the conditional estimate $\tilde{\theta}_1$ and to calculate the values of θ_1 which produce a specified sum of squares. By specifying the sum of squares to be that corresponding to a contour value, it is possible to generate the exact coordinates of points on the contour.) Let θ_2 go from 0.12 to 0.3 in steps of 0.01, and use contour values corresponding to 50, 75, and 95% confidence levels. Mark the location of the minimum on the plot.

2.6.2. Compare these contours with those in Figure 2.23. Which data set suffers most from nonlinearity?

2.6.3. Since the data are from the same type of experiment with the same model, how can this difference be explained?

2.7 Plot the point corresponding to $\boldsymbol{\theta}^0$ and the increment $\boldsymbol{\delta}^0$ from Problem 2.5 on the contour plot from Problem 2.6. Mark the points corresponding to the values $\lambda = 0.25, 0.5$, and 0.75 on the increment vector. Is a step factor less than 1 necessary in this case?

2.8

2.8.1. Use the data, model, and starting values from Problem 2.5 in a nonlinear estimation routine to obtain the least squares parameter estimates.

2.8.2. Calculate and plot the linear approximation joint and marginal inference regions on the plot from Problem 2.6.

2.8.3. Are the linear approximation inference regions accurate in this case?

Appendix A Data Sets Used in Examples

A.1 PCB

Data on the concentrations of polychlorinated biphenyl (PCB) residues in a series of lake trout from Cayuga Lake, NY, were reported in Bache et al. (1972) and are reproduced in Table A.1. The ages of the fish were accurately known, because the fish are annually stocked as yearlings and distinctly marked as to year class. Each whole fish was mechanically chopped, ground, and thoroughly mixed, and 5-gram samples taken. The samples were treated and PCB residues in parts per million (ppm) were estimated using column chromatography.

A linear model

$$f(x, \boldsymbol{\beta}) = \beta_1 + \beta_2 x$$

is proposed where f is predicted ln(PCB concentration) and x is $\sqrt[3]{\text{age}}$.

Table A.1 PCB concentration versus age for lake trout.

Age (years)	PCB Conc. (ppm)	Age (years)	PCB Conc. (ppm)
1	0.6	6	3.4
1	1.6	6	9.7
1	0.5	6	8.6
1	1.2	7	4.0
2	2.0	7	5.5
2	1.3	7	10.5
2	2.5	8	17.5
3	2.2	8	13.4
3	2.4	8	4.5
3	1.2	9	30.4
4	3.5	11	12.4
4	4.1	12	13.4
4	5.1	12	26.2
5	5.7	12	7.4

A.2 RUMFORD

Data on the amount of heat generated by friction were obtained by Count Rumford in 1798. A bore was fitted into a stationary cylinder and pressed against the bottom by means of a screw. The bore was turned by a team of horses for 30 minutes, after which Rumford "suffered the thermometer to remain in its place nearly three quarters of an hour, observing and noting down, at small intervals of time, the temperature indicated by it" (Roller, 1950). (See Table A.2)

A model based on Newton's law of cooling was proposed as

$$f(x, \theta) = 60 + 70e^{-\theta x}$$

where f is predicted temperature and x is time.

A.3 PUROMYCIN

Data on the "velocity" of an enzymatic reaction were obtained by Treloar (1974). The number of counts per minute of radioactive product from the reaction was measured as a function of substrate concentration in parts per million (ppm) and from these counts the initial rate, or "velocity," of the reac-

Table A.2 Temperature versus time for Rumford cooling experiment.

Time (min)	Temperature (°F)	Time (min)	Temperature (°F)
4	126	24	115
5	125	28	114
7	123	31	113
12	120	34	112
14	119	37.5	111
16	118	41	110
20	116		

Reprinted with permission from "The Early Development of the Concepts of Temperature and Heat: The Rise and Decline of the Caloric Theory." by Duane Roller, Harvard University Press, 1950.

tion was calculated (counts/min^2). The experiment was conducted once with the enzyme treated with Puromycin, [(a) in Table A.3] and once with the enzyme untreated (b). The velocity is assumed to depend on the substrate concentration according to the Michaelis–Menten equation. It was hypothesized that the ultimate velocity parameter (θ_1) should be affected by introduction of the Puromycin, but not the half-velocity parameter (θ_2).

The Michaelis–Menten model is

$$f(x, \boldsymbol{\theta}) = \frac{\theta_1 x}{\theta_{2+x}}$$

where f is predicted velocity and x is substrate concentration.

