Part III

Nonlinear Regression

Chapter 13

Introduction to Nonlinear Regression and Neural Networks

The linear regression models considered up to this point are generally satisfactory approximations for most regression applications. There are occasions, however, when an empirically indicated or a theoretically justified nonlinear regression model is more appropriate. For example, growth from birth to maturity in human subjects typically is nonlinear in nature, characterized by rapid growth shortly after birth, pronounced growth during puberty, and a leveling off sometime before adulthood. In another example, dose-response relationships tend to be nonlinear with little or no change in response for low dose levels of a drug, followed by rapid S-shaped changes occurring in the more active dose region, and finally with dose response leveling off as it reaches a saturated level. We shall consider in this chapter and the next some nonlinear regression models, how to obtain estimates of the regression parameters in such models, and how to make inferences about these regression parameters.

In this chapter, we introduce exponential nonlinear regression models and present the basic methods of nonlinear regression. We also introduce neural network models, which are now widely used in data mining applications. In Chapter 14, we present logistic regression models and consider their uses when the response variable is binary or categorical with more than two levels.

13.1 Linear and Nonlinear Regression Models

Linear Regression Models

In previous chapters, we considered linear regression models, i.e., models that are linear in the parameters. Such models can be represented by the general linear regression model (6.7):

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \tag{13.1}$$

Linear regression models, as we have seen, include not only first-order models in $p-1$ predictor variables but also more complex models. For instance, a polynomial regression model in one or more predictor variables is linear in the parameters, such as the following model in two predictor variables with linear, quadratic, and interaction terms:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \beta_3 X_{i2} + \beta_4 X_{i2}^2 + \beta_5 X_{i1} X_{i2} + \varepsilon_i \tag{13.2}$$

Also, models with transformed variables that are linear in the parameters belong to the class of linear regression models, such as the following model:

$$\log_{10} Y_i = \beta_0 + \beta_1 \sqrt{X_{i1}} + \beta_2 \exp(X_{i2}) + \varepsilon_i \tag{13.3}$$

In general, we can state a linear regression model in the form:

$$Y_i = f(\mathbf{X}_i, \boldsymbol{\beta}) + \varepsilon_i \tag{13.4}$$

where $\mathbf{X}_i$ is the vector of the observations on the predictor variables for the ith case:

$$\mathbf{X}_i = \begin{bmatrix} 1 \\ X_{i1} \\ \vdots \\ X_{i,p-1} \end{bmatrix} \tag{13.4a}$$

$\boldsymbol{\beta}$ is the vector of the regression coefficients in (6.18c), and $f(\mathbf{X}_i, \boldsymbol{\beta})$ represents the expected value $E\{Y_i\}$, which for linear regression models equals according to (6.54):

$$f(\mathbf{X}_i, \boldsymbol{\beta}) = \mathbf{X}_i'\boldsymbol{\beta} \tag{13.4b}$$

Nonlinear Regression Models

Nonlinear regression models are of the same basic form as that in (13.4) for linear regression models:

$$Y_i = f(\mathbf{X}_i, \boldsymbol{\gamma}) + \varepsilon_i \tag{13.5}$$

An observation Y_i is still the sum of a mean response $f(\mathbf{X}_i, \boldsymbol{\gamma})$ given by the nonlinear response function $f(\mathbf{X}, \boldsymbol{\gamma})$ and the error term ε_i. The error terms usually are assumed to have expectation zero, constant variance, and to be uncorrelated, just as for linear regression models. Often, a normal error model is utilized which assumes that the error terms are independent normal random variables with constant variance.

The parameter vector in the response function $f(\mathbf{X}, \boldsymbol{\gamma})$ is now denoted by $\boldsymbol{\gamma}$ rather than $\boldsymbol{\beta}$ as a reminder that the response function here is nonlinear in the parameters. We present now two examples of nonlinear regression models that are widely used in practice.

Exponential Regression Models. One widely used nonlinear regression model is the exponential regression model. When there is only a single predictor variable, one form of this regression model with normal error terms is:

$$Y_i = \gamma_0 \exp(\gamma_1 X_i) + \varepsilon_i \tag{13.6}$$

where:

γ_0 and γ_1 are parameters
X_i are known constants
ε_i are independent $N(0, \sigma^2)$

The response function for this model is:

$$f(\mathbf{X}, \boldsymbol{\gamma}) = \gamma_0 \exp(\gamma_1 X) \tag{13.7}$$

Note that this model is not linear in the parameters γ_0 and γ_1.

A more general nonlinear exponential regression model in one predictor variable with normal error terms is:

$$Y_i = \gamma_0 + \gamma_1 \exp(\gamma_2 X_i) + \varepsilon_i \tag{13.8}$$

where the error terms are independent normal with constant variance σ^2. The response function for this regression model is:

$$f(\mathbf{X}, \boldsymbol{\gamma}) = \gamma_0 + \gamma_1 \exp(\gamma_2 X) \tag{13.9}$$

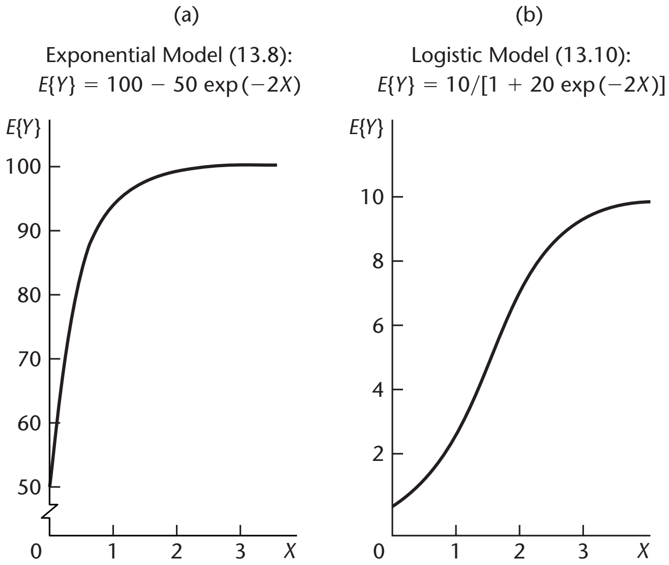
Exponential regression model (13.8) is commonly used in growth studies where the rate of growth at a given time X is proportional to the amount of growth remaining as time increases, with γ_0 representing the maximum growth value. Another use of this regression model is to relate the concentration of a substance (Y) to elapsed time (X). Figure 13.1a shows the response function (13.9) for parameter values $\gamma_0 = 100$, $\gamma_1 = -50$, and $\gamma_2 = -2$. We shall discuss exponential regression models (13.6) and (13.8) in more detail later in this chapter.

Logistic Regression Models. Another important nonlinear regression model is the *logistic regression model*. This model with one predictor variable and normal error terms is:

$$Y_i = \frac{\gamma_0}{1 + \gamma_1 \exp(\gamma_2 X_i)} + \varepsilon_i \tag{13.10}$$

where the error terms ε_i are independent normal with constant variance σ^2. The response

FIGURE 13.1 **Plots of Exponential and Logistic Response Functions.**



function here is:

$$f(\mathbf{X}, \boldsymbol{\gamma}) = \frac{\gamma_0}{1 + \gamma_1 \exp(\gamma_2 X)} \tag{13.11}$$

Note again that this response function is not linear in the parameters γ_0, γ_1, and γ_2.

This logistic regression model has been used in population studies to relate, for instance, number of species (Y) to time (X). Figure 13.1b shows the logistic response function (13.11) for parameter values $\gamma_0 = 10$, $\gamma_1 = 20$, and $\gamma_2 = -2$. Note that the parameter $\gamma_0 = 10$ represents the maximum growth value here.

Logistic regression model (13.10) is also widely used when the response variable is qualitative. An example of this use of the logistic regression model is predicting whether a household will purchase a new car this year (will, will not) on the basis of the predictor variables age of presently owned car, household income, and size of household. In this use of logistic regression models, the response variable (will, will not purchase car, in our example) is qualitative and will be represented by a 0, 1 indicator variable. Consequently, the error terms are not normally distributed here with constant variance. Logistic regression models and their use when the response variable is qualitative will be discussed in detail in Chapter 14.

General Form of Nonlinear Regression Models. As we have seen from the two examples of nonlinear regression models, these models are similar in general form to linear regression models. Each Y_i observation is postulated to be the sum of a mean response $f(\mathbf{X}_i, \boldsymbol{\gamma})$ based on the given nonlinear response function and a random error term ε_i. Furthermore, the error terms ε_i are often assumed to be independent normal random variables with constant variance.

An important difference of nonlinear regression models is that the number of regression parameters is not necessarily directly related to the number of X variables in the model. In linear regression models, if there are $p - 1$ X variables in the model, then there are p regression coefficients in the model. For the exponential regression model in (13.8), there is one X variable but three regression coefficients. The same is found for logistic regression model (13.10). Hence, we now denote the number of X variables in the nonlinear regression model by q, but we continue to denote the number of regression parameters in the response function by p. In the exponential regression model (13.6), for instance, there are $p = 2$ regression parameters and $q = 1$ X variable.

Also, we shall define the vector $\mathbf{X}_i$ of the observations on the X variables without the initial element 1. The general form of a nonlinear regression model is therefore expressed as follows:

$$Y_i = f(\mathbf{X}_i, \boldsymbol{\gamma}) + \varepsilon_i \tag{13.12}$$

where:

$$\underset{q\times 1}{\mathbf{X}_i} = \begin{bmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{iq} \end{bmatrix} \qquad \underset{p\times 1}{\boldsymbol{\gamma}} = \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_{p-1} \end{bmatrix} \tag{13.12a}$$

Comment

Nonlinear response functions that can be linearized by a transformation are sometimes called *intrinsically linear* response functions. For example, the exponential response function:

$$f(\mathbf{X}, \boldsymbol{\gamma}) = \gamma_0[\exp(\gamma_1 X)]$$

is an intrinsically linear response function because it can be linearized by the logarithmic transformation:

$$\log_e f(\mathbf{X}, \boldsymbol{\gamma}) = \log_e \gamma_0 + \gamma_1 X$$

This transformed response function can be represented in the linear model form:

$$g(\mathbf{X}, \boldsymbol{\gamma}) = \beta_0 + \beta_1 X$$

where $g(\mathbf{X}, \boldsymbol{\gamma}) = \log_e f(\mathbf{X}, \boldsymbol{\gamma})$, $\beta_0 = \log_e \gamma_0$, and $\beta_1 = \gamma_1$.

Just because a nonlinear response function is intrinsically linear does not necessarily imply that linear regression is appropriate. The reason is that the transformation to linearize the response function will affect the error term in the model. For example, suppose that the following exponential regression model with normal error terms that have constant variance is appropriate:

$$Y_i = \gamma_0 \exp(\gamma_1 X_i) + \varepsilon_i$$

A logarithmic transformation of Y to linearize the response function will affect the normal error term ε_i so that the error term in the linearized model will no longer be normal with constant variance. Hence, it is important to study any nonlinear regression model that has been linearized for appropriateness; it may turn out that the nonlinear regression model is preferable to the linearized version. ■

Estimation of Regression Parameters

Estimation of the parameters of a nonlinear regression model is usually carried out by the method of least squares or the method of maximum likelihood, just as for linear regression models. Also as in linear regression, both of these methods of estimation yield the same parameter estimates when the error terms in nonlinear regression model (13.12) are independent normal with constant variance.

Unlike linear regression, it is usually not possible to find analytical expressions for the least squares and maximum likelihood estimators for nonlinear regression models. Instead, numerical search procedures must be used with both of these estimation procedures, requiring intensive computations. The analysis of nonlinear regression models is therefore usually carried out by utilizing standard computer software programs.

Example

To illustrate the fitting and analysis of nonlinear regression models in a simple fashion, we shall use an example where the model has only two parameters and the sample size is reasonably small. In so doing, we shall be able to explain the concepts and procedures without overwhelming the reader with details.

