

Chapter 6

Process Modeling

with Single-Response Data

The development of a process model typically goes through several stages, outlined in Figure 4.1 and elaborated in this chapter and the next. Typical stages include model formulation; collection of data from new experiments or existing sources; parameter estimation; testing and discrimination of various postulated models; and extensions of the database with sequentially designed experiments. Computational aids are presented in this book for several of these tasks, but prudent investigators will also use diagnostic plots, physical and chemical clues, and experience as aids in checking the data and constructing good candidate models.

The statistical investigation of a model begins with the estimation of its parameters from observations. Chapters 4 and 5 give some background for this step. For single-response observations with independent normal error distributions and given relative precisions, Bayes' theorem leads to the famous method of least squares. Multiresponse observations need more detailed treatment, to be discussed in Chapter 7.

Gauss (1799) gave the first published least-squares solution, but without describing his method, which he identified later as least squares (Plackett 1972, pp. 240 and 246). Legendre (1805) presented least squares as a curve-fitting method, and Gauss (1809) derived the method from elegant probabilistic arguments. Later, setting probabilities aside, Gauss (1823, 1828) showed that linear least squares gives the parameter estimates with least variance among all linear combinations of the observations. Controversy ensued from Gauss' book of 1809, in which he referred to least squares as *principium nostrum* ("our principle") and said he had used it since 1795; Legendre refused to share credit for the discovery. Fascinating accounts of this dispute and the history of least squares are given by Plackett (1972), Stigler (1981, 1986), and G. W. Stewart (1995).

This chapter uses Gauss' 1809 treatment of nonlinear least squares (submitted in 1806, but delayed by the publisher's demand that it be translated into Latin). Gauss weighted the observations according to their precision, as we do in Sections 6.1 and 6.2. He provided *normal equations* for parameter estimation, as we do in Section 6.3, with iteration for models nonlinear in the parameters. He gave efficient algorithms for the parameter

estimates and their variances; we follow his lead in Sections 6.4 and 6.6. Developments since his time have included goodness-of-fit testing (Section 6.5), interval estimation for parameters and functions (Section 6.6), model discrimination (Section 6.6), and design of the next experiment for optimal estimation and/or model discrimination (Section 6.7). These developments are outlined in this chapter and implemented conveniently in the software package GREGPLUS; see Appendix C.

6.1 THE OBJECTIVE FUNCTION $S(\boldsymbol{\theta})$

An objective function $S(\boldsymbol{\theta})$ is presented here for use in Bayesian estimation of the parameter vector $\boldsymbol{\theta}$ in a mathematical model

$$\begin{aligned} y_u &= f_u(\boldsymbol{\xi}_u, \boldsymbol{\theta}) + \varepsilon_u \\ &= f_u(\boldsymbol{\theta}) + \varepsilon_u \quad u = 1, \dots, n \end{aligned} \quad (6.1-1)$$

of single-response events $u = 1, \dots, n$. Such data give one response value y_u per event at settings $\boldsymbol{\xi}_u$ of the independent variables. A value y_u may be directly observed or may be a chosen function of observations from that event; in either case, corrections for any known causes of systematic error should be included in y_u . Each value y_u is then modeled as an expectation function $f_u(\boldsymbol{\xi}_u, \boldsymbol{\theta})$ plus an independent random error ε_u with probability density

$$p(\varepsilon_u | \sigma_u) = \frac{1}{\sqrt{2\pi}\sigma_u} \exp \left[-\frac{\varepsilon_u^2}{2\sigma_u^2} \right] \quad u = 1, \dots, n \quad (6.1-2)$$

based on the central limit theorem of Chapter 4 and on the postulate that the proposed model form $f(\boldsymbol{\xi}_u, \boldsymbol{\theta})$ is true. The predictive probability density¹ for the u th event is then

$$p(y_u | \boldsymbol{\theta}, \sigma_u) = \frac{1}{\sqrt{2\pi}\sigma_u} \exp \left\{ -\frac{[y_u - f_u(\boldsymbol{\theta})]^2}{2\sigma_u^2} \right\} \quad u = 1, \dots, n \quad (6.1-3)$$

and the predictive probability density for the full data vector \mathbf{y} is the product of these independent functions:

$$p(\mathbf{y} | \boldsymbol{\theta}, \sigma_1, \dots, \sigma_n) = \left[\prod_{u=1}^n \frac{1}{\sqrt{2\pi}\sigma_u} \right] \exp \left\{ -\sum_{u=1}^n \frac{[y_u - f_u(\boldsymbol{\theta})]^2}{2\sigma_u^2} \right\} \quad (6.1-4)$$