A.4 BOD

Data on biochemical oxygen demand (BOD) were obtained by Marske (1967). To determine the BOD, a sample of stream water was taken, injected with soluble organic matter, inorganic nutrients, and dissolved oxygen, and subdivided into BOD bottles. Each bottle was innoculated with a mixed culture of microorganisms, sealed, and incubated at constant temperature, and then the bottles were opened periodically and analyzed for dissolved oxygen concentration, from which the BOD was calculated in milligrams per liter (mg/l). (See Table A.4) The values shown are the averages of two analyses on each bottle.

A model was derived based on exponential decay with a fixed rate constant as

$$f(x, \boldsymbol{\theta}) = \theta_1(1 - e^{\theta_2 x})$$

where f is predicted biochemical oxygen demand and x is time.

Table A.3 Reaction velocity versus substrate concentration for the Puromycin experiment.

Substrate Concentration (ppm)	Velocity (counts/min^2) (a) Treated	(b) Untreated
0.02	76	67
	47	51
0.06	97	84
	107	86
0.11	123	98
	139	115
0.22	159	131
	152	124
0.56	191	144
	201	158
1.10	207	160
	200	

Table A.4 Biochemical oxygen demand versus time.

Time (days)	Biochemical Oxygen Demand (mg/l)	Time (days)	Biochemical Oxygen Demand (mg/l)
1	8.3	4	16.0
2	10.3	5	15.6
3	19.0	7	19.8

Table A.5 Reaction rate for isomerization of n-pentane to isopentane.

Partial Pressure (psia)			Reaction Rate
Hydrogen	n-Pentane	Isopentane	(hr^{-1})
205.8	90.9	37.1	3.541
404.8	92.9	36.3	2.397
209.7	174.9	49.4	6.694
401.6	187.2	44.9	4.722
224.9	92.7	116.3	0.593
402.6	102.2	128.9	0.268
212.7	186.9	134.4	2.797
406.2	192.6	134.9	2.451
133.3	140.8	87.6	3.196
470.9	144.2	86.9	2.021
300.0	68.3	81.7	0.896
301.6	214.6	101.7	5.084
297.3	142.2	10.5	5.686
314.0	146.7	157.1	1.193
305.7	142.0	86.0	2.648
300.1	143.7	90.2	3.303
305.4	141.1	87.4	3.054
305.2	141.5	87.0	3.302
300.1	83.0	66.4	1.271
106.6	209.6	33.0	11.648
417.2	83.9	32.9	2.002
251.0	294.4	41.5	9.604
250.3	148.0	14.7	7.754
145.1	291.0	50.2	11.590

A.5 ISOMERIZATION

Data on the reaction rate of the catalytic isomerization of n-pentane to isopentane versus the partial pressures of hydrogen, n-pentane, and isopentane were given in Carr (1960) and are reproduced in Table A.5. Isomerization is a chemical process in which a complex chemical is converted into more simple units, called isomers: catalytic isomerization employs catalysts to speed the reaction. The reaction rate depends on various factors, such as partial pressures of the products and the concentration of the catalyst. The differential

Table A.6 Relative concentrations of products versus time for thermal isomerization of α-pinene at 189.5°C.

Time (min)	α-Pinene (%)	Dipentene (%)	Alloocimene (%)	Pyronene (%)	Dimer (%)
1230	88.35	7.3	2.3	0.4	1.75
3060	76.4	15.6	4.5	0.7	2.8
4920	65.1	23.1	5.3	1.1	5.8
7800	50.4	32.9	6.0	1.5	9.3
10680	37.5	42.7	6.0	1.9	12.0
15030	25.9	49.1	5.9	2.2	17.0
22620	14.0	57.4	5.1	2.6	21.0
36420	4.5	63.1	3.8	2.9	25.7

reaction rate was expressed as grams of isopentane produced per gram of catalyst per hour (hr^{-1}), and the instantaneous partial pressure of a component was calculated as the mole fraction of the component times the total pressure, in pounds per square inch absolute (psia).

A common form of model for the reaction rate is the Hougen–Watson model (Hougen and Watson, 1947), of which the following is a special case,

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{\theta_1 \theta_3 (x_2 - x_3/1.632)}{1 + \theta_2 x_1 + \theta_3 x_2 + \theta_4 x_3}$$

where f is predicted reaction rate, x_1 is partial pressure of hydrogen, x_2 is partial pressure of isopentane, and x_3 is partial pressure of n-pentane.