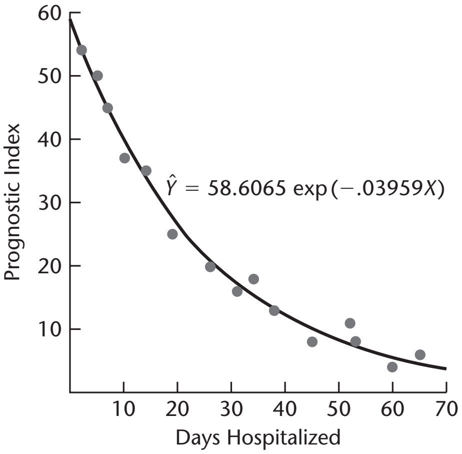
A hospital administrator wished to develop a regression model for predicting the degree of long-term recovery after discharge from the hospital for severely injured patients. The predictor variable to be utilized is number of days of hospitalization (X), and the response variable is a prognostic index for long-term recovery (Y), with large values of the index reflecting a good prognosis. Data for 15 patients were studied and are presented in Table 13.1. A scatter plot of the data is shown in Figure 13.2. Related earlier studies reported in the literature found the relationship between the predictor variable and the response variable to be exponential. Hence, it was decided to investigate the appropriateness of the two-parameter nonlinear exponential regression model (13.6):

$$Y_i = \gamma_0 \exp(\gamma_1 X_i) + \varepsilon_i \tag{13.13}$$

TABLE 13.1
Data—Severely Injured Patients Example.

Patient i	Days Hospitalized X_i	Prognostic Index Y_i
1	2	54
2	5	50
3	7	45
4	10	37
5	14	35
6	19	25
7	26	20
8	31	16
9	34	18
10	38	13
11	45	8
12	52	11
13	53	8
14	60	4
15	65	6

FIGURE 13.2
Scatter Plot and Fitted Nonlinear Regression Function—Severely Injured Patients Example.



where the ε_i are independent normal with constant variance. If this model is appropriate, it is desired to estimate the regression parameters γ_0 and γ_1.

13.2 Least Squares Estimation in Nonlinear Regression

We noted in Chapter 1 that the method of least squares for simple linear regression requires the minimization of the criterion Q in (1.8):

$$Q = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2 \tag{13.14}$$

Those values of β_0 and β_1 that minimize Q for the given sample observations (X_i, Y_i) are the least squares estimates and are denoted by b_0 and b_1.

We also noted in Chapter 1 that one method for finding the least squares estimates is by use of a numerical search procedure. With this approach, Q in (13.14) is evaluated for different values of β_0 and β_1, varying β_0 and β_1 systematically until the minimum value of Q is found. The values of β_0 and β_1 that minimize Q are the least squares estimates b_0 and b_1.

A second method for finding the least squares estimates is by means of the least squares normal equations. Here, the least squares normal equations are found analytically by differentiating Q with respect to β_0 and β_1 and setting the derivatives equal to zero. The solution of the normal equations yields the least squares estimates.

As we saw in Chapter 6, these procedures extend directly to multiple linear regression, for which the least squares criterion is given in (6.22). The concepts of least squares estimation for linear regression also extend directly to nonlinear regression models. The least squares criterion again is:

$$Q = \sum_{i=1}^{n} [Y_i - f(\mathbf{X}_i, \boldsymbol{\gamma})]^2 \tag{13.15}$$

where $f(\mathbf{X}_i, \boldsymbol{\gamma})$ is the mean response for the ith case according to the nonlinear response function $f(\mathbf{X}, \boldsymbol{\gamma})$. The least squares criterion Q in (13.15) must be minimized with respect to the nonlinear regression parameters $\gamma_0, \gamma_1, \dots, \gamma_{p-1}$ to obtain the least squares estimates. The same two methods for finding the least squares estimates—numerical search and normal equations—may be used in nonlinear regression. A difference from linear regression is that the solution of the normal equations usually requires an iterative numerical search procedure because analytical solutions generally cannot be found.

Example

The response function in the severely injured patients example is seen from (13.13) to be:

$$f(\mathbf{X}, \boldsymbol{\gamma}) = \gamma_0 \exp(\gamma_1 X)$$

Hence, the least squares criterion Q here is:

$$Q = \sum_{i=1}^{n} [Y_i - \gamma_0 \exp(\gamma_1 X_i)]^2$$

We can see that the method of maximum likelihood leads to the same criterion here when the error terms ε_i are independent normal with constant variance by considering the likelihood function:

$$L(\boldsymbol{\gamma}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} [Y_i - \gamma_0 \exp(\gamma_1 X_i)]^2\right]$$

Just as for linear regression, maximizing this likelihood function with respect to the regression parameters γ_0 and γ_1 is equivalent to minimizing the sum in the exponent, so that the maximum likelihood estimates are the same here as the least squares estimates.

We now discuss how to obtain the least squares estimates, first by use of the normal equations and then by direct numerical search procedures.

Solution of Normal Equations

To obtain the normal equations for a nonlinear regression model:

$$Y_i = f(\mathbf{X}_i, \boldsymbol{\gamma}) + \varepsilon_i$$

we need to minimize the least squares criterion Q:

$$Q = \sum_{i=1}^{n} [Y_i - f(\mathbf{X}_i, \boldsymbol{\gamma})]^2$$

with respect to $\gamma_0, \gamma_1, \ldots, \gamma_{p-1}$. The partial derivative of Q with respect to γ_k is:

$$\frac{\partial Q}{\partial \gamma_k} = \sum_{i=1}^{n} -2[Y_i - f(\mathbf{X}_i, \boldsymbol{\gamma})] \left[\frac{\partial f(\mathbf{X}_i, \boldsymbol{\gamma})}{\partial \gamma_k}\right] \tag{13.16}$$

When the p partial derivatives are each set equal to 0 and the parameters γ_k are replaced by the least squares estimates g_k, we obtain after some simplification the p normal equations:

$$\sum_{i=1}^{n} Y_i \left[\frac{\partial f(\mathbf{X}_i, \boldsymbol{\gamma})}{\partial \gamma_k}\right]_{\boldsymbol{\gamma}=\mathbf{g}} - \sum_{i=1}^{n} f(\mathbf{X}_i, \mathbf{g}) \left[\frac{\partial f(\mathbf{X}_i, \boldsymbol{\gamma})}{\partial \gamma_k}\right]_{\boldsymbol{\gamma}=\mathbf{g}} = 0 \qquad k = 0, 1, \ldots, p-1 \tag{13.17}$$

where $\mathbf{g}$ is the vector of the least squares estimates g_k:

$$\underset{p\times 1}{\mathbf{g}} = \begin{bmatrix} g_0 \\ g_1 \\ \vdots \\ g_{p-1} \end{bmatrix} \tag{13.18}$$

Note that the terms in brackets in (13.17) are the partial derivatives in (13.16) with the parameters γ_k replaced by the least squares estimates g_k.

The normal equations (13.17) for nonlinear regression models are nonlinear in the parameter estimates g_k and are usually difficult to solve, even in the simplest of cases. Hence, numerical search procedures are ordinarily required to obtain a solution of the normal equations iteratively. To make things still more difficult, multiple solutions may be possible.

Example

In the severely injured patients example, the mean response for the ith case is:

$$f(\mathbf{X}_i, \boldsymbol{\gamma}) = \gamma_0 \exp(\gamma_1 X_i) \tag{13.19}$$

Hence, the partial derivatives of $f(\mathbf{X}_i, \boldsymbol{\gamma})$ are:

$$\frac{\partial f(\mathbf{X}_i, \boldsymbol{\gamma})}{\partial \gamma_0} = \exp(\gamma_1 X_i) \tag{13.20a}$$

$$\frac{\partial f(\mathbf{X}_i, \boldsymbol{\gamma})}{\partial \gamma_1} = \gamma_0 X_i \exp(\gamma_1 X_i) \tag{13.20b}$$

Replacing γ_0 and γ_1 in (13.19), (13.20a), and (13.20b) by the respective least squares estimates g_0 and g_1, the normal equations (13.17) therefore are:

$$\begin{aligned}\sum Y_i \exp(g_1 X_i) - \sum g_0 \exp(g_1 X_i)\exp(g_1 X_i) &= 0\\ \sum Y_i g_0 X_i \exp(g_1 X_i) - \sum g_0 \exp(g_1 X_i) g_0 X_i \exp(g_1 X_i) &= 0\end{aligned}$$

Upon simplification, the normal equations become:

$$\begin{aligned}\sum Y_i \exp(g_1 X_i) - g_0 \sum \exp(2g_1 X_i) &= 0\\ \sum Y_i X_i \exp(g_1 X_i) - g_0 \sum X_i \exp(2g_1 X_i) &= 0\end{aligned}$$

These normal equations are not linear in g_0 and g_1, and no closed-form solution exists. Thus, numerical methods will be required to find the solution for the least squares estimates iteratively.

Direct Numerical Search—Gauss-Newton Method

In many nonlinear regression problems, it is more practical to find the least squares estimates by direct numerical search procedures rather than by first obtaining the normal equations and then using numerical methods to find the solution for these equations iteratively. The major statistical computer packages employ one or more direct numerical search procedures for solving nonlinear regression problems. We now explain one of these direct numerical search methods.

The *Gauss-Newton method,* also called the *linearization method,* uses a Taylor series expansion to approximate the nonlinear regression model with linear terms and then employs ordinary least squares to estimate the parameters. Iteration of these steps generally leads to a solution to the nonlinear regression problem.

The Gauss-Newton method begins with initial or starting values for the regression parameters $\gamma_0, \gamma_1, \ldots, \gamma_{p-1}$. We denote these by $g_0^{(0)}, g_1^{(0)}, \ldots, g_{p-1}^{(0)}$, where the superscript in parentheses denotes the iteration number. The starting values $g_k^{(0)}$ may be obtained from previous or related studies, theoretical expectations, or a preliminary search for parameter values that lead to a comparatively low criterion value Q in (13.15). We shall later discuss in more detail the choice of the starting values.

Once the starting values for the parameters have been obtained, we approximate the mean responses $f(\mathbf{X}_i, \boldsymbol{\gamma})$ for the n cases by the linear terms in the Taylor series expansion around the starting values $g_k^{(0)}$. We obtain for the ith case:

$$f(\mathbf{X}_i, \boldsymbol{\gamma}) \approx f\left(\mathbf{X}_i, \mathbf{g}^{(0)}\right) + \sum_{k=0}^{p-1}\left[\frac{\partial f(\mathbf{X}_i, \boldsymbol{\gamma})}{\partial \gamma_k}\right]_{\boldsymbol{\gamma}=\mathbf{g}^{(0)}}\left(\gamma_k - g_k^{(0)}\right) \tag{13.21}$$

where:

$$\underset{p\times 1}{\mathbf{g}^{(0)}} = \begin{bmatrix} g_0^{(0)} \\ g_1^{(0)} \\ \vdots \\ g_{p-1}^{(0)} \end{bmatrix} \tag{13.21a}$$

Note that $\mathbf{g}^{(0)}$ is the vector of the parameter starting values. The terms in brackets in (13.21) are the same partial derivatives of the regression function we encountered earlier in the normal equations (13.17), but here they are evaluated at $\gamma_k = g_k^{(0)}$ for $k = 0, 1, \ldots, p-1$.