¹ This density times $d\varepsilon$ is the probability that a prospective observation y_u at $\boldsymbol{\xi}_u$ will fall between $f(\boldsymbol{\xi}_u, \boldsymbol{\theta}) - d\varepsilon/2$ and $f(\boldsymbol{\xi}_u, \boldsymbol{\theta}) + d\varepsilon/2$ if the model form is true.

Relative weights are assigned to the values y_u in the forms

$$w_u = \sigma^2 / \sigma_u^2 \quad u = 1, \dots, n \quad (6.1-5)$$

where σ^2 is the variance of observations of unit weight and σ_u^2 is the variance assigned to observations at ξ_u . Unit weights ($w_u = 1$) are appropriate for data of uniform precision; other weightings are discussed in Section 6.2. Equation (6.1-4) then takes the form

$$p(\mathbf{y}|\boldsymbol{\theta}, \sigma, \mathbf{w}) = \left[\prod_{u=1}^n \frac{\sqrt{w_u}}{\sqrt{2\pi\sigma}} \right] \exp \left[-\frac{S(\boldsymbol{\theta})}{2\sigma^2} \right] \quad (6.1-6)$$

with the weighted sum-of-squares function

$$\begin{aligned} S(\boldsymbol{\theta}) &= \sum_{u=1}^n (\sqrt{w_u} [y_u - f_u(\boldsymbol{\theta})])^2 \\ &= \sum_{u=1}^n [Y_u - F_u(\boldsymbol{\theta})]^2 \\ &= \sum_{u=1}^n E_u(\boldsymbol{\theta})^2 \end{aligned} \quad (6.1-7)$$

Here the notations

$$Y_u = \sqrt{w_u} y_u; \quad F_u(\boldsymbol{\theta}) = \sqrt{w_u} f_u(\boldsymbol{\theta}); \quad E_u(\boldsymbol{\theta}) = Y_u - F_u(\boldsymbol{\theta}) \quad (6.1-8)$$

have reduced $S(\boldsymbol{\theta})$ to a simple sum of squares of the weighted errors $E_u(\boldsymbol{\theta})$. Correspondingly, Eq. (6.1-3) yields the following predictive probability density for a weighted observation Y_u ,

$$p(Y_u|\boldsymbol{\theta}, \sigma) = \frac{p(y_u|\boldsymbol{\theta}, \sigma)}{\sqrt{w_u}} = (\sqrt{2\pi\sigma})^{-1} \exp \left[-\frac{E_u^2}{2\sigma^2} \right] \quad u = 1, \dots, n \quad (6.1-9)$$

and for a vector Y_1, \dots, Y_n of weighted independent observations, it gives

$$p(\mathbf{Y}|\boldsymbol{\theta}, \sigma) = (\sqrt{2\pi\sigma})^{-n} \exp \left[-\frac{S(\boldsymbol{\theta})}{2\sigma^2} \right] \quad (6.1-10)$$

When \mathbf{Y} is given instead of $\boldsymbol{\theta}$ and σ , we call this function the *likelihood*,

$$\ell(\boldsymbol{\theta}, \sigma|\mathbf{Y}) = (\sqrt{2\pi\sigma})^{-n} \exp \left[-\frac{S(\boldsymbol{\theta})}{2\sigma^2} \right] \quad (6.1-11)$$

in the manner of Eq. (5.1-6).