A.6 α-PINENE

Data on the thermal isomerization of α-pinene, a component of turpentine, were reported in Fuguitt and Hawkins (1947). In this experiment, the relative concentrations (%) of α-pinene and three by-products were measured at each of eight times, and the relative concentration of a fourth by-product was imputed from the other concentrations. (See Table A.6) The initial concentration of α-pinene was 100%.

A linear kinetic model, shown in Figure A.1, was proposed in Box, Hunter, MacGregor and Erjavec (1973). This model provides for the production of dipentene and alloocimene, which in turn yields α- and β-pyronene and a dimer.

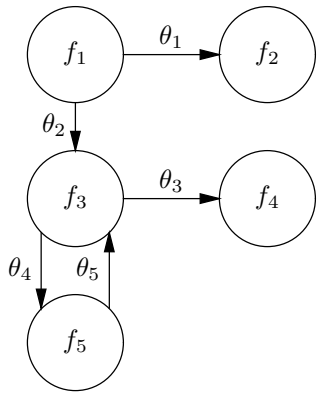

Fig. A.1 System diagram for α-pinene model where f_1 is α-pinene concentration, f_2 is dipentene concentration, f_3 is alloocimene concentration, f_4 is pyronene concentration, and f_5 is dimer concentration.

Table A.7 Sulfisoxazole concentration versus time.

Time (min)	Sulfisoxazole Conc. (μg/ml)	Time (min)	Sulfisoxazole Conc. (μg/ml)
0.25	215.6	3.00	101.2
0.50	189.2	4.00	88.0
0.75	176.0	6.00	61.6
1.00	162.8	12.00	22.0
1.50	138.6	24.00	4.4
2.00	121.0	48.00	0.1

A.7 SULFISOXAZOLE

Data on the metabolism of sulfisoxazole were obtained by Kaplan, Weinfeld, Abruzzo and Lewis (1972) and are reproduced in Table A.7. In this experiment, sulfisoxazole was administered to a subject intravenously, blood samples were taken at specified times, and the concentration of sulfisoxazole in the plasma in micrograms per milliliter (μg/ml) was measured.

For the intravenous data, a 2-compartment model was proposed, which we write as a sum of two exponentials,

$$f(x, \boldsymbol{\theta}) = \theta_1 e^{-\theta_2 x} + \theta_3 e^{-\theta_4 x}$$

where f is predicted sulfisoxazole concentration and x is time.

A.8 LUBRICANT

Data on the kinematic viscosity of a lubricant, in stokes, as a function of temperature (°C), and pressure in atmospheres (atm), were obtained (see Table A.8) and an empirical model was proposed for the logarithm of the viscosity, as discussed in Linssen (1975).

The proposed model is

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{\theta_1}{\theta_2 + x_1} + \theta_3 x_2 + \theta_4 x_2^2 + \theta_5 x_2^3 + (\theta_6 + \theta_7 x_2^2) x_2 \exp\left(\frac{-x_1}{\theta_8 + \theta_9 x_2^2}\right)$$

where f is predicted ln(viscosity), x_1 is temperature, and x_2 is pressure.

A.9 CHLORIDE

Data on the rate of transport of sulfite ions from blood cells suspended in a salt solution were obtained by W. H. Dennis and P. Wood at the University of Wisconsin, and analyzed by Sredni (1970). The chloride concentration (%) was determined from a continuous curve generated from electrical potentials. (See Table A.9)

A model was derived from the theory of ion transport as

$$f(x, \boldsymbol{\theta}) = \theta_1(1 - \theta_2 e^{-\theta_3 x})$$

where f is predicted chloride concentration and x is time.

A.10 ETHYL ACRYLATE

Data on the metabolism of ethyl acrylate were obtained by giving rats a bolus of radioactively tagged ethyl acrylate (Watts, deBethizy and Stiratelli, 1986). Each rat was given a measured dose of the compound via stomach intubation and placed in an enclosed cage from which the air could be drawn through a bubble chamber. The exhalate was bubbled through the chamber, and at a specified time the bubble chamber was replaced by a fresh one, so that the measured response was the accumulated CO_2 during the collection interval. The response reported in Table A.10 is the average, for nine rats, of the amount of accumulated CO_2 normalized by actual dose, in units of grams CO_2 per gram acrylate per gram rat. An empirical model with three exponential terms was determined from inspection of plots of the data and physical reasoning. Logarithms of the integrated function were fitted to logarithms of the data, using the refinements of Section 3.9.