Let us now simplify the notation as follows:

$$f_i^{(0)} = f\left(\mathbf{X}_i, \mathbf{g}^{(0)}\right) \tag{13.22a}$$

$$\beta_k^{(0)} = \gamma_k - g_k^{(0)} \tag{13.22b}$$

$$D_{ik}^{(0)} = \left[\frac{\partial f(\mathbf{X}_i, \boldsymbol{\gamma})}{\partial \gamma_k}\right]_{\boldsymbol{\gamma}=\mathbf{g}^{(0)}} \tag{13.22c}$$

The Taylor approximation (13.21) for the mean response for the ith case then becomes in this notation:

$$f(\mathbf{X}_i, \boldsymbol{\gamma}) \approx f_i^{(0)} + \sum_{k=0}^{p-1} D_{ik}^{(0)} \beta_k^{(0)}$$

and an approximation to the nonlinear regression model (13.12):

$$Y_i = f(\mathbf{X}_i, \boldsymbol{\gamma}) + \varepsilon_i$$

is:

$$Y_i \approx f_i^{(0)} + \sum_{k=0}^{p-1} D_{ik}^{(0)} \beta_k^{(0)} + \varepsilon_i \tag{13.23}$$

When we shift the $f_i^{(0)}$ term to the left and denote the difference $Y_i - f_i^{(0)}$ by $Y_i^{(0)}$, we obtain the following linear regression model approximation:

$$Y_i^{(0)} \approx \sum_{k=0}^{p-1} D_{ik}^{(0)} \beta_k^{(0)} + \varepsilon_i \qquad i = 1, \ldots, n \tag{13.24}$$

where:

$$Y_i^{(0)} = Y_i - f_i^{(0)} \tag{13.24a}$$

Note that the linear regression model approximation (13.24) is of the form:

$$Y_i = \beta_0 X_{i0} + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

The responses $Y_i^{(0)}$ in (13.24) are residuals, namely, the deviations of the observations around the nonlinear regression function with the parameters replaced by the starting estimates. The X variables observations $D_{ik}^{(0)}$ are the partial derivatives of the mean response evaluated for each of the n cases with the parameters replaced by the starting estimates. Each regression coefficient $\beta_k^{(0)}$ represents the difference between the true regression parameter and the initial estimate of the parameter. Thus, the regression coefficients represent the adjustment amounts by which the initial regression coefficients must be corrected. The purpose of fitting the linear regression model approximation (13.24) is therefore to estimate the regression coefficients $\beta_k^{(0)}$ and use these estimates to adjust the initial starting estimates of the regression parameters. In fitting this linear regression approximation, note that there is no intercept term in the model. Use of a computer multiple regression package therefore requires a specification of no intercept.

We shall represent the linear regression model approximation (13.24) in matrix form as follows:

$$\mathbf{Y}^{(0)} \approx \mathbf{D}^{(0)}\boldsymbol{\beta}^{(0)} + \boldsymbol{\varepsilon} \tag{13.25}$$

where:

(13.25a) $$\underset{n\times 1}{\mathbf{Y}^{(0)}} = \begin{bmatrix} Y_1 - f_1^{(0)} \\ \vdots \\ Y_n - f_n^{(0)} \end{bmatrix}$$

(13.25b) $$\underset{n\times p}{\mathbf{D}^{(0)}} = \begin{bmatrix} D_{10}^{(0)} & \cdots & D_{1,p-1}^{(0)} \\ \vdots & & \vdots \\ D_{n0}^{(0)} & \cdots & D_{n,p-1}^{(0)} \end{bmatrix}$$

(13.25c) $$\underset{p\times 1}{\boldsymbol{\beta}^{(0)}} = \begin{bmatrix} \beta_0^{(0)} \\ \vdots \\ \beta_{p-1}^{(0)} \end{bmatrix}$$

(13.25d) $$\underset{n\times 1}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Note again that the approximation model (13.25) is precisely in the form of the general linear regression model (6.19), with the **D** matrix of partial derivatives now playing the role of the **X** matrix (but without a column of 1s for the intercept). We can therefore estimate the parameters $\boldsymbol{\beta}^{(0)}$ by ordinary least squares and obtain according to (6.25):

$$\mathbf{b}^{(0)} = \left(\mathbf{D}^{(0)\prime}\mathbf{D}^{(0)}\right)^{-1}\mathbf{D}^{(0)\prime}\mathbf{Y}^{(0)} \tag{13.26}$$

where $\mathbf{b}^{(0)}$ is the vector of the least squares estimated regression coefficients. As we noted earlier, an ordinary multiple regression computer program can be used to obtain the estimated regression coefficients $b_k^{(0)}$, with a specification of no intercept.

We then use these least squares estimates to obtain revised estimated regression coefficients $g_k^{(1)}$ by means of (13.22b):

$$g_k^{(1)} = g_k^{(0)} + b_k^{(0)}$$

where $g_k^{(1)}$ denotes the revised estimate of γ_k at the end of the first iteration. In matrix form, we represent the revision process as follows:

$$\mathbf{g}^{(1)} = \mathbf{g}^{(0)} + \mathbf{b}^{(0)} \tag{13.27}$$

At this point, we can examine whether the revised regression coefficients represent adjustments in the proper direction. We shall denote the least squares criterion measure Q in (13.15) evaluated for the starting regression coefficients $\mathbf{g}^{(0)}$ by $SSE^{(0)}$; it is:

$$SSE^{(0)} = \sum_{i=1}^{n}\left[Y_i - f\left(\mathbf{X}_i, \mathbf{g}^{(0)}\right)\right]^2 = \sum_{i=1}^{n}\left(Y_i - f_i^{(0)}\right)^2 \tag{13.28}$$

At the end of the first iteration, the revised estimated regression coefficients are $\mathbf{g}^{(1)}$, and the least squares criterion measure evaluated at this stage, now denoted by $SSE^{(1)}$, is:

$$SSE^{(1)} = \sum_{i=1}^{n}\left[Y_i - f\left(\mathbf{X}_i, \mathbf{g}^{(1)}\right)\right]^2 = \sum_{i=1}^{n}\left(Y_i - f_i^{(1)}\right)^2 \tag{13.29}$$

If the Gauss-Newton method is working effectively in the first iteration, $SSE^{(1)}$ should be smaller than $SSE^{(0)}$ since the revised estimated regression coefficients $\mathbf{g}^{(1)}$ should be better estimates.

Note that the nonlinear regression functions $f(\mathbf{X}_i, \mathbf{g}^{(0)})$ and $f(\mathbf{X}_i, \mathbf{g}^{(1)})$ are used in calculating $SSE^{(0)}$ and $SSE^{(1)}$, and not the linear approximations from the Taylor series expansion.

The revised regression coefficients $\mathbf{g}^{(1)}$ are not, of course, the least squares estimates for the nonlinear regression problem because the fitted model (13.25) is only an approximation of the nonlinear model. The Gauss-Newton method therefore repeats the procedure just described, with $\mathbf{g}^{(1)}$ now used for the new starting values. This produces a new set of revised estimates, denoted by $\mathbf{g}^{(2)}$, and a new least squares criterion measure $SSE^{(2)}$. The iterative process is continued until the differences between successive coefficient estimates $\mathbf{g}^{(s+1)} - \mathbf{g}^{(s)}$ and/or the difference between successive least squares criterion measures $SSE^{(s+1)} - SSE^{(s)}$ become negligible. We shall denote the final estimates of the regression coefficients simply by **g** and the final least squares criterion measure, which is the error sum of squares, by SSE.

The Gauss-Newton method works effectively in many nonlinear regression applications. In some instances, however, the method may require numerous iterations before converging, and in a few cases it may not converge at all.

Example

In the severely injured patients example, the initial values of the parameters γ_0 and γ_1 were obtained by noting that a logarithmic transformation of the response function linearizes it:

$$\log_e \gamma_0[\exp(\gamma_1 X)] = \log_e \gamma_0 + \gamma_1 X$$

Hence, a linear regression model with a transformed Y variable was fitted as an initial approximation to the exponential model:

$$Y_i' = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where:

$$Y_i' = \log_e Y_i$$
$$\beta_0 = \log_e \gamma_0$$
$$\beta_1 = \gamma_1$$

This linear regression model was fitted by ordinary least squares and yielded the estimated regression coefficients $b_0 = 4.0371$ and $b_1 = -.03797$ (calculations not shown). Hence, the initial starting values are $g_0^{(0)} = \exp(b_0) = \exp(4.0371) = 56.6646$ and $g_1^{(0)} = b_1 = -.03797$.

The least squares criterion measure at this stage requires evaluation of the nonlinear regression function (13.7) for each case, utilizing the starting parameter values $g_0^{(0)}$ and $g_1^{(0)}$. For instance, for the first case, for which $X_1 = 2$, we obtain:

$$f\left(\mathbf{X}_1, \mathbf{g}^{(0)}\right) = f_1^{(0)} = g_0^{(0)} \exp\left(g_1^{(0)} X_1\right) = (56.6646) \exp[-.03797(2)] = 52.5208$$

TABLE 13.2 $\mathbf{Y}^{(0)}$ **and** $\mathbf{D}^{(0)}$ **Matrices—Severely Injured Patients Example.**

$$\underset{15\times 1}{\mathbf{Y}^{(0)}} = \begin{bmatrix} Y_1 - f_1^{(0)} \\ \cdot \\ \cdot \\ \cdot \\ Y_{15} - f_{15}^{(0)} \end{bmatrix} = \begin{bmatrix} Y_1 - g_0^{(0)} \exp(g_1^{(0)} X_1) \\ \cdot \\ \cdot \\ \cdot \\ Y_{15} - g_0^{(0)} \exp(g_1^{(0)} X_{15}) \end{bmatrix} = \begin{bmatrix} 1.4792 \\ 3.1337 \\ 1.5609 \\ -1.7624 \\ 1.6996 \\ -2.5422 \\ -1.1139 \\ -1.4629 \\ 2.4172 \\ -\ .3871 \\ -2.2625 \\ 3.1327 \\ .4259 \\ -1.8063 \\ 1.1977 \end{bmatrix}$$

$$\underset{15\times 2}{\mathbf{D}^{(0)}} = \begin{bmatrix} \exp(g_1^{(0)} X_1) & g_0^{(0)} X_1 \exp(g_1^{(0)} X_1) \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \exp(g_1^{(0)} X_{15}) & g_0^{(0)} X_{15} \exp(g_1^{(0)} X_{15}) \end{bmatrix} = \begin{bmatrix} .92687 & 105.0416 \\ .82708 & 234.3317 \\ .76660 & 304.0736 \\ .68407 & 387.6236 \\ .58768 & 466.2057 \\ .48606 & 523.3020 \\ .37261 & 548.9603 \\ .30818 & 541.3505 \\ .27500 & 529.8162 \\ .23625 & 508.7088 \\ .18111 & 461.8140 \\ .13884 & 409.0975 \\ .13367 & 401.4294 \\ .10247 & 348.3801 \\ .08475 & 312.1510 \end{bmatrix}$$

Since $Y_1 = 54$, the deviation from the mean response is:

$$Y_1^{(0)} = Y_1 - f_1^{(0)} = 54 - 52.5208 = 1.4792$$

Note again that the deviation $Y_1^{(0)}$ is the residual for case 1 at the initial fitting stage since $f_1^{(0)}$ is the estimated mean response when the initial estimates $\mathbf{g}^{(0)}$ of the parameters are employed. The stage 0 residuals for this and the other sample cases are presented in Table 13.2 and constitute the $\mathbf{Y}^{(0)}$ vector.

