CHAPTER 12

INTRODUCTION TO NONLINEAR REGRESSION

Linear regression models provide a rich and flexible framework that suits the needs of many analysts. However, linear regression models are not appropriate for all situations. There are many problems in engineering and the sciences where the response variable and the predictor variables are related through a known **nonlinear** function. This leads to a **nonlinear regression model**. When the method of least squares is applied to such models, the resulting normal equations are nonlinear and, in general, difficult to solve. The usual approach is to directly minimize the residual sum of squares by an iterative procedure. In this chapter we describe estimating the parameters in a nonlinear regression model and show how to make appropriate inferences on the model parameters. We also illustrate computer software for noulinear regression.

12.1 LINEAR AND NONLINEAR REGRESSION MODELS

12.1.1 Linear Regression Models

In previous chapters we have concentrated on the **linear regression model**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \tag{12.1}$$

These models include not only the first-order relationships, such as Eq. (12.1), but also polynomial models and other more complex relationships. In fact, we could write the linear regression model as

Introduction to Linear Regression Analysis, Fifth Edition. Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining.

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_r z_r + \varepsilon \tag{12.2}$$

where z_i represents any **function** of the original regressors $x_1, x_2, \ldots, x_k$, including transformations such as $\exp(x_i)$, $\sqrt{x_i}$, and $\sin(x_i)$. These models are called **linear** regression models because they are **linear in the unknown parameters**, the β_j, $j = 1, 2, \ldots, k$.

We may write the linear regression model (12.1) in a general form as

$$\begin{aligned} y &= \mathbf{x}'\boldsymbol{\beta} + \varepsilon \\ &= f(\mathbf{x}, \boldsymbol{\beta}) + \varepsilon \end{aligned} \tag{12.3}$$

where $\mathbf{x}' = [1, x_1, x_2, \ldots, x_k]$. Since the expected value of the model errors is zero, the expected value of the response variable is

$$\begin{aligned} E(y) &= E[f(\mathbf{x}, \boldsymbol{\beta}) + \varepsilon] \\ &= f(\mathbf{x}, \boldsymbol{\beta}) \end{aligned}$$

We usually refer to $f(\mathbf{x}, \boldsymbol{\beta})$ as the **expectation function** for the model. Obviously, the expectation function here is just a linear function of the unknown parameters.

12.1.2 Nonlinear Regression Models

There are many situations where a linear regression model may not be appropriate. For example, the engineer or scientist may have direct knowledge of the form of the relationship between the response variable and the regressors, perhaps from the theory underlying the phenomena. The true relationship between the response and the regressors may be a differential equation or the solution to a differential equation. Often, this will lead to a model of nonlinear form.

Any model that is not linear in the unknown parameters is a **nonlinear regression model**. For example, the model

$$y = \theta_1 e^{\theta_2 x} + \varepsilon \tag{12.4}$$

is not linear in the unknown parameters θ_1 and θ_2. We will use the symbol θ to represent a parameter in a nonlinear model to emphasize the difference between the linear and the nonlinear case.

In general, we will write the nonlinear regression model as

$$y = f(\mathbf{x}, \boldsymbol{\theta}) + \varepsilon \tag{12.5}$$

where $\boldsymbol{\theta}$ is a $p \times 1$ vector of unknown parameters and ε is an uncorrelated random-error term with $E(\varepsilon) = 0$ and $\text{Var}(\varepsilon) = \sigma^2$. We also typically assume that the errors are normally distributed, as in linear regression. Since

$$\begin{aligned} E(y) &= E[f(\mathbf{x}, \boldsymbol{\theta}) + \varepsilon] \\ &= f(\mathbf{x}, \boldsymbol{\theta}) \end{aligned} \tag{12.6}$$

we call $f(\mathbf{x}, \boldsymbol{\theta})$ the **expectation function** for the nonlinear regression model. This is very similar to the linear regression case, except that now the expectation function is a **nonlinear** function of the parameters.

In a nonlinear regression model, at least one of the derivatives of the expectation function with respect to the parameters depends on at least one of the parameters. In linear regression, these derivatives are **not** functions of the unknown parameters. To illustrate these points, consider a linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

with expectation function $f(\mathbf{x}, \boldsymbol{\beta}) = \beta_0 + \sum_{j=1}^{k} \beta_j x_j$ Now

$$\frac{\partial f(\mathbf{x}, \boldsymbol{\beta})}{\partial \beta_j} = x_j, \quad j = 0, 1, \ldots, k$$

where $x_0 \equiv 1$. Notice that in the linear case the derivatives are **not** functions of the β's

Now consider the nonlinear model

$$\begin{aligned} y &= f(x, \boldsymbol{\theta}) + \varepsilon \\ &= \theta_1 e^{\theta_2 x} + \varepsilon \end{aligned}$$

The derivatives of the expectation function with respect to θ_1 and θ_2 are

$$\frac{\partial f(x, \boldsymbol{\theta})}{\partial \theta_1} = e^{\theta_2 x} \quad \text{and} \quad \frac{\partial f(x, \boldsymbol{\theta})}{\partial \theta_2} = \theta_1 x e^{\theta_2 x}$$

Since the derivatives are a function of the unknown parameters θ_1 and θ_2, the model is nonlinear.

12.2 ORIGINS OF NONLINEAR MODELS

Nonlinear regression models often strike people as being very ad hoc because these models typically involve mathematical functions that are nonintuitive to people outside of the specific application area. Too often, people fail to appreciate the scientific theory underlying these nonlinear regression models. The scientific method uses mathematical models to describe physical phenomena. In many cases, the theory describing the physical relationships involves the solution of a set of differential equations, especially whenever rates of change are the basis for the mathematical model. This section outlines how the differential equations that form the heart of the theory describing physical behavior lead to nonlinear models. We discuss two examples. The first example deals with reaction rates and is more straightforward. The second example gives more details about the underlying theory to illustrate why nonlinear regression models have their specific forms. Our key point is that nonlinear regression models are almost always deeply rooted in the appropriate science.