Table A.8 Logarithm of lubricant viscosity versus pressure and temperature.

T = 0°C Pressure (atm)	T = 25°C ln[viscosity (s)]	Pressure (atm)	ln[viscosity (s)]
1.000	5.10595	1.000	4.54223
740.803	6.38705	805.500	5.82452
1407.470	7.38511	1505.920	6.70515
363.166	5.79057	2339.960	7.71659
1.000	5.10716	422.941	5.29782
805.500	6.36113	1168.370	6.22654
1868.090	7.97329	2237.290	7.57338
3285.100	10.47250	4216.890	10.3540
3907.470	11.92720	5064.290	11.9844
4125.470	12.42620	5280.880	12.4435
2572.030	9.15630	3647.270	9.52333
		2813.940	8.34496
T = 37.8°C Pressure (atm)	T = 98.9°C ln[viscosity (s)]	Pressure (atm)	ln[viscosity (s)]
516.822	5.17275	1.000	3.38099
1737.990	6.64963	685.950	4.45783
1008.730	5.80754	1423.640	5.20675
2749.240	7.74101	2791.430	6.29101
1375.820	6.23206	4213.370	7.32719
191.084	4.66060	2103.670	5.76988
1.000	4.29865	402.195	4.08766
2922.940	7.96731	1.000	3.37417
4044.600	9.34225	2219.700	5.83919
4849.800	10.51090	3534.750	6.72635
5605.780	11.82150	4937.710	7.76883
6273.850	13.06800	6344.170	8.91362
3636.720	8.80445	7469.350	9.98334
1948.960	6.85530	5640.940	8.32329
1298.470	6.11898	4107.890	7.13210

Table A.9 Chloride ion concentration versus time.

Time (min)	Conc. (%)	Time (min)	Conc. (%)	Time (min)	Conc. (%)
2.45	17.3	4.25	22.6	6.05	26.6
2.55	17.6	4.35	22.8	6.15	27.0
2.65	17.9	4.45	23.0	6.25	27.0
2.75	18.3	4.55	23.2	6.35	27.0
2.85	18.5	4.65	23.4	6.45	27.0
2.95	18.9	4.75	23.7	6.55	27.3
3.05	19.0	4.85	24.0	6.65	27.8
3.15	19.3	4.95	24.2	6.75	28.1
3.25	19.8	5.05	24.5	6.85	28.1
3.35	19.9	5.15	25.0	6.95	28.1
3.45	20.2	5.25	25.4	7.05	28.4
3.55	20.5	5.35	25.5	7.15	28.6
3.65	20.6	5.45	25.9	7.25	29.0
3.75	21.1	5.55	25.9	7.35	29.2
3.85	21.5	5.65	26.3	7.45	29.3
3.95	21.9	5.75	26.2	7.55	29.4
4.05	22.0	5.85	26.5	7.65	29.4
4.15	22.3	5.95	26.5	7.75	29.4

Reproduced from J. Sredni, "Problems of Design, Estimation, and Lack of Fit in Model Building," Ph.D. Thesis, University of Wisconsin–Madison, 1970, with permission of the author.

Table A.10 Collection intervals and averages of normalized exhaled CO_2.

Collection Interval (hr) Start	CO_2 Length	(g)
0.0	0.25	0.01563
0.25	0.25	0.04190
0.5	0.25	0.05328
0.75	0.25	0.05226
1.0	0.5	0.08850
1.5	0.5	0.06340
2.0	2.0	0.13419
4.0	2.0	0.04502
6.0	2.0	0.02942
8.0	16.0	0.02716
24.0	24.0	0.01037
48.0	24.0	0.00602

Reproduced with permission.

The integrated model is written

$$\begin{aligned} F(\boldsymbol{x}, \boldsymbol{\theta}) &= -\frac{\theta_4 + \theta_5}{\theta_1} e^{-\theta_1 x_1}(1 - e^{-\theta_1 x_2}) \\ &\quad + \frac{\theta_4}{\theta_2} e^{-\theta_2 x_1}(1 - e^{-\theta_2 x_2}) + \frac{\theta_5}{\theta_3} e^{-\theta_3 x_1}(1 - e^{-\theta_3 x_2}) \end{aligned}$$

where F is predicted CO_2 exhaled during an interval, x_1 is interval starting time, and x_2 is interval duration.