The least squares criterion measure at this initial stage then is simply the sum of the squared stage 0 residuals:

$$\begin{aligned} SSE^{(0)} &= \sum \left(Y_i - f_i^{(0)}\right)^2 = \sum \left(Y_i^{(0)}\right)^2 \\ &= (1.4792)^2 + \cdots + (1.1977)^2 = 56.0869 \end{aligned}$$

To revise the initial estimates for the parameters, we require the $\mathbf{D}^{(0)}$ matrix and the $\mathbf{Y}^{(0)}$ vector. The latter was already obtained in the process of calculating the least squares criterion measure at stage 0. To obtain the $\mathbf{D}^{(0)}$ matrix, we need the partial derivatives of the regression function (13.19) evaluated at $\boldsymbol{\gamma} = \mathbf{g}^{(0)}$. The partial derivatives are given in (13.20). Table 13.2 shows the $\mathbf{D}^{(0)}$ matrix entries in symbolic form and also the numerical values. To illustrate the calculations for case 1, we know from Table 13.1 that $X_1 = 2$. Hence, evaluating the partial derivatives at $\mathbf{g}^{(0)}$, we find:

$$D_{10}^{(0)} = \left[\frac{\partial f(\mathbf{X}_1, \boldsymbol{\gamma})}{\partial \gamma_0}\right]_{\boldsymbol{\gamma}=\mathbf{g}^{(0)}} = \exp\left(g_1^{(0)} X_1\right) = \exp[-.03797(2)] = .92687$$

$$\begin{aligned} D_{11}^{(0)} = \left[\frac{\partial f(\mathbf{X}_1, \boldsymbol{\gamma})}{\partial \gamma_1}\right]_{\boldsymbol{\gamma}=\mathbf{g}^{(0)}} &= g_0^{(0)} X_1 \exp\left(g_1^{(0)} X_1\right) \\ &= 56.6646(2)\exp[-.03797(2)] = 105.0416 \end{aligned}$$

We are now ready to obtain the least squares estimates $\mathbf{b}^{(0)}$ by regressing the response variable $Y^{(0)}$ in Table 13.2 on the two X variables in $\mathbf{D}^{(0)}$ in Table 13.2, using regression with no intercept. A standard multiple regression computer program yielded $b_0^{(0)} = 1.8932$ and $b_1^{(0)} = -.001563$. Hence, the vector $\mathbf{b}^{(0)}$ of the estimated regression coefficients is:

$$\mathbf{b}^{(0)} = \begin{bmatrix} 1.8932 \\ -.001563 \end{bmatrix}$$

By (13.27), we now obtain the revised least squares estimates $\mathbf{g}^{(1)}$:

$$\mathbf{g}^{(1)} = \mathbf{g}^{(0)} + \mathbf{b}^{(0)} = \begin{bmatrix} 56.6646 \\ -.03797 \end{bmatrix} + \begin{bmatrix} 1.8932 \\ -.001563 \end{bmatrix} = \begin{bmatrix} 58.5578 \\ -.03953 \end{bmatrix}$$

Hence, $g_0^{(1)} = 58.5578$ and $g_1^{(1)} = -.03953$ are the revised parameter estimates at the end of the first iteration. Note that the estimated regression coefficients have been revised moderately from the initial values, as can be seen from Table 13.3a, which presents the estimated regression coefficients and the least squares criterion measures for the starting values and the first iteration. Note also that the least squares criterion measure has been reduced in the first iteration.

Iteration 2 requires that we now revise the residuals from the exponential regression function and the first partial derivatives, based on the revised parameter estimates $g_0^{(1)} = 58.5578$ and $g_1^{(1)} = -.03953$. For case 1, for which $Y_1 = 54$ and $X_1 = 2$, we obtain:

$$\begin{aligned} Y_1^{(1)} &= Y_1 - f_1^{(1)} = 54 - (58.5578)\exp[-.03953(2)] = -.1065 \\ D_{10}^{(1)} &= \exp\left(g_1^{(1)} X_1\right) = \exp[-.03953(2)] = .92398 \\ D_{11}^{(1)} &= g_0^{(1)} X_1 \exp\left(g_1^{(1)} X_1\right) = 58.5578(2)\exp[-.03953(2)] = 108.2130 \end{aligned}$$

By comparing these results with the comparable stage 0 results for case 1 in Table 13.2, we see that the absolute magnitude of the residual for case 1 is substantially reduced as a result of the stage 1 revised fit and that the two partial derivatives are changed to a moderate extent. After the revised residuals $Y_i^{(1)}$ and the partial derivatives $D_{i0}^{(1)}$ and $D_{i1}^{(1)}$ have been

TABLE 13.3 **Gauss-Newton Method Iterations and Final Nonlinear Least Squares Estimates—Severely Injured Patients Example.**

(a) Estimates of Parameters and Least Squares Criterion Measure

Iteration	g_0	g_1	*SSE*
0	56.6646	−.03797	56.0869
1	58.5578	−.03953	49.4638
2	58.6065	−.03959	49.4593
3	58.6065	−.03959	49.4593

(b) Final Least Squares Estimates

k	g_k	$s\{g_k\}$
0	58.6065	1.472
1	−.03959	.00171

$$MSE = \frac{49.4593}{13} = 3.80456$$

(c) Estimated Approximate Variance-Covariance Matrix of Estimated Regression Coefficients

$$\mathbf{s}^2\{\mathbf{g}\} = MSE(\mathbf{D'D})^{-1} = 3.80456 \begin{bmatrix} 5.696\text{E}{-1} & -4.682\text{E}{-4} \\ -4.682\text{E}{-4} & 7.697\text{E}{-7} \end{bmatrix}$$

$$= \begin{bmatrix} 2.1672 & -1.781\text{E}{-3} \\ -1.781\text{E}{-3} & 2.928\text{E}{-6} \end{bmatrix}$$

obtained for all cases, the revised residuals are regressed on the revised partial derivatives, using a no-intercept regression fit, and the estimated regression parameters are again revised according to (13.27).

This process was carried out for three iterations. Table 13.3a contains the estimated regression coefficients and the least squares criterion measure for each iteration. We see that while iteration 1 led to moderate revisions in the estimated regression coefficients and a substantially better fit according to the least squares criterion, iteration 2 resulted only in minor revisions of the estimated regression coefficients and little improvement in the fit. Iteration 3 led to no change in either the estimates of the coefficients or the least squares criterion measure.

Hence, the search procedure was terminated after three iterations. The final regression coefficient estimates therefore are $g_0 = 58.6065$ and $g_1 = -.03959$, and the fitted regression function is:

$$\hat{Y} = (58.6065)\exp(-.03959X) \tag{13.30}$$

The error sum of squares for this fitted model is $SSE = 49.4593$. Figure 13.2 on page 515 shows a plot of this estimated regression function, together with a scatter plot of the data. The fit appears to be a good one.

Comments

1. The choice of initial starting values is very important with the Gauss-Newton method because a poor choice may result in slow convergence, convergence to a local minimum, or even divergence.

Good starting values will generally result in faster convergence, and if multiple minima exist, will lead to a solution that is the global minimum rather than a local minimum. Fast convergence, even if the initial estimates are far from the least squares solution, generally indicates that the linear approximation model (13.25) is a good approximation to the nonlinear regression model. Slow convergence, on the other hand, especially from initial estimates reasonably close to the least squares solution, usually indicates that the linear approximation model is not a good approximation to the nonlinear model.

2. A variety of methods are available for obtaining starting values for the regression parameters. Often, related earlier studies can be utilized to provide good starting values for the regression parameters. Another possibility is to select p representative observations, set the regression function $f(\mathbf{X}_i, \boldsymbol{\gamma})$ equal to Y_i for each of the p observations (thereby ignoring the random error), solve the p equations for the p parameters, and use the solutions as the starting values, provided they lead to reasonably good fits of the observed data. Still another possibility is to do a grid search in the parameter space by selecting in a grid fashion various trial choices of **g**, evaluating the least squares criterion Q for each of these choices, and using as the starting values that **g** vector for which Q is smallest.

3. When using the Gauss-Newton or another direct search procedure, it is often desirable to try other sets of starting values after a solution has been obtained to make sure that the same solution will be found.

4. Some computer packages for nonlinear regression require that the user specify the starting values for the regression parameters. Others do a grid search to obtain starting values.

5. Most nonlinear computer programs have a library of commonly used regression functions. For nonlinear response functions not in the library and specified by the user, some computer programs using the Gauss-Newton method require the user to input also the partial derivatives of the regression function, while others numerically approximate partial derivatives from the regression function.

6. The Gauss-Newton method may produce iterations that oscillate widely or result in increases in the error sum of squares. Sometimes, these aberrations are only temporary, but occasionally serious convergence problems exist. Various modifications of the Gauss-Newton method have been suggested to improve its performance, such as the Hartley modification (Ref. 13.1).

7. Some properties that exist for linear regression least squares do not hold for nonlinear regression least squares. For example, the residuals do not necessarily sum to zero for nonlinear least squares. Additionally, the error sum of squares *SSE* and the regression sum of squares *SSR* do not necessarily sum to the total sum of squares *SSTO*. Consequently, the coefficient of multiple determination $R^2 = SSR/SSTO$ is not a meaningful descriptive statistic for nonlinear regression. ■

Other Direct Search Procedures

Two other direct search procedures, besides the Gauss-Newton method, that are frequently used are the method of steepest descent and the Marquardt algorithm. The *method of steepest descent* searches for the minimum least squares criterion measure Q by iteratively determining the direction in which the regression coefficients **g** should be changed. The method of steepest descent is particularly effective when the starting values $\mathbf{g}^{(0)}$ are not good, being far from the final values **g**.

The *Marquardt algorithm* seeks to utilize the best features of the Gauss-Newton method and the method of steepest descent, and occupies a middle ground between these two methods.

Additional information about direct search procedures can be found in specialized sources, such as References 13.2 and 13.3.

13.3 Model Building and Diagnostics

The model-building process for nonlinear regression models often differs somewhat from that for linear regression models. The reason is that the functional form of many nonlinear models is less suitable for adding or deleting predictor variables and curvature and interaction effects in the direct fashion that is feasible for linear regression models. Some types of nonlinear regression models do lend themselves to adding and deleting predictor variables in a direct fashion. We shall take up two such nonlinear regression models in Chapter 14, where we consider the logistic and Poisson multiple regression models.

Validation of the selected nonlinear regression model can be performed in the same fashion as for linear regression models.

Use of diagnostic tools to examine the appropriateness of a fitted model plays an important role in the process of building a nonlinear regression model. The appropriateness of a regression model must always be considered, whether the model is linear or nonlinear. Nonlinear regression models may not be appropriate for the same reasons as linear regression models. For example, when nonlinear growth models are used for time series data, there is the possibility that the error terms may be correlated. Also, unequal error variances are often present when nonlinear growth models with asymptotes are fitted, such as exponential models (13.6) and (13.8). Typically, the error variances for cases in the neighborhood of the asymptote(s) differ from the error variances for cases elsewhere.

When replicate observations are available and the sample size is reasonably large, the appropriateness of a nonlinear regression function can be tested formally by means of the lack of fit test for linear regression models in (6.68). This test will be an approximate one for nonlinear regression models, but the actual level of significance will be close to the specified level when the sample size is reasonably large. Thus, we calculate the pure error sum of squares by (3.16), obtain the lack of fit sum of squares by (3.24), and calculate test statistic (6.68b) in the usual fashion when performing a formal lack of fit test for a nonlinear response function.

Plots of residuals against time, against the fitted values, and against each of the predictor variables can be helpful in diagnosing departures from the assumed model, just as for linear regression models. In interpreting residual plots for nonlinear regression, one needs to remember that the residuals for nonlinear regression do not necessarily sum to zero.

If unequal error variances are found to be present, weighted least squares can be used in fitting the nonlinear regression model. Alternatively, transformations of the response variable can be investigated that may stabilize the variance of the error terms and also permit use of a linear regression model.