To apply Bayes' theorem, we need a prior probability density for the unknowns, $\boldsymbol{\theta}$ and σ . Treating $\boldsymbol{\theta}$ and σ as independent *a priori* and $p(\boldsymbol{\theta})$ as uniform over the permitted range of $\boldsymbol{\theta}$, we obtain the joint prior density

$$p(\boldsymbol{\theta}, \sigma) \propto \begin{cases} \sigma^{-1} & \text{for permitted values of } \boldsymbol{\theta} \\ 0 & \text{otherwise} \end{cases} \quad (6.1-12)$$

consistent with the Jeffreys prior of Eq. (5.5-12). Multiplication of the likelihood $\ell(\boldsymbol{\theta}, \sigma | \mathbf{Y})$ by this prior density, in accordance with Bayes' theorem, then gives the posterior density

$$p(\boldsymbol{\theta}, \sigma | \mathbf{Y}) \propto \sigma^{-(n+1)} \exp \left[-\frac{S(\boldsymbol{\theta})}{2\sigma^2} \right] \quad (6.1-13)$$

over the permitted range of $\boldsymbol{\theta}$. This probability density takes its largest value at the least sum of squares $S(\boldsymbol{\theta})$ in the permitted range of $\boldsymbol{\theta}$. Of course, one should examine not only the least-squares $\boldsymbol{\theta}$ value (the so-called *point estimate* $\hat{\boldsymbol{\theta}}$), but also its uncertainty as indicated by measures such as a 95% posterior probability interval for each estimated parameter θ_i . The measures provided by GREGPLUS are described in Section 6.6, along with corresponding results for auxiliary functions $\phi_i(\boldsymbol{\theta})$ that the user may define.

An alternate argument for minimizing $S(\boldsymbol{\theta})$ is to maximize the function $\ell(\boldsymbol{\theta}, \sigma | \mathbf{Y})$ given in Eq. (6.1-10). This *maximum likelihood* approach, advocated by Fisher (1925), gives the same point estimate $\hat{\boldsymbol{\theta}}$ as does the posterior density function in Eq. (6.1-13). The posterior density function is essential, however, for calculating posterior probabilities for regions of $\boldsymbol{\theta}$ and for rival models, as we do in later sections of this chapter.

The permitted region of $\boldsymbol{\theta}$ can take various forms. Our package GREGPLUS uses a rectangular region

$$\text{BNDLW}(i) \leq \theta_i \leq \text{BNDUP}(i) \quad i = 1, \dots, \text{NPAR} \quad (6.1-14)$$

for a model containing NPAR parameters.

Many methods are available for least-squares calculations. Models linear in $\boldsymbol{\theta}$ allow direct solutions; other models need iteration. The choice of iteration method depends on one's goal. For a mere curve-fit of the data, a direct search procedure such as that of Powell (1965) or of Nelder and Mead (1965) may suffice. But to determine the most important parameters and their most probable values, a method based on derivatives of $S(\boldsymbol{\theta})$ is essential and is followed here.

6.2 WEIGHTING AND OBSERVATION FORMS

Uniform weighting (known as *simple least squares*) is appropriate when the expected variances of the observations are equal and is commonly used when these values are unknown. GREGPLUS uses this weighting when called with $\text{JWT} = 0$; the values $\sqrt{w_u} = 1$ are then provided automatically.

Observations done with uniform *relative* precision are fairly common. For example, reported catalytic rate data often include adjustments of reaction temperature, catalyst activity, and other variables to standard values. These combined adjustments amount to a correction factor C_u for the catalyst mass in each event $u = 1, \dots, n$. Then the total rate of production of any product P in a reactor experiment can be expressed as

$$R_{Pu} \equiv \left(\frac{F \Delta x}{CW} \right)_u = f(\boldsymbol{\xi}_u, \boldsymbol{\theta}) + \varepsilon_u \quad u = 1, \dots, n \quad (6.2-1)$$

Here F is the total molar feed rate to the reactor, Δx is the production of the product P in moles per mole of feed, and W is the catalyst mass.

The function $y_u = \ln R_{Pu}$ has the following variance according to Eq. (4.D-4), with the errors in $\ln F_u$, $\ln \Delta x_u$, $\ln C_u$, and $\ln W_u$ regarded as random independent variables:

$$\text{Var}(\ln R_{Pu}) = \text{Var}(\ln F_u) + \text{Var}(\ln \Delta x_u) + \text{Var}(\ln C_u) + \text{Var}(\ln W_u) \quad (6.2-2)$$

Therefore, the weight w_u for each observation $y_u = \ln(R)_{Pu}$ satisfies

$$w_u = \frac{\sigma^2}{\text{Var}(\ln F_u) + \text{Var}(\ln \Delta x_u) + \text{Var}(\ln C_u) + \text{Var}(\ln W_u)} \quad (6.2-3)$$

according to Eq. (6.1-5). Simple weights $w_u = 1$ for all u are thus appropriate for observations modeled by Eq. (6.1-1) if the terms in Eq. (6.2-3) are independent of u . A somewhat more detailed weighting is used in Example 6.2.