A.11 SACCHARIN

Data on the metabolism of saccharin compounds were obtained by Renwick (1982). In this experiment, a rat received a single bolus of saccharin, and the amount of saccharin excreted was measured by collecting urine in contiguous time intervals. The measured response was the level of radioactivity of the urine, which was converted to amount of saccharin in micrograms (μg). (See Table A.11.) An empirical compartment model with two exponential terms was determined from inspection of plots of the data. Logarithms of the integrated function were fitted to logarithms of the data, using the refinements of Section 3.9.

Table A.11 Collection intervals and excreted saccharin amounts.

Collection Interval (min) Start	Saccharin Length	(μg)
0	5	7518
5	10	6275
15	15	4989
30	15	2580
45	15	1485
60	15	861
75	15	561
90	15	363
105	15	300

From "Pharmacokinetics in Toxicology," by A. G. Renwick, in *Principles and Methods of Toxicology*, A. Wallace Hayes, Ed., Raven Press, 1982. Reprinted with permission of the publisher.

The integrated model is written

$$F(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{\theta_3}{\theta_1} e^{-\theta_1 x_1}(1 - e^{-\theta_1 x_2}) + \frac{\theta_4}{\theta_2} e^{-\theta_2 x_1}(1 - e^{-\theta_2 x_2})$$

where F is predicted saccharin excreted during an interval, x_1 is interval starting time, and x_2 is interval duration.

A.12 NITRITE UTILIZATION

Data on the utilization of nitrite in bush beans as a function of light intensity were obtained by J.R. Elliott and D.R. Peirson of Wilfrid Laurier University. Portions of primary leaves from three 16-day-old bean plants were subjected to eight levels of light intensity measured in microeinsteins per square metre per second (μE/m^{2}s) and the nitrite utilization in nanomoles of NO_2^- per gram per hour (nmol/ghr) was measured. The experiment was repeated on a different day. (See Table A.12)

An empirical model was suggested to satisfy the requirements of zero nitrite utilization at zero light intensity and approach to an asymptote as light intensity increased. Two models were fitted which rose to a peak and then began to decline, as described in Section 3.12. These models are

$$f(x, \boldsymbol{\theta}) = \frac{\theta_1 x}{\theta_2 + x + \theta_3 x^2}$$

Table A.12 Nitrite utilization versus light intensity.

Light (μE/m^{2}s)	Nitrite Utilization Intensity (nmol/ghr) Day 1	Day 2
2.2	256	549
	685	1550
	1537	1882
5.5	2148	1888
	2583	3372
	3376	2362
9.6	3634	4561
	4960	4939
	3814	4356
17.5	6986	7548
	6903	7471
	7636	7642
27.0	9884	9684
	11597	8988
	10221	8385
46.0	17319	13505
	16539	15324
	15047	15430
94.0	19250	17842
	20282	18185
	18357	17331
170.0	19638	18202
	19043	18315
	17475	15605

and

$$f(x, \boldsymbol{\theta}) = \theta_1(e^{-\theta_3 x} - e^{-\theta_2 x})$$

where f is predicted nitrite utilization and x is light intensity.

A.13 S-PMMA

Data on the dielectric behavior of syndiotactic poly(methylmethacrylate) (s-PMMA) were obtained by Havriliak, Jr. and Negami (1967). A disk of the polymer was inserted between the two metal electrodes of a dielectric cell which formed one arm of a four-armed electrical bridge. The bridge was powered by an oscillating voltage whose frequency f could be changed from 5 to 500000 hertz (Hz), and bridge balance was achieved using capacitance and conductance standards. The complex dielectric constant was calculated using changes from the standards relative to the cell dielectric constant. Measurements were made by simultaneously adjusting the capacitance (real) and the conductance (imaginary) arms of the bridge when it was excited at a specific frequency. The measured responses were the relative capacitance and relative conductance (dimensionless). (See Table A.13)

The model is an empirical generalization of two models based on theory. It is written

$$f(x, \boldsymbol{\theta}) = \theta_2 + \frac{\theta_1 - \theta_2}{\left[1 + (i2\pi x e^{-\theta_3})^{\theta_4}\right]^{\theta_5}}$$

where f is predicted relative complex impedance and x is frequency.

A.14 TETRACYCLINE

Data on the metabolism of tetracycline were presented in Wagner (1967). In this experiment, a tetracycline compound was administered orally to a subject and the concentration of tetracycline hydrochloride in the serum in micrograms per milliliter (μg/ml) was measured over a period of 16 hours. (See Table A.14)

A 2-compartment model was proposed, and dead time was incorporated as

$$f(x, \boldsymbol{\theta}) = \theta_3[e^{-\theta_1(x-\theta_4)} - e^{-\theta_2(x-\theta_4)}]$$

where f is predicted tetracycline hydrochloride concentration and x is time.