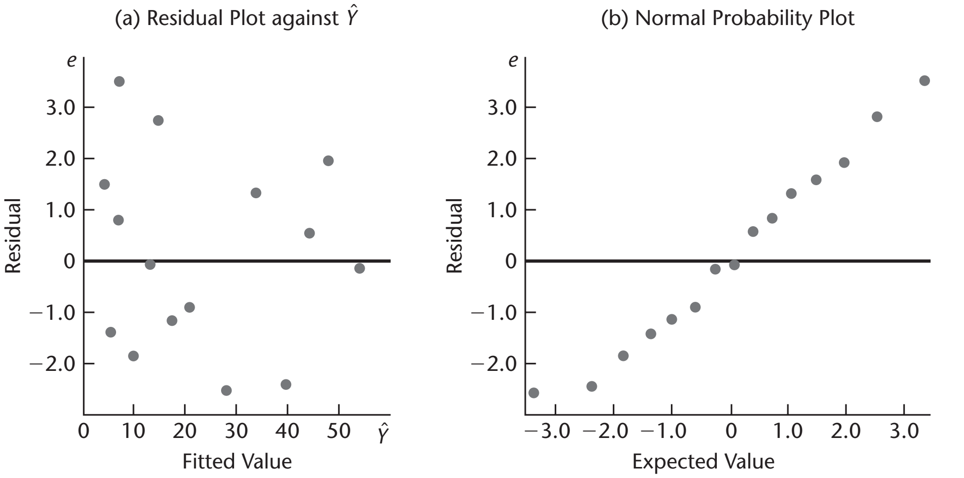
Example

In the severely injured patients example, the residuals were obtained by use of the fitted nonlinear regression function (13.30):

$$e_i = Y_i - (58.6065)\exp(-.03959X_i)$$

A plot of the residuals against the fitted values is shown in Figure 13.3a, and a normal probability plot of the residuals is shown in Figure 13.3b. These plots do not suggest any serious departures from the model assumptions. The residual plot against the fitted values in Figure 13.3a does raise the question whether the error variance may be somewhat larger for cases with small fitted values near the asymptote. The Brown-Forsythe test (3.9) was

FIGURE 13.3 **Diagnostic Residual Plots—Severely Injured Patients Example.**



conducted. Its P-value is .64, indicating that the residuals are consistent with constancy of the error variance.

On the basis of these, as well as some other diagnostics, it was concluded that exponential regression model (13.13) is appropriate for the data.

13.4 Inferences about Nonlinear Regression Parameters

Exact inference procedures about the regression parameters are available for linear regression models with normal error terms for any sample size. Unfortunately, this is not the case for nonlinear regression models with normal error terms, where the least squares and maximum likelihood estimators for any given sample size are not normally distributed, are not unbiased, and do not have minimum variance.

Consequently, inferences about the regression parameters in nonlinear regression are usually based on large-sample theory. This theory tells us that the least squares and maximum likelihood estimators for nonlinear regression models with normal error terms, when the sample size is large, are approximately normally distributed and almost unbiased, and have almost minimum variance. This large-sample theory also applies when the error terms are not normally distributed.

Before presenting details about large-sample inferences for nonlinear regression, we need to consider first how the error term variance σ^2 is estimated for nonlinear regression models.

Estimate of Error Term Variance

Inferences about nonlinear regression parameters require an estimate of the error term variance σ^2. This estimate is of the same form as for linear regression, the error sum of squares again being the sum of the squared residuals:

$$MSE = \frac{SSE}{n-p} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-p} = \frac{\sum[Y_i - f(\mathbf{X}_i, \mathbf{g})]^2}{n-p} \qquad \textbf{(13.31)}$$

Here **g** is the vector of the final parameter estimates, so that the residuals are the deviations around the fitted nonlinear regression function using the final estimated regression coefficients **g**. For nonlinear regression, *MSE* is not an unbiased estimator of σ^2, but the bias is small when the sample size is large.

Large-Sample Theory

When the error terms are independent and normally distributed and the sample size is reasonably large, the following theorem provides the basis for inferences for nonlinear regression models:

> When the error terms ε_i are independent $N(0, \sigma^2)$ and the sample size n is reasonably large, the sampling distribution of **g** is approximately normal. The expected value of the mean vector is approximately: **(13.32)**
>
> $$\mathbf{E}\{\mathbf{g}\} \approx \boldsymbol{\gamma} \tag{13.32a}$$
>
> The approximate variance-covariance matrix of the regression coefficients is estimated by:
>
> $$\mathbf{s}^2\{\mathbf{g}\} = MSE(\mathbf{D}'\mathbf{D})^{-1} \tag{13.32b}$$

Here **D** is the matrix of partial derivatives evaluated at the final least squares estimates **g**, just as $\mathbf{D}^{(0)}$ in (13.25b) is the matrix of partial derivatives evaluated at $\mathbf{g}^{(0)}$. Note that the estimated approximate variance-covariance matrix $\mathbf{s}^2\{\mathbf{g}\}$ is of exactly the same form as the one for linear regression in (6.48), with **D** again playing the role of the **X** matrix.

Thus, when the sample size is large and the error terms are independent normal with constant variance, the least squares estimators in **g** for nonlinear regression are approximately normally distributed and almost unbiased. They also have near minimum variance, since the variance-covariance matrix in (13.32b) estimates the minimum variances. We should add that theorem (13.32) holds even if the error terms are not normally distributed.

As a result of theorem (13.32), inferences for nonlinear regression parameters are carried out in the same fashion as for linear regression when the sample size is reasonably large. Thus, an interval estimate for a regression parameter is carried out by (6.50) and a test by (6.51). The needed estimated variance is obtained from the matrix $\mathbf{s}^2\{\mathbf{g}\}$ in (13.32b). These inference procedures when applied to nonlinear regression are only approximate, to be sure, but the approximation often is very good. For some nonlinear regression models, the sample size can be quite small for the large-sample approximation to be good. For other nonlinear regression models, however, the sample size may need to be quite large.

When Is Large-Sample Theory Applicable?

Ideally, we would like a rule that would tell us when the sample size in any given nonlinear regression application is large enough so that the large-sample inferences based on asymptotic theorem (13.32) are appropriate. Unfortunately, no simple rule exists that tells us when it is appropriate to use the large-sample inference methods and when it is not appropriate. However, a number of guidelines have been developed that are helpful in assessing the appropriateness of using the large-sample inference procedures in a given application.

1. Quick convergence of the iterative procedure in finding the estimates of the nonlinear regression parameters is often an indication that the linear approximation in (13.25) to

the nonlinear regression model is a good approximation and hence that the asymptotic properties of the regression estimates are applicable. Slow convergence suggests caution and consideration of other guidelines before large-sample inferences are employed.

2. Several measures have been developed for providing guidance about the appropriateness of the use of large-sample inference procedures. Bates and Watts (Ref. 13.4) developed curvature measures of nonlinearity. These indicate the extent to which the nonlinear regression function fitted to the data can be reasonably approximated by the linear approximation in (13.25). Box (Ref. 13.5) obtained a formula for estimating the bias of the estimated regression coefficients. A small bias supports the appropriateness of the large-sample inference procedures. Hougaard (Ref. 13.6) developed an estimate of the skewness of the sampling distributions of the estimated regression coefficients. An indication of little skewness supports the approximate normality of the sampling distributions and consequently the applicability of the large-sample inference procedures.

3. Bootstrap sampling described in Chapter 11 provides a direct means of examining whether the sampling distributions of the nonlinear regression parameter estimates are approximately normal, whether the variances of the sampling distributions are near the variances for the linear approximation model, and whether the bias in each of the parameter estimates is fairly small. If so, the sampling behavior of the nonlinear regression estimates is said to be *close-to-linear* and the large-sample inference procedures may appropriately be used. Nonlinear regression estimates whose sampling distributions are not close to normal, whose variances are much larger than the variances for the linear approximation model, and for which there is substantial bias are said to behave in a *far-from-linear* fashion and the large-sample inference procedures are then not appropriate.

Once many bootstrap samples have been obtained and the nonlinear regression parameter estimates calculated for each sample, the bootstrap sampling distribution for each parameter estimate can be examined to see if it is near normal. The variances of the bootstrap distributions of the estimated regression coefficients can be obtained next to see if they are close to the large-sample variance estimates obtained by (13.32b). Similarly, the bootstrap confidence intervals for the regression coefficients can be obtained and compared with the large-sample confidence intervals. Good agreement between these intervals again provides support for the appropriateness of the large-sample inference procedures. In addition, the difference between each final regression parameter estimate and the mean of its bootstrap sampling distribution is an estimate of the bias of the regression estimate. Small or negligible biases of the nonlinear regression estimates support the appropriateness of the large-sample inference procedures.

Remedial Measures. When the diagnostics suggest that large-sample inference procedures are not appropriate in a particular instance, remedial measures should be explored. One possibility is to reparameterize the nonlinear regression model. For example, studies have shown that for the nonlinear model:

$$Y_i = \gamma_0 X_i/(\gamma_1 + X_i) + \varepsilon_i$$

the use of large-sample inference procedures is often not appropriate. However, the following reparameterization:

$$Y_i = X_i/(\theta_1 X_i + \theta_2) + \varepsilon_i$$

where $\theta_1 = 1/\gamma_0$ and $\theta_2 = \gamma_1/\gamma_0$, yields identical fits and generally involves no problems in using large-sample inference procedures for moderate sample sizes (see Ref. 13.7 for details).

Another remedial measure is to use the bootstrap estimates of precision and confidence intervals instead of the large-sample inferences. However, when the linear approximation in (13.25) is not a close approximation to the nonlinear regression model, convergence may be very slow and bootstrap estimates of precision and confidence intervals may be difficult to obtain. Still another remedial measure that is sometimes available is to increase the sample size.

Example

For the severely injured patients example, we know from Table 13.3a on page 524 that the final error sum of squares is $SSE = 49.4593$. Since $p = 2$ parameters are present in the nonlinear response function (13.19), we obtain:

$$MSE = \frac{SSE}{n-p} = \frac{49.4593}{15-2} = 3.80456$$

Table 13.3b presents this mean square, and Table 13.3c contains the large-sample estimated variance-covariance matrix of the estimated regression coefficients. The matrix $(\mathbf{D}'\mathbf{D})^{-1}$ is based on the final regression coefficient estimates **g** and is shown without computational details.

We see from Table 13.3c that $s^2\{g_0\} = 2.1672$ and $s^2\{g_1\} = .000002928$. The estimated standard deviations of the regression coefficients are given in Table 13.3b.

To check on the appropriateness of the large-sample variances of the estimated regression coefficients and on the applicability of large-sample inferences in general, we have generated 1,000 bootstrap samples of size 15. The fixed X sampling procedure was used since the exponential model appears to fit the data well and the error term variance appears to be fairly constant. Histograms of the resulting bootstrap sampling distributions of g_0^* and g_1^* are shown in Figure 13.4, together with some characteristics of these distributions. We see that the g_0^* distribution is close to normal. The g_1^* distribution suggests that the sampling distribution may be slightly skewed to the left, but the departure from normality does not appear to be great. The means of the distribution, denoted by $\bar{g}_0^*$ and $\bar{g}_1^*$, are very close to the final least squares estimates, indicating that the bias in the estimates is negligible:

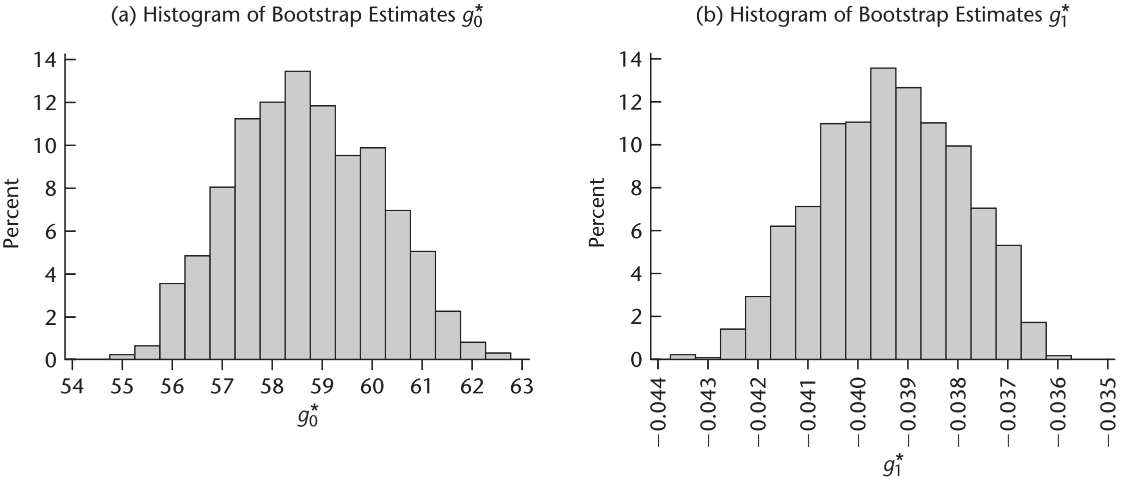
$$\begin{aligned} \bar{g}_0^* &= 58.67 \qquad & \bar{g}_1^* &= -.03936 \\ g_0 &= 58.61 \qquad & g_1 &= -.03959 \end{aligned}$$

Furthermore, the standard deviations of the bootstrap sampling distributions are very close to the large-sample standard deviations in Table 13.3b:

$$\begin{aligned} s^*\{g_0^*\} &= 1.423 \qquad & s^*\{g_1^*\} &= .00142 \\ s\{g_0\} &= 1.472 \qquad & s\{g_1\} &= .00171 \end{aligned}$$

These indications all point to the appropriateness of large-sample inferences here, even though the sample size ($n = 15$) is not very large.