Example 12.1

We first consider formally incorporating the effect of temperature into a second-order reaction kinetics model. For example, the hydrolysis of ethyl acetate is well modeled by a second-order kinetics model. Let A_t be the amount of ethyl acetate at time t. The second-order model is

$$\frac{dA_t}{dt} = -kA_t^2$$

where k is the rate constant. Rate constants depend on temperature, which we will incorporate into our model later. Let A_0 be the amount of ethyl acetate at time zero. The solution to the rate equation is

$$\frac{1}{A_t} = \frac{1}{A_0} + kt$$

With some algebra, we obtain

$$A_t = \frac{A_0}{1 + A_0 tk}$$

We next consider the impact of temperature on the rate constant. The Arrhenius equation states

$$k = C_1 \exp\left(-\frac{E_a}{RT}\right)$$

where E_a is the activation energy and C_1 is a constant. Substituting the Arrhenius equation into the rate equation yields

$$A_t = \frac{A_0}{1 + A_0 t C_1 \exp(-E_a / RT)}$$

Thus, an appropriate nonlinear regression model is

$$A_t = \frac{\theta_1}{1 + \theta_2 t \exp(-\theta_3 / T)} + \varepsilon_t \tag{12.7}$$

where $\theta_1 = A_0$, $\theta_2 = C_1 A_0$, and $\theta_3 = E_a/R$. ■

Example 12.2

We next consider the Clausius–Clapeyron equation, which is an important result in physical chemistry and chemical engineering. This equation describes the relationship of vapor pressure and temperature.

Vapor pressure is the physical property which explains why puddles of water evaporate away. Stable liquids at a given temperature are those that have achieved an equilibrium with their vapor phase. The vapor pressure is the partial pressure of the vapor phase at this equilibrium. If the vapor pressure equals the ambient pressure, then the liquid boils. Puddles evaporate when the partial pressure of the water vapor in the ambient atmosphere is less than the vapor pressure of water at that temperature. The nonequilibrium condition presented by this difference between the actual partial pressure and the vapor pressure causes the puddle's water to evaporate over time.

The chemical theory that describes the behavior at the vapor–liquid interface notes that at equilibrium the Gibbs free energies of both the vapor and liquid phases must be equal. The Gibbs free energy G is given by

$$G = U + PV - TS = H - TS$$

where U is the "internal energy," P is the pressure, V is the volume, T is the "absolute" temperature, S is the entropy, and $H = U + PV$ is the enthalpy. Typically, in thermodynamics, we are more interested in the change in Gibbs free energy than its absolute value. As a result, the actual value of U is often of limited interest. The derivation of the Clausius–Clapeyron equation also makes use of the ideal gas law,

$$PV = RT$$

where R is the ideal gas constant.

Consider the impact of a slight change in the temperature when holding the volume fixed. From the ideal gas law, we observe that an increase in the temperature necessitates an increase in the pressure. Let dG be the resulting differential in the Gibbs free energy. We note that

$$\begin{aligned} dG &= \left(\frac{\partial G}{\partial P}\right)_T dP + \left(\frac{\partial G}{\partial T}\right)_P dT \\ &= V dP - S dT \end{aligned}$$

Let the subscript 1 denote the liquid phase and the subscript v denote the vapor phase. Thus, G_1 and G_v are the Gibbs free energies of the liquid and vapor phases, respectively. If we maintain the vapor–liquid equilibrium as we change the temperature and pressure, then

$$\begin{aligned} dG_1 &= dG_v \\ V_1 dP - S_1 dT &= V_v dP - S_v dT \end{aligned}$$

Rearranging, we obtain

$$\frac{dP}{dT} = \frac{S_v - S_1}{V_v - V_1} \tag{12.8}$$

We observe that the volume occupied by the vapor is much larger than the volume occupied by the liquid. Effectively, the difference is so large that we can treat V_1 as zero. Next, we observe that entropy is defined by

$$dS = \frac{dQ}{T}$$

where Q is the heat exchanged reversibly between the system and its surroundings. For our vapor–liquid equilibrium situation, the net heat exchanged is H_{vap}, which is the heat of vaporization at temperature T. Thus,

$$S_{\text{v}} - S_1 = \frac{H_{\text{vap}}}{T}$$

We then can rewrite (12.8) as

$$\frac{dP}{dT} = \frac{H_{\text{vap}}}{VT}$$

From the ideal gas law,

$$V = \frac{RT}{P}$$

We then may rewrite (12.8) as

$$\frac{dP}{dT} = \frac{PH_{\text{vap}}}{RT^2}$$

Rearranging, we obtain,

$$\frac{dP}{P} = \frac{H_{\text{vap}}dT}{RT^2}$$

Integrating, we obtain

$$\ln(P) = \mathcal{C} - \mathcal{C}_1\frac{1}{T} \tag{12.9}$$

where $\mathcal{C}$ is an integration constant and

$$\mathcal{C}_1 = \frac{H_{\text{vap}}}{R}$$

We can reexpress (12.9) as

$$P = \mathcal{C}_0 + \mathcal{C}\exp\left(-\frac{\mathcal{C}_1}{T}\right) \tag{12.10}$$

where C_0 is another integration constant. Equation (12.9) suggests a simple linear regression model of the form

$$\ln(P)_i = \beta_0 + \beta_1 \frac{1}{T_i} + \varepsilon_i \tag{12.11}$$

Equation (12.10) on the other hand, suggests a nonlinear regression model of the form

$$P_i = \theta_1 \exp\left(\frac{\theta_2}{T_i}\right) + \varepsilon_i \tag{12.12}$$