A.15 OIL SHALE

Data on the pyrolysis of oil shale were obtained by Hubbard and Robinson (1950) and are reproduced in Table A.15. Oil shale contains organic matter

Table A.13 Real and imaginary dielectric constant versus frequency for s-PMMA at 86.7°F.

Frequency (Hz)	Relative Impedance Real	Frequency Imag	Relative Impedance (Hz)	Real	Imag
30	4.220	0.136	3000	3.358	0.305
50	4.167	0.167	5000	3.258	0.289
70	4.132	0.188	7000	3.193	0.277
100	4.038	0.212	10000	3.128	0.255
150	4.019	0.236	15000	3.059	0.240
200	3.956	0.257	20000	2.984	0.218
300	3.884	0.276	30000	2.934	0.202
500	3.784	0.297	50000	2.876	0.182
700	3.713	0.309	70000	2.838	0.168
1000	3.633	0.311	100000	2.798	0.153
1500	3.540	0.314	150000	2.759	0.139
2000	3.433	0.311			

From "Analytic Representation of Dielectric Constants: A Complex Multiresponse Problem," by S. Havriliak, Jr. and D. G. Watts, in *Design, Data, and Analysis* , Colin L. Mallows, Ed., Wiley, 1987. Reprinted with permission of the publisher.

Table A.14 Tetracycline concentration versus time.

Time (hr)	Tetracycline Conc. (μg/ml)	Time (hr)	Tetracycline Conc. (μg/ml)
1	0.7	8	0.8
2	1.2	10	0.6
3	1.4	12	0.5
4	1.4	16	0.3
6	1.1		

From "Use of Computers in Pharmacokinetics," by J.G. Wagner, in *Journal of Clinical Pharmacology and Therapeutics*, 1967, **8**, 201. Reprinted with permission of the publisher.

Table A.15 Relative concentration of bitumen and oil versus time and temperature for pyrolysis of oil shale.

T = 673K			T = 698K		
Time (min)	Concentration (%) Bitumen	Oil	Time (min)	Concentration (%) Bitumen	Oil
5	0.0	0.0	5.0	6.5	0.0
7	2.2	0.0	7.0	14.4	1.4
10	11.5	0.7	10.0	18.0	10.8
15	13.7	7.2	12.5	16.5	14.4
20	15.1	11.5	15.0	29.5	21.6
25	17.3	15.8	17.5	23.7	30.2
30	17.3	20.9	20.0	36.7	33.1
40	20.1	26.6	25.0	27.3	40.3
50	20.1	32.4	30.0	16.5	47.5
60	22.3	38.1	40.0	7.2	55.4
80	20.9	43.2	50.0	3.6	56.8
100	11.5	49.6	60.0	2.2	59.7
120	6.5	51.8			
150	3.6	54.7			
T = 723K			T = 748K		
Time (min)	Concentration (%) Bitumen	Oil	Time (min)	Concentration (%) Bitumen	Oil
5.0	8.6	0.0	3.0	0.7	0.0
7.5	15.8	2.9	4.5	17.3	2.9
8.0	25.9	16.5	5.0	23.0	17.3
9.0	25.2	24.4	5.5	24.4	20.9
10.0	26.6	29.5	6.0	23.0	25.9
11.0	33.8	35.2	6.5	33.1	29.5
12.5	25.9	39.5	7.0	31.6	33.8
15.0	20.1	45.3	8.0	20.9	45.3
17.5	12.9	43.1	9.0	10.1	53.2
17.5	9.3	54.6	10.0	4.3	58.2
20.0	3.6	59.7	12.5	0.7	57.5
20.0	2.2	53.9	15.0	0.7	61.1
T = 773K			T = 798K		
Time (min)	Concentration (%) Bitumen	Oil	Time (min)	Concentration (%) Bitumen	Oil
3.0	6.5	0.0	3.00	25.2	20.9
4.0	24.4	23.0	3.25	33.1	25.2
4.5	26.6	32.4	3.50	21.6	17.3
5.0	25.9	37.4	4.00	20.9	36.7
5.5	17.3	45.3	5.00	4.3	56.8
6.0	21.6	45.3	7.00	0.0	61.8
6.5	1.4	57.5			
10.0	0.0	60.4			

From "A Thermal Decomposition Study of Colorado Oil Shale," Hubbard, A.B. and Robinson, W.E., U.S. Bureau of Mines, Rept. Invest. No. 4744, 1950.