FIGURE 13.4 **Bootstrap Sampling Distributions—Severely Injured Patients Example.**



$\bar{g}_0^* = 58.67$

$s^*\{g_0^*\} = 1.423$

$g_0^*(.025) = 56.044$

$g_0^*(.975) = 61.436$

$\bar{g}_1^* = -.03936$

$s^*\{g_1^*\} = .00142$

$g_1^*(.025) = -.04207$

$g_1^*(.975) = -.03681$

Interval Estimation of a Single γ_k

Based on large-sample theorem (13.32), the following approximate result holds when the sample size is large and the error terms are normally distributed:

$$\frac{g_k - \gamma_k}{s\{g_k\}} \sim t(n-p) \qquad k = 0, 1, \ldots, p-1 \tag{13.33}$$

where $t(n-p)$ is a t variable with $n-p$ degrees of freedom. Hence, approximate $1-\alpha$ confidence limits for any single γ_k are formed by means of (6.50):

$$g_k \pm t(1-\alpha/2; n-p)s\{g_k\} \tag{13.34}$$

where $t(1-\alpha/2;\, n-p)$ is the $(1-\alpha/2)100$ percentile of the t distribution with $n-p$ degrees of freedom.

Example

For the severely injured patients example, it is desired to estimate γ_1 with a 95 percent confidence interval. We require $t(.975; 13) = 2.160$, and find from Table 13.3b that $g_1 = -.03959$ and $s\{g_1\} = .00171$. Hence, the confidence limits are $-.03959 \pm 2.160(.00171)$, and the approximate 95 percent confidence interval for γ_1 is:

$$-.0433 \leq \gamma_1 \leq -.0359$$

Thus, we can conclude with approximate 95 percent confidence that γ_1 is between $-.0433$ and $-.0359$. To confirm the appropriateness of this large-sample confidence interval, we shall obtain the 95 percent bootstrap confidence interval for γ_1. Using (11.58) and the results in Figure 13.4b, we obtain:

$$d_1 = g_1 - g_1^*(.025) = -.03959 + .04207 = .00248$$
$$d_2 = g_1^*(.975) - g_1 = -.03681 + .03959 = .00278$$

The reflection method confidence limits by (11.59) then are:

$$g_1 - d_2 = -.03959 - .00278 = -.04237$$
$$g_1 + d_1 = -.03959 + .00248 = -.03711$$

Hence, the 95 percent bootstrap confidence interval is $-.0424 \leq \gamma_1 \leq -.0371$. This confidence interval is very close to the large-sample confidence interval, again supporting the appropriateness of large-sample inference procedures here.

Simultaneous Interval Estimation of Several γ_k

Approximate joint confidence intervals for several regression parameters in nonlinear regression can be developed by the Bonferroni procedure. If m parameters are to be estimated with approximate family confidence coefficient $1 - \alpha$, the joint Bonferroni confidence limits are:

$$g_k \pm Bs\{g_k\} \tag{13.35}$$

where:

$$B = t(1 - \alpha/2m; n - p) \tag{13.35a}$$

Example

In the severely injured patients example, it is desired to obtain simultaneous interval estimates for γ_0 and γ_1 with an approximate 90 percent family confidence coefficient. With the Bonferroni procedure we therefore require separate confidence intervals for the two parameters, each with a 95 percent statement confidence coefficient. We have already obtained a confidence interval for γ_1 with a 95 percent statement confidence coefficient. The approximate 95 percent statement confidence limits for γ_0, using the results in Table 13.3b, are $58.6065 \pm 2.160(1.472)$ and the confidence interval for γ_0 is:

$$55.43 \leq \gamma_0 \leq 61.79$$

Hence, the joint confidence intervals with approximate family confidence coefficient of 90 percent are:

$$55.43 \leq \gamma_0 \leq 61.79$$
$$-.0433 \leq \gamma_1 \leq -.0359$$

Test Concerning a Single γ_k

A large-sample test concerning a single γ_k is set up in the usual fashion. To test:

$$\begin{aligned} H_0&: \gamma_k = \gamma_{k0} \\ H_a&: \gamma_k \neq \gamma_{k0} \end{aligned} \tag{13.36a}$$

where γ_{k0} is the specified value of γ_k, we may use the t^* test statistic based on (6.49) when n is reasonably large:

$$t^* = \frac{g_k - \gamma_{k0}}{s\{g_k\}} \tag{13.36b}$$

The decision rule for controlling the risk of making a Type I error at approximately α then is:

$$\begin{aligned} &\text{If } |t^*| \leq t(1-\alpha/2; n-p), \text{ conclude } H_0 \\ &\text{If } |t^*| > t(1-\alpha/2; n-p), \text{ conclude } H_a \end{aligned} \tag{13.36c}$$

Example

In the severely injured patients example, we wish to test:

$$H_0\text{: } \gamma_0 = 54$$
$$H_a\text{: } \gamma_0 \neq 54$$

The test statistic (13.36b) here is:

$$t^* = \frac{58.6065 - 54}{1.472} = 3.13$$

For $\alpha = .01$, we require $t(.995; 13) = 3.012$. Since $|t^*| = 3.13 > 3.012$, we conclude H_a, that $\gamma_0 \neq 54$. The approximate two-sided P-value of the test is .008.

Test Concerning Several γ_k

When a large-sample test concerning several γ_k simultaneously is desired, we use the same approach as for the general linear test, first fitting the full model and obtaining $SSE(F)$, then fitting the reduced model and obtaining $SSE(R)$, and finally calculating the same test statistic (2.70) as for linear regression:

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div MSE(F) \tag{13.37}$$

For large n, this test statistic is distributed approximately as $F(df_R - df_F, df_F)$ when H_0 holds.

13.5 Learning Curve Example

We now present a second example, to provide an additional illustration of the nonlinear regression concepts developed in this chapter. An electronics products manufacturer undertook the production of a new product in two locations (location A: coded $X_1 = 1$, location B: coded $X_1 = 0$). Location B has more modern facilities and hence was expected to be more efficient than location A, even after the initial learning period. An industrial engineer calculated the expected unit production cost for a modern facility after learning has occurred. Weekly unit production costs for each location were then expressed as a fraction of this expected cost. The reciprocal of this fraction is a measure of relative efficiency, and this relative efficiency measure was utilized as the response variable (Y) in the study.

It is well known that efficiency increases over time when a new product is produced, and that the improvements eventually slow down and the process stabilizes. Hence, it was decided to employ an exponential model with an upper asymptote for expressing the relation between relative efficiency (Y) and time (X_2), and to incorporate a constant effect for the difference in the two production locations. The model decided on was:

$$Y_i = \gamma_0 + \gamma_1 X_{i1} + \gamma_3 \exp(\gamma_2 X_{i2}) + \varepsilon_i \tag{13.38}$$

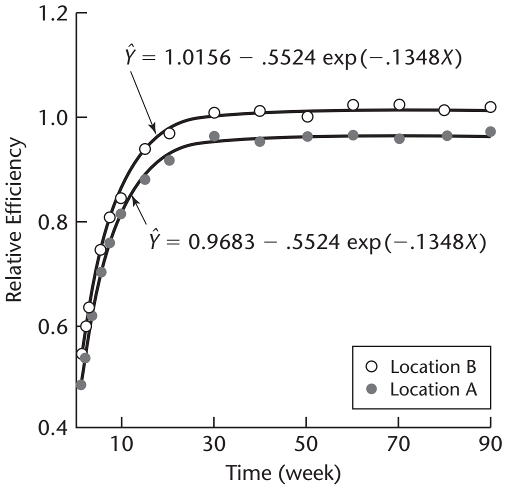
When γ_2 and γ_3 are negative, γ_0 is the upper asymptote for location B as X_2 gets large, and $\gamma_0 + \gamma_1$ is the upper asymptote for location A. The parameters γ_2 and γ_3 reflect the speed of learning, which was expected to be the same in the two locations.

While weekly data on relative production efficiency for each location were available, we shall only use observations for selected weeks during the first 90 weeks of production to simplify the presentation. A portion of the data on location, week, and relative efficiency is presented in Table 13.4; a plot of the data is shown in Figure 13.5. Note that learning was relatively rapid in both locations, and that the relative efficiency in location B toward the

TABLE 13.4 Data—Learning Curve Example.

Observation	Location	Week	Relative Efficiency
i	X_{i1}	X_{i2}	Y_i
1	1	1	.483
2	1	2	.539
3	1	3	.618
...	...	...	...
13	1	70	.960
14	1	80	.967
15	1	90	.975
16	0	1	.517
17	0	2	.598
18	0	3	.635
...	...	...	...
28	0	70	1.028
29	0	80	1.017
30	0	90	1.023

FIGURE 13.5 Scatter Plot and Fitted Nonlinear Regression Functions—Learning Curve Example.



end of the 90-week period even exceeded 1.0; i.e., the actual unit costs at this stage were lower than the industrial engineer's expected unit cost.

Regression model (13.38) is nonlinear in the parameters γ_2 and γ_3. Hence, a direct numerical search estimation procedure was to be employed, for which starting values for the parameters are needed. These were developed partly from past experience, partly from analysis of the data. Previous studies indicated that γ_3 should be in the neighborhood of $-.5$, so $g_3^{(0)} = -.5$ was used as the starting value. Since the difference in the relative efficiencies between locations A and B for a given week tended to average $-.0459$ during the 90-week period, a starting value $g_1^{(0)} = -.0459$ was specified. The largest observed relative efficiency for location B was 1.028, so that a starting value $g_0^{(0)} = 1.025$ was felt to be reasonable. Only a starting value for γ_2 remains to be found. This was chosen by selecting a typical relative efficiency observation in the middle of the time period, $Y_{24} = 1.012$, and equating it to the response function with $X_{24,1} = 0$, $X_{24,2} = 30$, and the starting values for the other regression coefficients (thus ignoring the error term):

$$1.012 = 1.025 - (.5)\exp(30\gamma_2)$$

Solving this equation for γ_2, the starting value $g_2^{(0)} = -.122$ was obtained. Tests for several other representative observations yielded similar starting values, and $g_2^{(0)} = -.122$ was therefore considered to be a reasonable initial value.