It is important to note that there are subtle, yet profound differences between these two possible models. We discuss some of the possible differences between linear and nonlinear models in Section 12.4. ■

12.3 NONLINEAR LEAST SQUARES

Suppose that we have a sample of n observations on the response and the regressors, say $y_i, x_{i1}, x_{i2}, \ldots, x_{ik}$, for $i = 1, 2, \ldots, n$. We have observed previously that the method of least squares in linear regression involves minimizing the least-squares function

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[y_i - \left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \right) \right]^2$$

Because this is a linear regression model, when we differentiate $S(\boldsymbol{\beta})$ with respect to the unknown parameters and equate the derivatives to zero, the resulting normal equations are **linear** equations, and consequently, they are easy to solve.

Now consider the nonlinear regression situation. The model is

$$y_i = f(\mathbf{x}_i, \boldsymbol{\theta}) + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

where now $\mathbf{x}_i' = [1, x_{i1}, x_{i2}, \ldots, x_{ik}]$ for $i = 1, 2, \ldots, n$. The least-squares function is

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n} [y_i - f(\mathbf{x}_i, \boldsymbol{\theta})]^2 \tag{12.13}$$

To find the least-squares estimates we must differentiate Eq. (12.13) with respect to each element of $\boldsymbol{\theta}$. This will provide a set of p normal equations for the nonlinear regression situation. The normal equations are

$$\sum_{i=1}^{n} [y_i - f(\mathbf{x}_i, \boldsymbol{\theta})] \left[\frac{\partial f(\mathbf{x}_i, \boldsymbol{\theta})}{\partial \theta_j} \right]_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} = 0 \quad \text{for } j = 1, 2, \ldots, p \tag{12.14}$$

In a nonlinear regression model the derivatives in the large square brackets will be functions of the unknown parameters. Furthermore, the expectation function is also a nonlinear function, so the normal equations can be very difficult to solve.

Example 12.3 Normal Equations for a Nonlinear Model

Consider the nonlinear regression model in Eq. (12.4):

$$y = \theta_1 e^{\theta_2 x} + \varepsilon$$

The least-squares normal equations for this model are

$$\sum_{i=1}^{n} \left[y_i - \hat{\theta}_1 e^{\hat{\theta}_2 x_i} \right] e^{\hat{\theta}_2 x_i} = 0$$

$$\sum_{i=1}^{n} \left[y_i - \hat{\theta}_1 e^{\hat{\theta}_2 x_i} \right] \hat{\theta}_1 x_i e^{\hat{\theta}_2 x_i} = 0 \tag{12.15}$$

After simplification, the normal equations are

$$\sum_{i=1}^{n} y_i e^{\hat{\theta}_2 x_i} - \hat{\theta}_1 \sum_{i=1}^{n} e^{2\hat{\theta}_2 x_i} = 0$$

$$\sum_{i=1}^{n} y_i x_i e^{\hat{\theta}_2 x_i} - \hat{\theta}_1 \sum_{i=1}^{n} x_i e^{2\hat{\theta}_2 x_i} = 0 \tag{12.16}$$

These equations are not linear in $\hat{\theta}_1$ and $\hat{\theta}_2$, and no simple closed-form solution exists. In general, **iterative methods** must be used to find the values of $\hat{\theta}_1$ and $\hat{\theta}_2$. To further complicate the problem, sometimes there are multiple solutions to the normal equations. That is, there are multiple stationary values for the residual sum of squares function $S(\boldsymbol{\theta})$. ■

Geometry of Linear and Nonlinear Least Squares Examining the geometry of the least-squares problem is helpful in understanding the complexities introduced by a nonlinear model. For a given sample, the residual-sum-of-squares function $S(\boldsymbol{\theta})$ depends only on the model parameters $\boldsymbol{\theta}$. Thus, in the parameter space (the space defined by the $\theta_1, \theta_2, \ldots, \theta_p$), we can represent the function $S(\boldsymbol{\theta})$ with a contour plot, where each contour on the surface is a line of constant residual sum of squares.

Suppose the regression model is linear; that is, the parameters are $\boldsymbol{\theta} = \boldsymbol{\beta}$, and the residual-sum-of-squares function is $\boldsymbol{S}(\boldsymbol{\beta})$ Figure 12.1*a* shows the contour plot for this situation. If the model is linear in the unknown parameters, the contours are ellipsoidal and have a unique global minimum at the least-squares estimator $\hat{\boldsymbol{\beta}}$.

When the model is nonlinear, the contours will often appear as in Figure 12.1*b*. Notice that these contours are not elliptical and are in fact quite elongated and irregular in shape. A "banana-shape" appearance is very typical. The specific shape and orientation of the residual sum of squares contours depend on the form of the nonlinear model and the sample of data that have been obtained. Often the surface will be very elongated near the optimum, so many solutions for $\boldsymbol{\theta}$ will produce a residual sum of squares that is close to the global minimum. This results in a problem that is **ill-conditioned**, and in such problems it is often difficult to find the global minimum for $\boldsymbol{\theta}$. In some situations, the contours may be so irregular that there are several local minima and perhaps more than one global minimum. Figure 12.1*c* shows a situation where there is one local minimum and a global minimum.