The remaining sections of this chapter are implemented in the package GREGPLUS, described and demonstrated in Appendix C.

6.3 PARAMETRIC SENSITIVITIES; NORMAL EQUATIONS

Gauss (1809) used a Newton-like iteration scheme to minimize $S(\boldsymbol{\theta})$ for nonlinear models; a single iteration suffices for linear models. He approximated the departures of the data from a model as linear expansions $\tilde{E}_u(\boldsymbol{\theta})$ around the starting point $\boldsymbol{\theta}^k$ of the current iteration:

$$\begin{aligned} E_u(\boldsymbol{\theta}) &\approx \tilde{E}_u(\boldsymbol{\theta}) \equiv Y_u - F_u(\boldsymbol{\theta}^k) - \sum_{r=1}^p \left[\frac{\partial F_u}{\partial \theta_r} \Big|_{\boldsymbol{\theta}^k} \right] (\theta_r - \theta_r^k) \\ &\equiv E_u(\boldsymbol{\theta}^k) - \sum_{r=1}^p X_{ur} (\theta_r - \theta_r^k) \quad (u = 1, \dots, n) \end{aligned} \quad (6.3-1)$$

Here X_{ur} denotes $\partial F_u / \partial \theta_r$ evaluated at $\boldsymbol{\theta}^k$. This set of equations can be expressed concisely as

$$\mathbf{E}(\boldsymbol{\theta}) \approx \tilde{\mathbf{E}}(\boldsymbol{\theta}) \equiv \mathbf{E}^k - \mathbf{X}(\boldsymbol{\theta} - \boldsymbol{\theta}^k) \quad (6.3-2)$$

in which $\mathbf{E}(\boldsymbol{\theta})$ is the column vector of functions $E_u(\boldsymbol{\theta})$, $\tilde{\mathbf{E}}(\boldsymbol{\theta})$ is the local linear approximation to $\mathbf{E}(\boldsymbol{\theta})$, and \mathbf{X} is the local matrix of *parametric sensitivities* of the model. The resulting approximation for the error sum of squares is

$$\begin{aligned}\tilde{S}(\boldsymbol{\theta}) &= \tilde{\mathbf{E}}^T(\boldsymbol{\theta})\tilde{\mathbf{E}}(\boldsymbol{\theta}) \\ &= [\mathbf{E}^k - \mathbf{X}(\boldsymbol{\theta} - \boldsymbol{\theta}^k)]^T [\mathbf{E}^k - \mathbf{X}(\boldsymbol{\theta} - \boldsymbol{\theta}^k)] \\ &= \mathbf{E}^{kT}\mathbf{E}^k - 2(\boldsymbol{\theta} - \boldsymbol{\theta}^k)^T \mathbf{X}^T \mathbf{E}^k + (\boldsymbol{\theta} - \boldsymbol{\theta}^k)^T \mathbf{X}^T \mathbf{X} (\boldsymbol{\theta} - \boldsymbol{\theta}^k)\end{aligned}\quad (6.3-3)$$

and its gradient $\partial\tilde{S}/\partial\boldsymbol{\theta}^T$ is

$$\nabla\tilde{S}(\boldsymbol{\theta}) = \mathbf{0} - 2\mathbf{X}^T \mathbf{E}^k + 2\mathbf{X}^T \mathbf{X} (\boldsymbol{\theta} - \boldsymbol{\theta}^k) \quad (6.3-4)$$

Setting $\nabla\tilde{S} = \mathbf{0}$ gives the equation system

$$\mathbf{X}^T \mathbf{X} (\boldsymbol{\theta}^* - \boldsymbol{\theta}^k) = \mathbf{X}^T \mathbf{E}^k \quad (6.3-5)$$

for the locus of level points $\boldsymbol{\theta}^*$ of the current function $\tilde{S}(\boldsymbol{\theta})$. If the matrix $\mathbf{X}^T \mathbf{X}$ is nonsingular, then it is also positive definite (see Problem 6.A) and the locus is the global minimum of $\tilde{S}(\boldsymbol{\theta})$; this outcome is favored by use of well-designed experiments and by absence of unnecessary parameters. A method will be given in Section 6.4 to find a local minimum \tilde{S} within the useful range of Eq. (6.3-2).