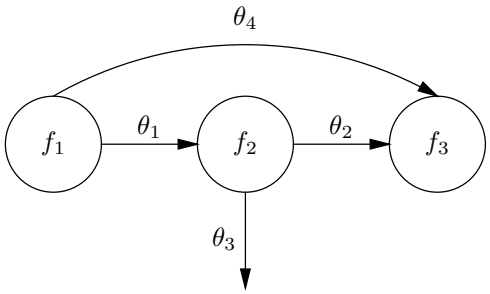

Fig. A.2 System diagram for oil shale model where f_1 is kerogen, f_2 is bitumen, and f_3 is oil.

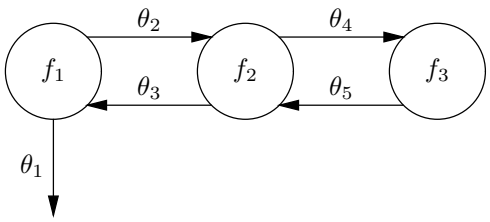

Fig. A.3 System diagram for the tetracycline model where f_1 is the concentration in the sampled compartment. The other compartments do not have a physical interpretation.

which is organically bonded to the structure of the rock: to extract oil from the rock, heat is applied, and so the technique is called pyrolysis. During pyrolysis, the benzene organic material, called kerogen, decomposes chemically to oil and bitumen, and there are unmeasured by-products of insoluble organic residues and light gases. The responses measured were the concentrations of oil and bitumen (%). The initial concentration of kerogen was 100%. Ziegel and Gorman (1980) proposed a linear kinetic model with the system diagram in Figure A.2.

A.16 LIPOPROTEINS

Data on lipoprotein metabolism were reported in Anderson (1983). The response was the concentration, in percent, of a tracer in the serum of a baboon given a bolus injection. Measurements were made at half-day and day intervals. (See Table A.16) An empirical compartment model with two exponential terms was proposed, based on inspection of plots of the data. The system diagram of the final 3-compartment catenary model fitted in Section 5.4 is given in Figure A.3. (A mammary model was also fitted, as discussed in Section 5.4.) It is assumed that the initial concentration in compartment 1 is 100% and that the only response measured is the concentration in compartment 1.

Table A.16 Lipoprotein tracer concentration versus time.

Time (days)	Tracer Conc. (%)	Time (days)	Tracer Conc. (%)
0.5	46.10	5.0	3.19
1.0	25.90	6.0	2.40
1.5	17.00	7.0	1.82
2.0	12.10	8.0	1.41
3.0	7.22	9.0	1.00
4.0	4.51	10.0	0.94

From *Compartmental Modeling and Tracer Kinetics*, D. H. Anderson, p 211, 1983, Springer–Verlag. Reproduced with permission of the author and the publisher.

References

Anderson, D. H. (1983). *Compartmental Modeling and Tracer Kinetics*, Springer-Verlag, Berlin.

Ansley, C. F. (1985). Quick proofs of some regression theorems via the QR algorithm, *American Statistician* **39**: 55–59.

Bache, C. A., Serum, J. W., Youngs, W. D. and Lisk, D. J. (1972). Polychlorinated Biphenyl residues: Accumulation in Cayuga Lake trout with age, *Science* **117**: 1192–1193.

Bard, Y. (1974). *Nonlinear Parameter Estimation*, Academic Press, New York.

Bates, D. M. and Watts, D. G. (1981). A relative offset orthogonality convergence criterion for nonlinear least squares, *Technometrics* **23**: 179–183.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley, New York.

Box, G. E. P. (1960). Fitting empirical data, *Annals of the New York Academy of Sciences* **86**: 792–816.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations, *Journal of the Royal Statistical Society, Ser. B* **26**: 211–252.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*, Addison-Wesley, Reading, MA.

Box, G. E. P. and Tidwell, P. W. (1962). Transformations of the independent variables, *Technometrics* **4**: 531–550.

Box, G. E. P., Hunter, W. G. and Hunter, J. S. (1978). *Statistics for Experimenters*, Wiley, New York.

Box, G. E. P., Hunter, W. G., MacGregor, J. F. and Erjavec, J. (1973). Some problems associated with the analysis of multiresponse models, *Technometrics* **15**(1): 33–51.

Carr, N. L. (1960). Kinetics of catalytic isomerization of n-pentane, *Industrial and Engineering Chemistry* **52**: 391–396.