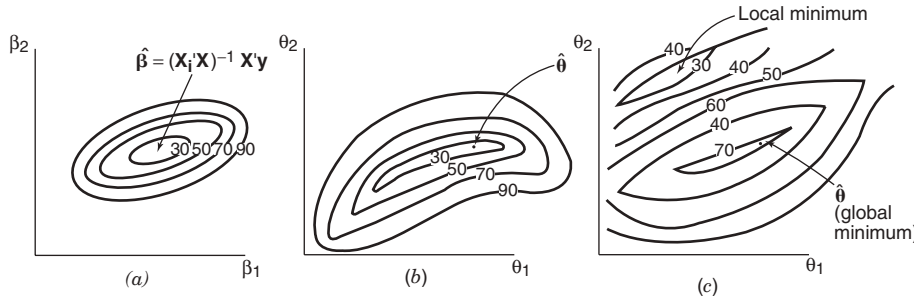

Figure 12.1 Contours of the residual-sum-of-squares function: (*a*) linear model; (*b*) nonlinear model; (*c*) nonlinear model with local and global minima.

Maximum-Likelihood Estimation We have concentrated on least squares in the nonlinear case. If the error terms in the model are normally and independently distributed with constant variance, application of the method of maximum likelihood to the estimation problem will lead to least squares. For example, consider the model in Eq. (12.4):

$$y_i = \theta_1 e^{\theta_2 x_i} + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{12.17}$$

If the errors are normally and independently distributed with mean zero and variance σ^2, then the likelihood function is

$$L(\theta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[y_i - \theta_1 e^{\theta_2 x_i} \right]^2 \right] \tag{12.18}$$

Clearly, maximizing this likelihood function is equivalent to minimizing the residual sum of squares. Therefore, in the normal-theory case, least-squares estimates are the same as maximum-likelihood estimates.

12.4 TRANFORMATION TO A LINEAR MODEL

It is sometimes useful to consider a **transformation** that induces linearity in the model expectation function. For example, consider the model

$$\begin{aligned} y &= f(x, \boldsymbol{\theta}) + \varepsilon \\ &= \theta_1 e^{\theta_2 x} + \varepsilon \end{aligned} \tag{12.19}$$

The Clausius–Clapeyron equation (12.12) is an example of this model. Now since $E(y) = f(x, \boldsymbol{\theta}) = \theta_1 e^{\theta_2 x}$, we can linearize the expectation function by taking logarithms,

$$\ln E(y) = \ln \theta_1 + \theta_2 x$$

which we saw in Eq. (12.11) in our derivation of the Clausius–Clapeyron equation. Therefore, it is tempting to consider rewriting the model as

$$\begin{aligned} \ln y &= \ln \theta_1 + \theta_2 x + \varepsilon \\ &= \beta_0 + \beta_1 x + \varepsilon \end{aligned} \tag{12.20}$$

and using simple **linear** regression to estimate β_0 and β_1. However, the linear least-squares estimates of the parameters in Eq. (12.20) will not in general be equivalent to the nonlinear parameter estimates in the original model (12.19). The reason is that in the **original nonlinear model** least squares implies minimization of the sum of squared residuals on y, whereas in the **transformed model** (12.20) we are minimizing the sum of squared residuals on $\ln y$.

Note that in Eq. (12.19) the error structure is **additive**, so taking logarithms **cannot** produce the model in Eq. (12.20). If the error structure is **multiplicative**, say

$$y = \theta_1 e^{\theta_2 x} \varepsilon \tag{12.21}$$

then taking logarithms will be appropriate, since

$$\begin{aligned} \ln y &= \ln \theta_1 + \theta_2 x + \ln \varepsilon \\ &= \beta_0 + \beta_1 x + \varepsilon^* \end{aligned} \tag{12.22}$$

and if ε* follows a normal distribution, all the standard linear regression model properties and associated inference will apply.

A nonlinear model that can be transformed to an equivalent linear form is said to be **intrinsically linear**. However, the issue often revolves around the error structure, namely, do the standard assumptions on the errors apply to the original nonlinear model or to the linearized one? This is sometimes not an easy question to answer.

Example 12.4 The Puromycin Data

Bates and Watts [1988] use the **Michaelis–Menten** model for chemical kinetics to relate the initial velocity of an enzymatic reaction to the substrate concentration x. The model is

$$y = \frac{\theta_1 x}{x + \theta_2} + \varepsilon \tag{12.23}$$

The data for the initial rate of a reaction for an enzyme treated with puromycin are shown in Table 12.1 and plotted in Figure 12.2.

We note that the expectation function can be linearized easily, since

$$\begin{aligned} \frac{1}{f(x, \boldsymbol{\theta})} &= \frac{x + \theta_2}{\theta_1 x} = \frac{1}{\theta_1} + \frac{\theta_2}{\theta_1} \frac{1}{x} \\ &= \beta_0 + \beta_1 x \end{aligned}$$

so we are tempted to fit the **linear** model

$$y^* = \beta_0 + \beta_1 u + \varepsilon$$

TABLE 12.1 Reaction Velocity and Substrate Concentration for Puromycin Experiment

Substrate Concentration (ppm)	Velocity [(counts/min)/min]	
0.02	47	76
0.06	97	107
0.11	123	139
0.22	152	159
0.56	191	201
1.10	200	207

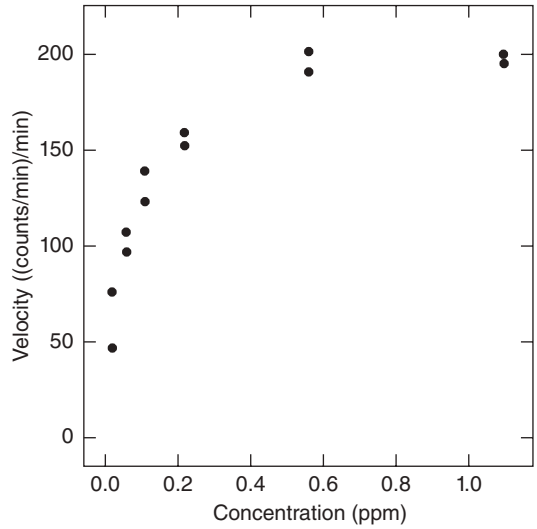

Figure 12.2 Plot of reaction velocity versus substrate concentration for the puromycin experiment. (Adapted from Bates and Watts [1988], with permission of the publisher.)

where $y^* = 1/y$ and $u = 1/x$. The resulting least-squares fit is

$$\hat{y}^* = 0.005107 + 0.0002472u$$

Figure 12.3*a* shows a scatterplot of the transformed data y^* and u with the straight-line fit superimposed. As there are replicates in the data, it is easy to see from Figure 12.2 that the variance of the original data is approximately constant, while Figure 12.3*a* indicates that in the transformed scale the constant-variance assumption is unreasonable.