The system (6.3-5) can also be written as

$$\mathbf{X}^T [\mathbf{E}^k - \mathbf{X}(\boldsymbol{\theta}^* - \boldsymbol{\theta}^k)] = \mathbf{0} \quad (6.3-6)$$

Thus, the correction vector $[\mathbf{E}^k - \mathbf{X}(\boldsymbol{\theta}^* - \boldsymbol{\theta}^k)]$ is orthogonal (normal) to each column vector of \mathbf{X} . The rows of Eq. (6.3-5) are accordingly known as the *normal equations* of the given problem.

If the model functions $F_u(\boldsymbol{\theta})$ all are linear in $\boldsymbol{\theta}$, then the sensitivity matrix \mathbf{X} is constant and a single application of Eq. (6.3-5) will give the least-squares solution. In practice, \mathbf{X} is often far from constant, making iteration necessary as described in the following section.

One can extend Eq. (6.3-1) to second order to get a fuller quadratic function $\tilde{S}(\boldsymbol{\theta})$. This gives the *full Newton* equations of the problem, whereas Eqs. (6.3-3)–(6.3-5) are called the *Gauss-Newton* equations. Either form can be expressed as a symmetric matrix expansion

$$\tilde{S}(\boldsymbol{\theta}) = S(\boldsymbol{\theta}^k) + \left[(\boldsymbol{\theta} - \boldsymbol{\theta}^k)^T \vdots -1 \right] \begin{bmatrix} \mathbf{A}_{\theta\theta} & \vdots & \mathbf{A}_{\theta L} \\ \dots\dots\dots & & \\ \mathbf{A}_{\theta L}^T & \vdots & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\theta} - \boldsymbol{\theta}^k \\ \dots\dots\dots \\ -1 \end{bmatrix} \quad (6.3-7)$$

with submatrices defined as follows

$$\mathbf{A}_{\theta L} = -\frac{1}{2} \frac{\partial S}{\partial \boldsymbol{\theta}} \Big|_{\boldsymbol{\theta}^k} = \mathbf{X}^T \mathbf{E}^k \quad (\text{Always}) \quad (6.3-8)$$

$$\mathbf{A}_{\theta\theta} = \frac{1}{2} \frac{\partial^2 \tilde{S}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \begin{cases} \mathbf{X}^T \mathbf{X} & (\text{Gauss-Newton}) \\ \mathbf{X}^T \mathbf{X} - \left\{ \frac{1}{2} \sum_{u=1}^n E_u \frac{\partial^2 F_u}{\partial \theta_r \partial \theta_s} \Big|_{\boldsymbol{\theta}^k} \right\} & (\text{full Newton}) \end{cases} \quad (6.3-9)$$

Here the sum within the curly brackets denotes a square matrix with the indicated elements. The normal equations¹ in either form give

$$\mathbf{A}_{\theta\theta}(\boldsymbol{\theta} - \boldsymbol{\theta}^k) = \mathbf{A}_{\theta L} \quad (6.3-10)$$

for the locus of level points of the quadratic function \tilde{S} . Maxima and saddle points of $\tilde{S}(\boldsymbol{\theta})$ are possible for the full Newton equations, because positive definiteness of $\mathbf{A}_{\theta\theta}$ is no longer guaranteed when the curly-bracketed term of Eq. (6.3-9) is included; however, GREGPLUS avoids such points by the solution method described below. Dennis, Gay, and Welsch (1981) give an adaptive algorithm that approximates the full Newton matrix $\mathbf{A}_{\theta\theta}$ by use of results from successive iterations.