With the four starting values $g_0^{(0)} = 1.025$, $g_1^{(0)} = -.0459$, $g_2^{(0)} = -.122$, and $g_3^{(0)} = -.5$, a computer package direct numerical search program was utilized to obtain the least squares estimates. The least squares regression coefficients stabilized after five iterations. The final estimates, together with the large-sample estimated standard deviations of their sampling distributions, are presented in Table 13.5, columns 1 and 2. The fitted regression function is:

$$\hat{Y} = 1.0156 - .04727X_1 - (.5524)\exp(-.1348X_2) \tag{13.39}$$

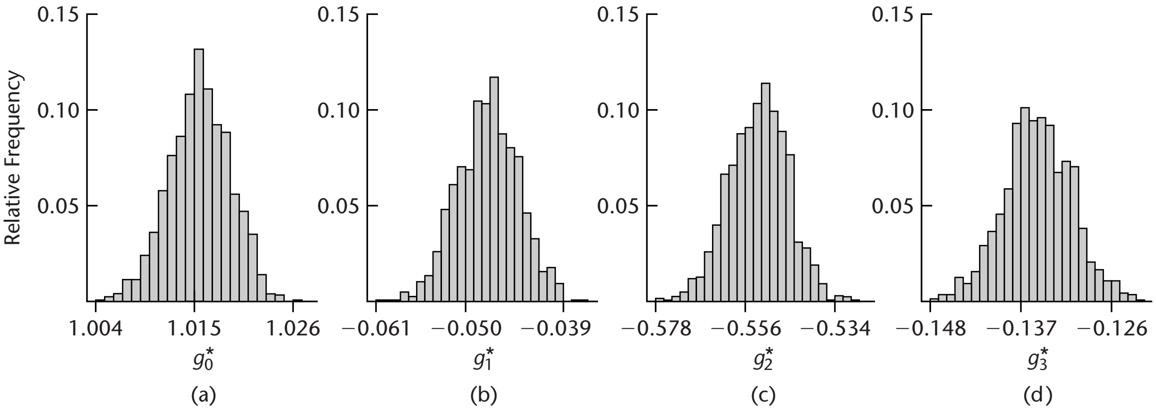
The error sum of squares is $SSE = .00329$, with $30-4=26$ degrees of freedom. Figure 13.5 presents the fitted regression functions for the two locations, together with a plot of the data. The fit seems to be quite good, and residual plots (not shown) did not indicate any noticeable departures from the assumed model.

In order to explore the applicability of large-sample inference procedures here, bootstrap fixed X sampling was employed. One thousand bootstrap samples of size 30 were generated.

TABLE 13.5 **Nonlinear Least Squares Estimates and Standard Deviations and Bootstrap Results—Learning Curve Example.**

	(1)	(2)	(3)	(4)
	Nonlinear Least Squares		Bootstrap	
k	g_k	$s\{g_k\}$	$\bar{g}_k^*$	$s^*\{g_k^*\}$
0	1.0156	.003672	1.015605	.003374
1	−.04727	.004109	−.04724	.003702
2	−.5524	.008157	−.55283	.007275
3	−.1348	.004359	−.13495	.004102

FIGURE 13.6 **MINITAB Histograms of Bootstrap Sampling Distributions—Learning Curve Example.**



The estimated bootstrap means and standard deviations for each of the sampling distributions are presented in Table 13.5, columns 3 and 4. Note first that each least squares estimate g_k in column 1 of Table 13.5 is very close to the mean $\bar{g}_k^*$ of its respective bootstrap sampling distribution in column 3, indicating that the estimates have very little bias. Note also that each large-sample standard deviation $s\{g_k\}$ in column 2 of Table 13.5 is fairly close to the respective bootstrap standard deviation $s^*\{g_k^*\}$ in column 4, again supporting the applicability of large-sample inference procedures here. Finally, we present in Figure 13.6 MINITAB plots of the histograms of the four bootstrap sampling distributions. They appear to be consistent with approximately normal sampling distributions. These results all indicate that the sampling behavior of the nonlinear regression estimates is close to linear and therefore support the use of large-sample inferences here.

There was special interest in the parameter γ_1, which reflects the effect of location. An approximate 95 percent confidence interval is to be constructed. We require $t(.975;26) = 2.056$. The estimated standard deviation from Table 13.5 is $s\{g_1\} = .004109$. Hence, the approximate 95 percent confidence limits for γ_1 are $-.04727 \pm 2.056(.004109)$, and the confidence interval for γ_1 is:

$$-.0557 \leq \gamma_1 \leq -.0388$$

An approximate 95 percent confidence interval for γ_1 by the bootstrap reflection method was also obtained for comparative purposes using (11.59). It is:

$$-.0547 \leq \gamma_1 \leq -.0400$$

This is very close to that obtained by large-sample inference procedures. Since γ_1 is seen to be negative, these confidence intervals confirm that location A with its less modern facilities tends to be less efficient.

Comments

1. When learning curve models are fitted to data constituting repeated observations on the same unit, such as efficiency data for the same production unit at different points in time, the error terms may be correlated. Hence, in these situations it is important to ascertain whether or not a model assuming

uncorrelated error terms is reasonable. In the learning curve example, a plot of the residuals against time order did not suggest any serious correlations among the error terms.

2. With learning curve models, it is not uncommon to find that the error variances are unequal. Again, therefore, it is important to check whether the assumption of constancy of error variance is reasonable. In the learning curve example, plots of the residuals against the fitted values and time did not suggest any serious heteroscedasticity problem. ■

13.6 Introduction to Neural Network Modeling

In recent years there has been an explosion in the amount of available data, made possible in part by the widespread availability of low-cost computer memory and automated data collection systems. The regression modeling techniques discussed to this point in this book typically were developed for use with data sets involving fewer than 1,000 observations and fewer than 50 predictors. Yet it is not uncommon now to be faced with data sets involving perhaps millions of observations and hundreds or thousands of predictors. Examples include point-of-sale data in marketing, credit card scoring data, on-line monitoring of production processes, optical character recognition, internet e-mail filtering data, microchip array data, and computerized medical record data. This exponential growth in available data has motivated researchers in the fields of statistics, artificial intelligence, and data mining to develop simple, flexible, powerful procedures for data modeling that can be applied to very large data sets. In this section we discuss one such technique, neural network modeling.

Neural Network Model

The basic idea behind the neural network approach is to model the response as a nonlinear function of various linear combinations of the predictors. Recall that our standard multiple regression model (6.7) involves just one linear combination of the predictors, namely $E\{Y_i\} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1}$. Thus, as we will demonstrate, the neural network model is simply a nonlinear statistical model that contains many more parameters than the corresponding linear statistical model. One result of this is that the models will typically be overparameterized, resulting in parameters that are uninterpretable, which is a major shortcoming of neural network modeling. An advantage of the neural network approach is that the resulting model will often perform better in predicting future responses than a standard regression model. Such models require large data sets, and are evaluated solely on their ability to predict responses in hold-out (validation) data sets.

In this section we describe the simplest, but most widely used, neural network model, the *single-hidden-layer, feedforward neural network*. This network is sometimes referred to as a *single-layer perceptron*. In a neural network model the ith response Y_i is modeled as a nonlinear function g_Y of m *derived predictor values*, $H_{i0}, H_{i1}, \ldots, H_{i,m-1}$:

$$Y_i = g_Y(\beta_0 H_{i0} + \beta_1 H_{i1} + \cdots + \beta_{i,m-1} H_{i,m-1}) + \varepsilon_i = g_Y(\mathbf{H}_i'\boldsymbol{\beta}) + \varepsilon_i \quad \textbf{(13.40)}$$

where:

$$\underset{m\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{m-1} \end{bmatrix} \qquad \underset{m\times 1}{\mathbf{H}_i} = \begin{bmatrix} H_{i0} \\ H_{i1} \\ \vdots \\ H_{i,m-1} \end{bmatrix} \quad \textbf{(13.40a)}$$

We take H_{i0} equal to 1 and for $j = 1, \ldots, p-1$, the jth derived predictor value for the ith observation, H_{ij}, is a nonlinear function g_j of a linear combination of the original predictors:

$$H_{ij} = g_j(\mathbf{X}_i'\boldsymbol{\alpha}_j) \qquad j = 1, \ldots, m-1 \tag{13.41}$$

where:

$$\underset{p\times 1}{\boldsymbol{\alpha}_j} = \begin{bmatrix} \alpha_{j0} \\ \alpha_{j1} \\ \vdots \\ \alpha_{j,p-1} \end{bmatrix} \qquad \underset{p\times 1}{\mathbf{X}_i} = \begin{bmatrix} X_{i0} \\ X_{i1} \\ \vdots \\ X_{i,p-1} \end{bmatrix} \tag{13.41a}$$

and where $X_{i0} = 1$. Note that $\mathbf{X}_i'$ is the ith row of the $\mathbf{X}$ matrix. Equations (13.40) and (13.41) together form the neural network model:

$$Y_i = g_Y(\mathbf{H}_i'\boldsymbol{\beta}) + \varepsilon_i = g_Y\left[\beta_0 + \sum_{j=1}^{m-1}\beta_j g_j(\mathbf{X}_i'\boldsymbol{\alpha}_j)\right] + \varepsilon_i \tag{13.42}$$

The m functions $g_Y, g_1, \ldots, g_{m-1}$ are called *activation functions* in the neural networks literature. To completely specify the neural network model, it is necessary to identify the m activation functions. A common choice for each of these functions is the logistic function:

$$g(Z) = \frac{1}{1+e^{-Z}} = [1+e^{-Z}]^{-1} \tag{13.43}$$

This function is flexible and can be adapted to a variety of circumstances.

As a simple example, consider the case of a single predictor, X_1. Then from (13.41), the jth derived predictor for the ith observation is:

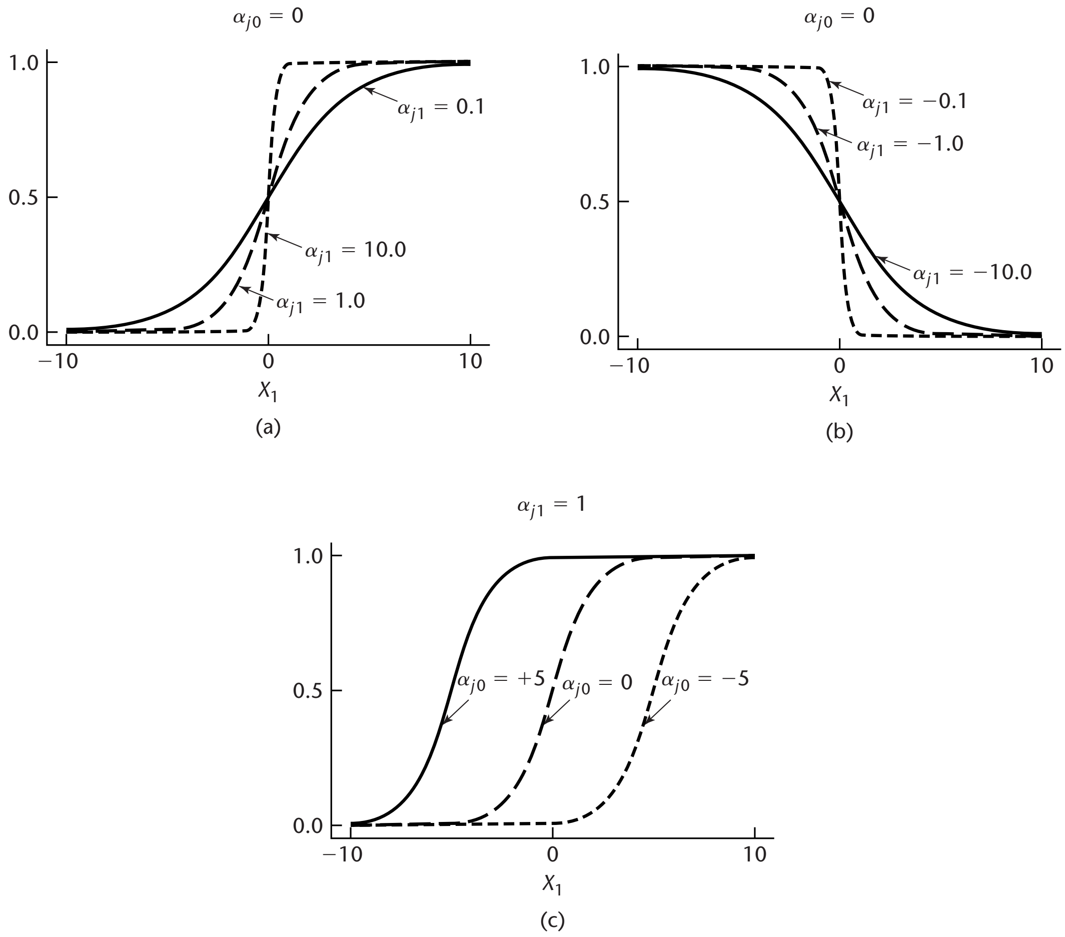
$$g_j(\mathbf{X}_i'\boldsymbol{\alpha}_j) = [1+\exp(-\alpha_{j0} - \alpha_{j1}X_{i1})]^{-1} \tag{13.44}$$

(Note that (13.44) is a reparameterization of (13.11), with $\gamma_0 = 1$, $\gamma_1 = e^{-\alpha_{j0}}$, and $\gamma_2 = -\alpha_{j1}$.) This function is shown in Figure 13.7 for various choices of α_{j0} and α_{j1}. In Figure 13.7a, the logistic function is plotted for fixed $\alpha_{j0} = 0$, and $\alpha_{j1} = .1$, 1, and 10. When $\alpha_{j1} = .1$, the logistic function is approximately linear over a wide range; when $\alpha_{j1} = 10$, the function is highly nonlinear in the center of the plot. Generally, relatively larger parameters (in absolute value) are required for highly nonlinear responses, and relatively smaller parameters result for approximately linear responses. Changing the sign of α_{j1} reverses the orientation of the logistic function, as shown in Figure 13.7b. Finally, for a given value of α_{j1}, the position of the logistic function along the X_1-axis is controlled by α_{j0}. In Figure 13.7c, the logistic function is plotted for fixed $\alpha_{j1} = 1$ and $\alpha_{j0} = -5$, 0, and 5. Note that all of the plots in Figure 13.7 reflect a characteristic *S*- or *sigmoidal*-shape, and the fact that the logistic function has a maximum of 1 and a minimum of 0.