Now since

$$\beta_0 = \frac{1}{\theta_1} \quad \text{and} \quad \beta_1 = \frac{\theta_2}{\theta_1}$$

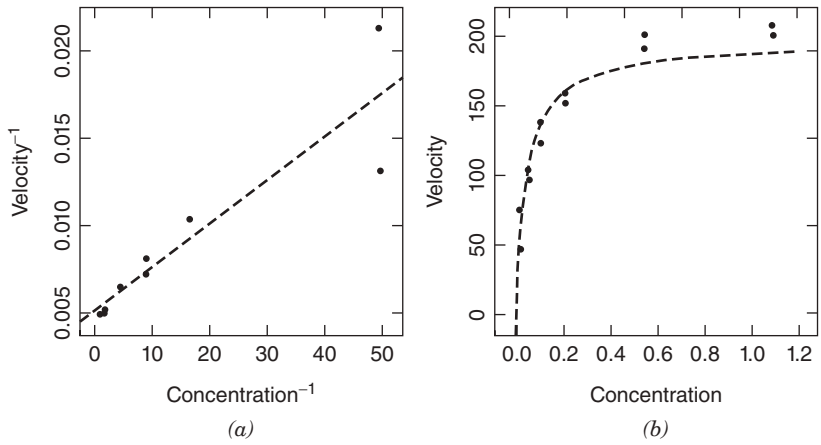

Figure 12.3 (*a*) Plot of inverse velocity versus inverse concentration for the puromycin data. (*b*) Fitted curve in the original scale. (Adapted from Bates and Watts [1988], with permission of the publisher.)

we have

$$0.005107 = \frac{1}{\hat{\theta}_1} \quad \text{and} \quad 0.0002472 = \frac{\hat{\theta}_2}{\hat{\theta}_1}$$

and so we can estimate θ_1 and θ_2 in the original model as

$$\hat{\theta}_1 = 195.81 \quad \text{and} \quad \hat{\theta}_2 = 0.04841$$

Figure 12.3*b* shows the fitted curve in the original scale along with the data. Observe from the figure that the fitted asymptote is too small. The variance at the replicated points has been distorted by the transformation, so runs with low concentration (high reciprocal concentration) dominate the least-squares fit, and as a result the model does not fit the data well at high concentrations. ■

12.5 PARAMETER ESTIMATION IN A NONLINEAR SYSTEM

12.5.1 Linearization

A method widely used in computer algorithms for nonlinear regression is **linearization** of the nonlinear function followed by the Gauss–Newton iteration method of parameter estimation. Linearization is accomplished by a **Taylor series expansion** of $f(\mathbf{x}_i, \boldsymbol{\theta})$ about the point $\boldsymbol{\theta}_0' = [\theta_{10}, \theta_{20}, \ldots, \theta_{p0}]$ with only the linear terms retained. This yields

$$f(\mathbf{x}_i, \boldsymbol{\theta}) = f(\mathbf{x}_i, \boldsymbol{\theta}_0) + \sum_{j=1}^{p} \left[\frac{\partial f(\mathbf{x}_i, \boldsymbol{\theta})}{\partial \theta_j} \right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\theta_j - \theta_{j0}) \tag{12.24}$$

If we set

$$f_i^0 = f(\mathbf{x}_i, \boldsymbol{\theta}_0)$$
$$\beta_j^0 = \theta_j - \theta_{j0}$$
$$Z_{ij}^0 = \left[\frac{\partial f(\mathbf{x}_i, \boldsymbol{\theta}_0)}{\partial \theta_j}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$$

we note that the nonlinear regression model can be written as

$$y_i - f_i^0 = \sum_{j=1}^{p} \beta_j^0 Z_{ij}^0 + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{12.25}$$

That is, we now have a linear regression model. We usually call $\boldsymbol{\theta}_0$ the starting values for the parameters.

We may write Eq. (12.25) as

$$\mathbf{y}_0 = \mathbf{Z}_0 \boldsymbol{\beta}_0 + \boldsymbol{\varepsilon} \tag{12.26}$$

so the estimate of $\boldsymbol{\beta}_0$ is

$$\hat{\boldsymbol{\beta}}_0 = (\mathbf{Z}_0' \mathbf{Z}_0)^{-1} \mathbf{Z}_0' \mathbf{y}_0 = (\mathbf{Z}_0' \mathbf{Z}_0)^{-1} \mathbf{Z}_0' (\mathbf{y} - \mathbf{f}_0) \tag{12.27}$$

Now since $\boldsymbol{\beta}_0 = \boldsymbol{\theta} - \boldsymbol{\theta}_0$, we could define

$$\hat{\boldsymbol{\theta}}_1 = \hat{\boldsymbol{\beta}}_0 + \boldsymbol{\theta}_0 \tag{12.28}$$

as revised estimates of $\boldsymbol{\theta}$. Sometimes $\hat{\boldsymbol{\beta}}_0$ is called the **vector of increments**. We may now place the revised estimates $\hat{\boldsymbol{\theta}}_1$ in Eq. (12.24) (in the same roles played by the initial estimates $\boldsymbol{\theta}_0$) and then produce another set of revised estimates, say $\hat{\boldsymbol{\theta}}_2$, and so forth.