The derivatives F_{u_r} are called the *first-order parametric sensitivities* of the model. Their direct computation via Newton's method is implemented in Subroutines DDAPLUS (Appendix B) and PDAPLUS. Finite-difference approximations are also provided as options in GREGPLUS to produce the matrix \mathbf{A} in either the Gauss-Newton or the full Newton form; these approximations are treated in Problems 6.B and 6.C.

The useful range of Eq. (6.3-7) can often be enhanced by writing the model in a different form. For example, the equation

$$\ln k_j = \ln A_j - E_j/RT \quad (6.3-11)$$

yields better-conditioned normal equations when rewritten in the form

$$\ln k_j = \ln k_{jB} + \frac{E_j}{R} \left[\frac{1}{T_B} - \frac{1}{T} \right] \quad (6.3-12)$$

used by Arrhenius (1889) and recommended by Box (1960). This parameterization is demonstrated in Example C.3, along with the robust form provided by Ratkowsky (1985) for heterogeneous reaction rate expressions.

¹ Since $\mathbf{A}_{\theta\theta}$ is real and symmetric, it belongs to the class of *normal* matrices [Stewart (1973), p. 288]. Thus, Eqs. (6.3-10) are properly called *normal equations*, whichever form of Eq. (6.3-9) is used.

6.4 CONSTRAINED MINIMIZATION OF $S(\theta)$

Models linear in θ , with unconstrained parameters, can be fitted directly by solving Eq. (6.3-5). Efficient algorithms and software for such problems are available [Lawson and Hanson (1974, 1995); Dongarra, Bunch, Moler, and Stewart (1979); Anderson et al., (1992)], and will not be elaborated here. We will focus on nonlinear models with bounded parameters, which are common in chemical kinetics and chemical reaction engineering.

Models nonlinear in θ need careful treatment. Direct iteration with Eq. (6.3-10) often fails because of the limited range of the expansions $\tilde{E}_u(\theta)$. Gauss-Newton iteration schemes with steps adjusted by line search work well [Booth, Box, Muller and Peterson (1958); Hartley (1961, 1964); Box and Kanemasu (1972, 1984); Bard (1974); Bock (1981)] when $A_{\theta\theta}$ is well-conditioned and θ unrestricted, but give difficulty otherwise.

Levenberg (1944) and Marquardt (1963) augmented $X^T X$ in Eq. (6.3-5) with a positive diagonal matrix λd , thus obtaining a nonsingular coefficient matrix for each iteration while shortening each correction step $\Delta\theta$. This approach has the difficulty that the true rank of the problem is concealed until near the end, when λ is reduced toward zero in an effort to recover Eq. (6.3-5). If $X^T X$ then turns out to be singular, the original problem is indeterminate (having too many parameters or inappropriate data) and the method gives no guidance for dealing with this.

Difficulties of this sort can be avoided by minimizing $\tilde{S}(\theta)$ of Eq. (6.3-7) over a trust region of θ in each iteration. A line search can be used to adjust the correction vector $\Delta\theta^k$ and the trust-region dimensions for the next iteration. Such a method is outlined below and used in GREGPLUS.

6.4.1 The Quadratic Programming Algorithm GRQP

A rectangular *trust region* of the form

$$|\theta_i - \text{PARB}(i)| \leq \begin{cases} |\text{CHMAX}(i)| & \text{if } \text{CHMAX}(i) \geq 0.0; \\ |\text{CHMAX}(i) * \text{PARB}(i)| & \text{if } \text{CHMAX}(i) < 0.0 \\ & \text{and } \text{PARB}(i) \neq 0.0; \\ \sqrt{0.1S(\theta^k)/A_{ii,\text{ref}}} & \text{if } \text{CHMAX}(i) < 0.0 \\ & \text{and } \text{PARB}(i) = 0.0 \\ & \text{and } A_{ii,\text{ref}} \neq 0.0; \\ 0.0 & \text{otherwise} \end{cases}$$

$$i = 1, \dots, \text{NPAR} \quad (6.4-1)$$

is used in Subroutine GRQP as the working range of the quadratic expansion $\tilde{S}(\theta)$ for the current iteration. Here PARB(i) and $A_{ii,\text{ref}}$ are the starting values of PAR(i) and A_{ii} for the iteration. The values CHMAX(i) are specified initially by the user and are adjustable by GREGPLUS at the end of each iteration.