Chambers, J. M. (1977). *Computational Methods for Data Analysis*, Wiley, New York.

Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983). *Graphical Methods for Data Analysis*, Wadsworth, Belmont, CA.

Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*, Chapman and Hall, London.

Daniel, C. and Wood, F. S. (1980). *Fitting Equations to Data: Computer Analysis of Multifactor Data*, 2nd edn, Wiley, New York.

Dongarra, J. J., Bunch, J. R., Moler, C. B. and Stewart, G. W. (1979). *Linpack Users' Guide*, SIAM, Philadelphia.

Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis*, 2nd edn, Wiley, New York.

Fisher, R. A. (1935). *Design of Experiments*, Oliver and Boyd, London.

Fuguitt, R. E. and Hawkins, J. E. (1947). Rate of the thermal isomerization of α-pinene in the liquid phase, *J. Amer. Chem. Soc.* **69**: 319–322.

Hartley, H. O. (1961). The modified Gauss-Newton method for the fitting of non-linear regression functions by least squares, *Techmometrics* **3**: 269–280.

Havriliak, Jr., S. and Negami, S. (1967). A complex plane representation of dielectric and mechanical relaxation processes in some polymers, *Polymer* **8**: 161–205.

Himmelblau, D. M. (1972). A uniform evaluation of unconstrained optimization techniques, *in* F. A. Lootsma (ed.), *Numerical Methods for Nonlinear Optimization*, Academic Press, London.

Hocking, R. R. (1983). Developments in linear regression methodology: 1959-1982, *Technometrics* **25**: 219–249.

Hougen, O. A. and Watson, K. M. (1947). *Chemical Reaction Principles*, Wiley, New York.

Hubbard, A. B. and Robinson, W. E. (1950). A thermal decomposition study of colorado oil shale, *Technical Report 4744*, U.S. Bureau of Mines.

Jennrich, R. I. and Sampson, P. F. (1968). An application of stepwise regression to non-linear estimation, *Techmometrics* **10**(1): 63–72.

Joiner, B. L. (1981). Lurking variables: Some examples, *American Statistician* **35**: 227–233.

Kaplan, S. A., Weinfeld, R. E., Abruzzo, C. W. and Lewis, M. (1972). Pharmacokinetic profile of sulfisoxazole following intravenous, intramuscular, and oral administration to man, *Journal of Pharmaceutical Sciences* **61**: 773–778.

Kennedy, Jr., W. J. and Gentle, J. E. (1980). *Statistical Computing*, Marcel Dekker, New York.

Linssen, H. N. (1975). Nonlinearity measures: a case study, *Statist. Neerland.* **29**: 93–99.

Marske, D. (1967). *Biochemical oxygen demand data interpretation using sum of squares surface*, PhD thesis, University of Wisconsin–Madison.

Montgomery, D. C. and Peck, E. A. (1982). *Introduction to Linear Regression Analysis*, Wiley, New York.

Ralston, M. L. and Jennrich, R. I. (1978). DUD, a derivative-free algorithm for nonlinear least squares, *Technometrics* **20**: 7–14.

Renwick, A. G. (1982). Pharmacokinetics in toxicology, *in* A. W. Hayes (ed.), *Principles and Methods of Toxicology*, Raven Press, New York.

Roller, D. (1950). *The Early Development of the Concepts of Temperature and Heat: The Rise and Decline of the Caloric Theory*, Harvard University Press, Cambridge, MA.

SAS (1985). *SAS User's Guide: Statistics*, version 5 edn.

Seber, G. A. F. (1977). *Linear Regression Analysis*, Wiley, New York.

Sredni, J. (1970). *Problems of Design, Estimation, and Lack of Fit in Model Building*, PhD thesis, University of Wisconsin–Madison.

Stewart, G. W. (1973). *Introduction to Matrix Computations*, Academic Press, New York.

Treloar, M. A. (1974). *Effects of puromycin on galactosyltransferase of golgi membranes*, Master's thesis, University of Toronto.

Wagner, J. G. (1967). Use of computers in pharmacokinetics, *Clin. Pharmacology and Therapeutics* **8**: 201.

Watts, D. G., deBethizy, D. and Stiratelli, R. G. (1986). Toxicity of ethyl acrylate, *Technical report*, Rohm and Haas Co., Spring House, PA.

Ziegel, E. R. and Gorman, J. W. (1980). Kinetic modelling with multiresponse data, *Technometrics* **22**: 139–151.