In general, we have at the kth iteration

$$\hat{\boldsymbol{\theta}}_{k+1} = \hat{\boldsymbol{\theta}}_k + \hat{\boldsymbol{\beta}}_k = \hat{\boldsymbol{\theta}}_k + (\mathbf{Z}_k' \mathbf{Z}_k)^{-1} \mathbf{Z}_k' (\mathbf{y} - \mathbf{f}_k) \tag{12.29}$$

where

$$\mathbf{Z}_k = [Z_{ij}^k]$$
$$\mathbf{f}_k = [f_1^k, f_2^k, \ldots, f_n^k]'$$
$$\hat{\boldsymbol{\theta}}_k = [\theta_{1k}, \theta_{2k}, \ldots, \theta_{pk}]'$$

This iterative process continues until convergence, that is, until

$$\left[\left(\hat{\theta}_{j,k+1} - \hat{\theta}_{jk}\right) / \hat{\theta}_{jk}\right] < \delta, \quad j = 1, 2, \ldots, p$$

where δ is some small number, say 1.0×10^{-6}. At each iteration the residual sum of squares $S(\hat{\boldsymbol{\theta}}_k)$ should be evaluated to ensure that a reduction in its value has been obtained.

Example 12.5 The Puromycin Data

Bates and Watts [1988] use the Gauss–Newton method to fit the Michaelis–Menten model to the puromycin data in Table 12.1 using the starting values $\theta_{10} = 205$ and $\theta_{20} = 0.08$. Later we will discuss how these starting values were obtained. At this starting point, the residual sum of squares $S(\boldsymbol{\theta}_0) = 3155$. The data, fitted values, residuals, and derivatives evaluated at each observation are shown in Table 12.2. To illustrate how the required quantities are calculated, note that

$$\frac{\partial f(x, \theta_1, \theta_2)}{\partial \theta_1} = \frac{x}{\theta_2 + x} \quad \text{and} \quad \frac{\partial f(x, \theta_1, \theta_2)}{\partial \theta_2} = \frac{-\theta_1 x}{(\theta_2 + x)^2}$$

and since the first observation on x is $x_1 = 0.02$, we have

$$Z_{11}^0 = \left.\frac{x_1}{\theta_2 + x}\right|_{\theta_2 = 0.08} = \frac{0.02}{0.08 + 0.02} = 0.2000$$

$$Z_{12}^0 = \left.\frac{-\theta_1 x_1}{(\theta_2 + x_1)^2}\right|_{\theta_1 = 205,\, \theta_2 = 0.08} = \frac{(-205)(0.02)}{(0.08 + 0.02)^2} = -410.00$$

The derivatives Z_{ij}^0 are now collected into the matrix $\mathbf{Z}_0$ and the vector of increments calculated from Eq. (12.27) as

$$\hat{\boldsymbol{\beta}}_0 = \begin{bmatrix} 8.03 \\ -0.017 \end{bmatrix}$$

TABLE 12.2 Data, Fitted Values, Residuals, and Derivatives for the Puromycin Data at $\hat{\boldsymbol{\theta}}_0' = [205, 0.08]$

i	x_i	y_i	f_i^0	$y_i - f_i^0$	Z_{i1}^0	Z_{i2}^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

The revised estimate $\hat{\boldsymbol{\theta}}_1$ from Eq. (12.28) is

$$\begin{aligned}\hat{\boldsymbol{\theta}}_1 &= \hat{\boldsymbol{\beta}}_0 + \boldsymbol{\theta}_0 \\ &= \begin{bmatrix} 8.03 \\ -0.017 \end{bmatrix} + \begin{bmatrix} 205.00 \\ 0.08 \end{bmatrix} = \begin{bmatrix} 213.03 \\ 0.063 \end{bmatrix}\end{aligned}$$

The residual sum of squares at this point is $S\left(\hat{\boldsymbol{\theta}}_1\right) = 1206$, which is considerably smaller than $S(\boldsymbol{\theta}_0)$. Therefore, $\hat{\boldsymbol{\theta}}_1$ is adopted as the revised estimate of $\boldsymbol{\theta}$, and another iteration would be performed.

The Gauss–Newton algorithm converged at $\hat{\boldsymbol{\theta}}' = [212.7, 0.0641]'$ with $S\left(\hat{\boldsymbol{\theta}}\right) = 1195$. Therefore, the fitted model obtained by linearization is

$$\hat{y} = \frac{\hat{\boldsymbol{\theta}}_1 x}{x + \hat{\boldsymbol{\theta}}_2} = \frac{212.7x}{x + 0.0641}$$

Figure 12.4 shows the fitted model. Notice that the nonlinear model provides a much better fit to the data than did the transformation followed by linear regression in Example 12.4 (compare Figures 12.4 and 12.3*b*).

Residuals can be obtained from a fitted nonlinear regression model in the usual way, that is,

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \ldots, n$$

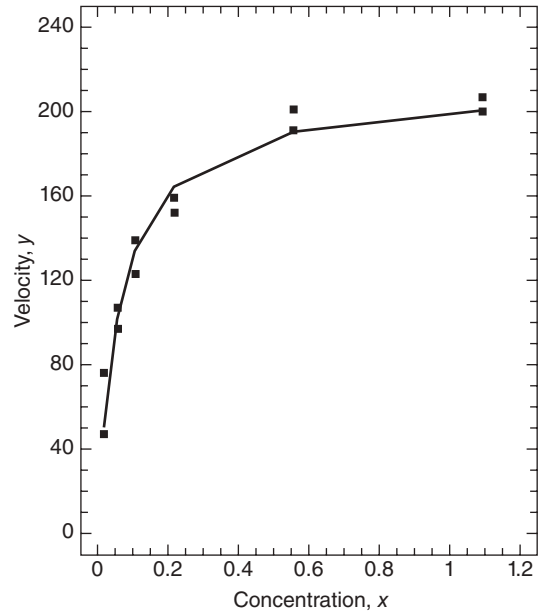

Figure 12.4 Plot of fitted nonlinear regression model, Example 12.5.

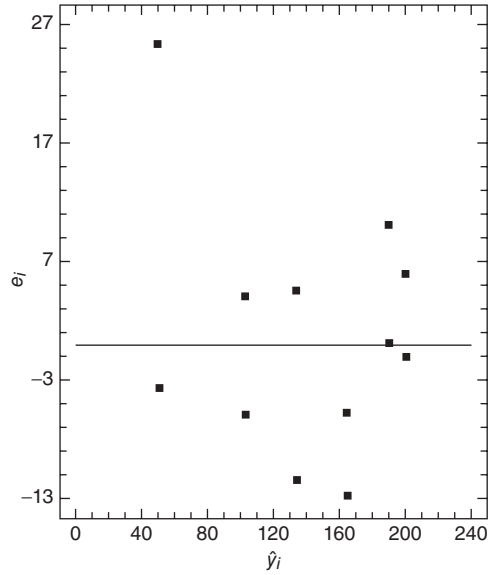

Figure 12.5 Plot of residuals versus predicted values, Example 12.5.

In this example the residuals are computed from

$$e_i = y_i - \frac{\hat{\theta}_1 x_i}{x_i + \hat{\theta}_2} = y_i - \frac{212.7x}{x_i + 0.0641}, \quad i = 1, 2, \ldots, 10$$

The residuals are plotted versus the predicted values in Figure 12.5. A normal probability plot of the residuals is shown in Figure 12.6. There is one moderately large residual; however, the overall fit is satisfactory, and the model seems to be a substantial improvement over that obtained by the transformation approach in Example 12.4. ■

Computer Programs Several PC statistics packages have the capability to fit nonlinear regression models. Both JMP and Minitab (version 16 and higher) have this capability. Table 12.3 is the output from JMP that results from fitting the Michaelis–Menten model to the puromycin data in Table 12.1. JMP required 13 iterations to converge to the final parameter estimates. The output provides the estimates of the model parameters, approximate standard errors of the parameter estimates, the error or residual sum of squares, and the correlation matrix of the parameter estimates. We make use of some of these quantities in later sections.

Estimation of σ^2 When the estimation procedure converges to a final vector of parameter estimates $\hat{\boldsymbol{\theta}}$, we can obtain an estimate of the error variance σ^2 from the residual mean square

$$\hat{\sigma}^2 = MS_{\text{Res}} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - p} = \frac{\sum_{i=1}^{n} \left[y_i - f\left(\mathbf{x}_i, \hat{\boldsymbol{\theta}}\right) \right]^2}{n - p} = \frac{S\left(\hat{\boldsymbol{\theta}}\right)}{n - p} \tag{12..30}$$

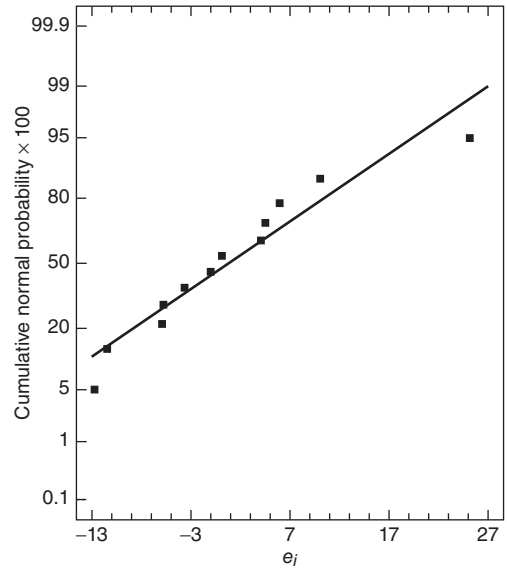

Figure 12.6 Normal probability plot of residuals, Example 12.5.

TABLE 12.3 JMP Output for Fitting the Michaelis–Menten Model to the Puromycin Data

Nonlinear Fit

Response: Velocity, Predictor: Michaelis Menten Model (2P)

Criterion	Current	Stop Limit
Iteration	13	60
Obj Change	2.001932e-12	1e-15
Relative Gradient	3.5267226e-7	0.000001
Gradient	0.0001344207	0.000001

Parameter	Current Value
theta1	212.68374295
theta2	0.0641212814
SSE	1195.4488144
N	12

Solution

SSE	DFE	MSE	RMSE
1195.4488144	10	119.54488	10.933658

Parameter	Estimate	ApproxStdErr
theta1	212.68374295	6.94715515
theta2	0.0641212814	0.00828095

Solved By: Analytic NR

Correlation of Estimates

	theta1	theta2
theta1	1.0000	0.7651
theta2	0.7651	1.0000

where p is the number of parameters in the nonlinear regression model. For the puromycin data in Example 12.5, we found that the residual sum of squares at the final iteration was $S(\hat{\boldsymbol{\theta}}) = 1195$ (also see the JMP output in Table 12.3), so the estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{S(\hat{\boldsymbol{\theta}})}{n-p} = \frac{1195}{12-2} = 119.5$$

We may also estimate the **asymptotic (large-sample) covariance matrix** of the parameter vector $\hat{\boldsymbol{\theta}}$ by

$$\text{Var}(\hat{\boldsymbol{\theta}}) = \sigma^2 (\mathbf{Z}'\mathbf{Z})^{-1} \tag{12.31}$$

where $\mathbf{Z}$ is the matrix of partial derivatives defined previously, evaluated at the final-iteration least-squares estimate $\hat{\boldsymbol{\theta}}$.

The covariance matrix of the $\hat{\boldsymbol{\theta}}$ vector for the Michaelis–Menten model in Example 12.5 is

$$\text{Var}(\hat{\boldsymbol{\theta}}) = \hat{\sigma}^2 (\mathbf{Z}'\mathbf{Z})^{-1} = 119.5 \begin{bmatrix} 0.4037 & 36.82\times10^{-5} \\ 36.82\times10^{-5} & 57.36\times10^{-8} \end{bmatrix}$$

The main diagonal elements of this matrix are approximate variances of the estimates of the regression coefficients. Therefore, approximate **standard errors** on the coefficients are

$$\text{se}(\hat{\theta}_1) = \sqrt{\text{Var}(\hat{\theta}_1)} = \sqrt{119.5(0.4037)} = 6.95$$

and

$$\text{se}(\hat{\theta}_2) = \sqrt{\text{Var}(\hat{\theta}_2)} = \sqrt{119.5(57.36\times10^{-8})} = 8.28\times10^{-3}$$

and the correlation between $\hat{\theta}_1$ and $\hat{\theta}_2$ is about

$$\frac{36.82\times10^{-5}}{\sqrt{0.4037(57.36\times10^{-8})}} = 0.77$$

These values agree closely with those reported in the JMP output, Table 12.3.

Graphical Perspective on Linearization We have observed that the residual-sum-of-squares function $S(\theta)$ for a nonlinear regression model is usually an irregular "banana-shaped" function, as shown in panels *b* and *c* of Figure 12.1. On the other hand, the residual-sum-of-squares function for linear least squares is very well behaved; in fact, it is elliptical and has the global minimum at the bottom of the "bowl." Refer to Figure 12.1*a*. The linearization technique converts the nonlinear regression problem into a sequence of linear ones, starting at the point $\boldsymbol{\theta}_0$.

The first iteration of linearization replaces the irregular contours with a set of elliptical contours. The irregular contours of $S(\boldsymbol{\theta})$ pass exactly through the starting

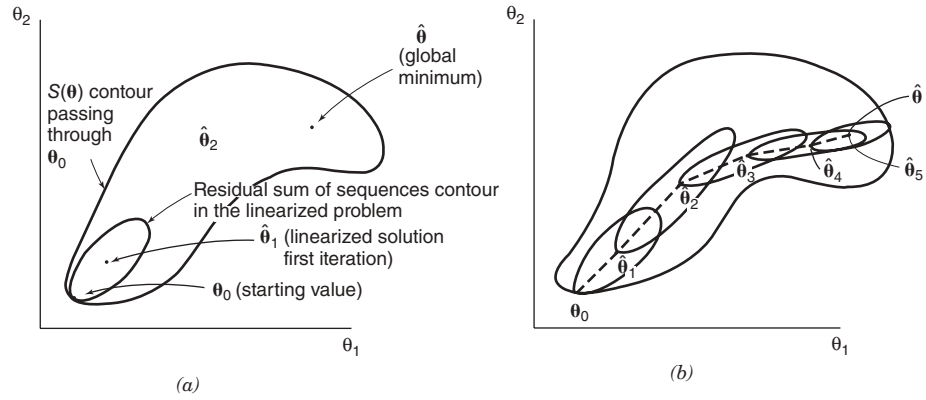

Figure 12.7 A geometric view of linearization: (*a*) the first iteration; (*b*) evolution of successive linearization iterations.

point $\boldsymbol{\theta}_0$, as shown in Figure 12.7*a*. When we solve the linearized problem, we are moving to the global minimum on the set of elliptical contours. This is done by ordinary linear least squares. Then the next iteration just repeats the process, starting at the new solution $\hat{\boldsymbol{\theta}}_1$. The eventual evolution of linearization is a sequence of linear problems for which the solutions "close in" on the global minimum of the nonlinear function. This is illustrated in Figure 12.7*b*. Provided that the nonlinear problem is not too ill-conditioned, either because of a poorly specified model or inadequate data, the linearization procedure should converge to a good estimate of the global minimum in a few iterations.

Linearization is facilitated by a good starting value $\boldsymbol{\theta}_0$, that is, one that is reasonably close to the global minimum. When $\boldsymbol{\theta}_0$ is close to $\hat{\boldsymbol{\theta}}$, the actual residual-sum-of-squares contours of the nonlinear problem are usually well-approximated by the contours of the linearized problem. We will discuss obtaining starting values in Section 12.5.3.

12.5.2 Other Parameter Estimation Methods

The basic linearization method described in Section 12.5.1 may converge very slowly in some problems. In other problems, it may generate a move in the wrong direction, with the residual-sum-of-squares function $S\left(\hat{\boldsymbol{\theta}}_k\right)$ actually **increasing** at the kth iteration. In extreme cases, it may fail to converge at all. Consequently, several other techniques for solving the nonlinear regression problem have been developed. Some of them are modifications and refinements of the linearization scheme. In this section we give a brief description of some of these procedures.

Method of Steepest Descent The method of steepest descent attempts to find the global minimum on the residual-sum-of-squares function by direct minimization. The objective is to move from an initial starting point $\boldsymbol{\theta}_0$ in a vector direction with components given by the derivatives of the residual-sum-of-squares function with respect to the elements of $\boldsymbol{\theta}$. Usually these derivatives are estimated by fitting a first-order or planar approximation around the point $\boldsymbol{\theta}_0$. The regression coefficients in the first-order model are taken as approximations to the first derivatives.