Substitution of g in (13.43) for each of $g_Y, g_1, \ldots, g_{m-1}$ in (13.42) yields the specific neural network model to be discussed in this section:

$$\begin{aligned} Y_i &= [1+\exp(-\mathbf{H}_i'\boldsymbol{\beta})]^{-1} + \varepsilon_i \\ &= \left[1+\exp\left[-\beta_0 - \sum_{j=1}^{m-1}\beta_j[1+\exp(-\mathbf{X}_i'\boldsymbol{\alpha}_j)]^{-1}\right]\right]^{-1} + \varepsilon_i \\ &= f(\mathbf{X}_i, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{m-1}, \boldsymbol{\beta}) + \varepsilon_i \end{aligned} \tag{13.45}$$

FIGURE 13.7 Various Logistic Activation Functions for Single Predictor.



where:

$\boldsymbol{\beta}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{m-1}$ are unknown parameter vectors
$\mathbf{X}_i$ is a vector of known constants
ε_i are residuals

Neural network model (13.45) is a special case of (13.12) and is therefore a nonlinear regression model. In principle, all of the methods discussed in this chapter for estimation, testing, and prediction with nonlinear models are applicable. Indeed, any nonlinear regression package can be used to estimate the unknown coefficients. Recall, however, that these models are generally overparameterized, and use of standard estimation methods will result in fitted models that have poor predictive ability. This is analogous to leaving too many unimportant predictors in a linear regression model. Special procedures for fitting model (13.45) that lead to better prediction will be considered later in this section.

Note that because the logistic activation function is bounded between 0 and 1, it is necessary to scale Y_i so that the scaled value, Y_i^{sc} also falls within these limits. This can be accomplished by using:

$$Y_i^{sc} = \frac{Y_i - Y_{\min}}{Y_{\max} - Y_{\min}}$$

where $Y_{\min}$ and $Y_{\max}$ are the minimum and maximum responses. It is also common practice to center and scale each of the predictors to have mean 0 and standard deviation 1. These transformations are generally handled automatically by neural network software.

Network Representation

Network diagrams are often used to depict a neural network model. Note that the standard linear regression function:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}$$

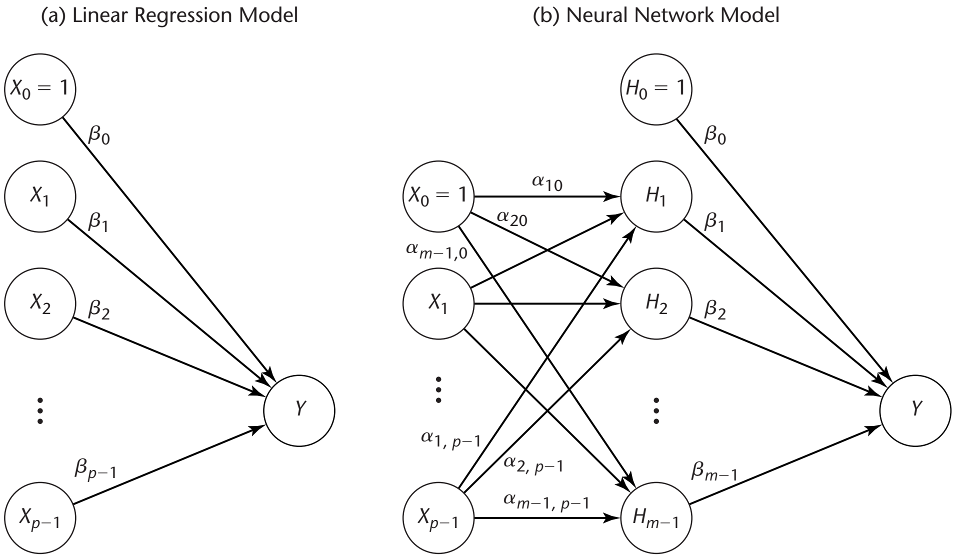
can be represented as a network as shown in Figure 13.8a. The link from each predictor X_i to the response is labeled with the corresponding regression parameter, β_i.

The feedforward, single-hidden-layer neural network model (13.45) is shown in Figure 13.8b. The predictor nodes are labeled $X_0, X_1, \ldots, X_{p-1}$ and are located on the left side of the diagram. In the center of the diagram are *m hidden nodes*. These nodes are linked to the p predictor nodes by relation (13.41); thus the links are labeled by using the α parameters. Finally, the hidden nodes are linked to the response Y by the β parameters.

Comments

1. Neural networks were first used as models for the human brain. The nodes represented neurons and the links between neurons represented synapses. A synapse would "fire" if the signal surpassed

FIGURE 13.8 **Network Representations of Linear Regression and Neural Network Models.**



a threshold. This suggested the use of step functions for the activation function, which were later replaced by smooth functions such as the logistic function.

2. The logistic activation function is sometimes replaced by a *radial basis function,* which is an n-dimensional normal probability density function. Details are provided in Reference 13.8. ■

Neural Network as Generalization of Linear Regression

It is easy to see that the standard multiple regression model is a special case of neural network model (13.45). If we choose for each of the activation functions $g_Y, g_1, \ldots, g_{m-1}$ the identity activation:

$$g(Z) = Z$$

we have:

$$E\{Y_i\} = \beta_0 + \beta_1 H_{i1} + \cdots + \beta_{m-1} H_{i,m-1} \tag{13.46a}$$

and:

$$H_{ij} = \alpha_{j0} + \alpha_{j1} X_{i1} + \cdots + \alpha_{j,p-1} X_{i,p-1} \tag{13.46b}$$

Substitution of (13.46b) into (13.46a) and rearranging yields:

$$\begin{aligned} E\{Y_i\} &= \left[\beta_0 + \sum_{j=1}^{m-1} \beta_j \alpha_{j0}\right] + \left[\sum_{j=1}^{m-1} \beta_j \alpha_{j1}\right] X_{i1} + \cdots + \left[\sum_{j=1}^{m-1} \beta_j \alpha_{j,p-1}\right] X_{i,p-1} \\ &= \beta_0^* + \beta_1^* X_{i1} + \cdots + \beta_{p-1}^* X_{i,p-1} \end{aligned} \tag{13.47}$$

where:

$$\begin{aligned} \beta_0^* &= \beta_0 + \sum_{j=1}^{m-1} \beta_j \alpha_{j0} \\ \beta_k^* &= \sum_{j=1}^{m-1} \beta_j \alpha_{jk} \qquad \text{for } k = 1, \ldots, p-1 \end{aligned} \tag{13.47a}$$

The neural network with identity activation functions thus reduces to the standard linear regression model.

There is a problem, however, with the interpretation of the neural network regression coefficients. If the regression function is given by $E\{Y_i\} = \beta_0^* + \beta_1^* X_{i1} + \cdots + \beta_p^* X_{i,p-1}$ as indicated in (13.47), then *any* set of neural network parameters satisfying the p equations in (13.47a) gives the correct model. Since there are many more neural network parameters than there are equations (or equivalently, β^* parameters) there are infinitely many sets of neural network parameters that lead to the correct model. Thus, any particular set of neural network parameters will have no intrinsic meaning in this case.

This overparameterization problem is somewhat reduced with the use of the logistic activation function in place of the identity function. Generally, however, if the number of hidden nodes is more than just a few, overparameterization will be present, and will lead to a fitted model with low predictive ability unless this issue is explicitly considered when the parameters are estimated. We now take up such estimation procedures.

Parameter Estimation: Penalized Least Squares

In Chapter 9 we considered model selection and validation. There, we observed that while R^2 never decreases with the addition of a new predictor, our ability to predict holdout responses in the validation stage can deteriorate if too many predictors are incorporated. Various model selection criteria, such as $R^2_{a,p}$, SBC_p, and AIC_p, have been adopted that contain penalties for the addition of predictors. We commented in Section 11.2 that ridge regression estimates can be obtained by the method of penalized least squares, which directly incorporates a penalty for the sum of squares of the regression coefficients. In order to control the level of overfitting, penalized least squares is frequently used for parameter estimation with neural networks.

The penalized least squares criterion is given by:

$$Q = \sum_{i=1}^{n} [Y_i - f(\mathbf{X}_i, \boldsymbol{\beta}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{m-1}]^2 + p_\lambda(\boldsymbol{\beta}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{m-1}) \qquad \textbf{(13.48)}$$

where the overfit penalty is:

$$p_\lambda(\boldsymbol{\beta}, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_{m-1}) = \lambda \left[\sum_{i=0}^{m-1} \beta_i^2 + \sum_{i=1}^{m-1} \sum_{j=0}^{p-1} \alpha_{ij}^2 \right] \qquad \textbf{(13.48a)}$$

Thus, the penalty is a positive constant, λ, times the sum of squares of the nonlinear regression coefficients. Note that the penalty is imposed not on the number of parameters $m + mp$, but on the total magnitude of the parameters. The *penalty weight* λ assigned to the regression coefficients governs the trade-off between overfitting and underfitting. If λ is large, the parameters estimates will be relatively small in absolute magnitude; if λ is small, the estimates will be relatively large. A "best" value for λ is generally between .001 and .1 and is chosen by cross-validation. For example, we may fit the model for a range of λ-values between .001 and .1, and choose the value that minimizes the total prediction error of the hold-out sample. The resulting parameter estimates are called shrinkage estimates because use of $\lambda > 0$ leads to reductions in their absolute magnitudes.

In Section 13.3 we described various search procedures, such as the Gauss-Newton method for finding nonlinear least squares estimates. Such methods can also be used with neural networks and penalized least squares criterion (13.48). We observed in Comment 1 on page 524, that the choice of starting values is important. Poor choice of starting values may lead to convergence to a local minimum (rather than the global minimum) when multiple minima exist. The problem of multiple minima is especially prevalent when fitting neural networks, due to the typically large numbers of parameters and the functional form of model (13.48). For this reason, it is common practice to fit the model many times (typically between 10 and 50 times) using different sets of randomly chosen starting values for each fit. The set of parameter estimates that leads to the lowest value of criterion function (13.48)—i.e., the best of the best—is chosen for further study. In the neural networks literature, finding a set of parameter values that minimize criterion (13.48) is referred to as *training the network*. The number of searches conducted before arriving at the final estimates is referred to as the number of *tours*.

Comment

Neural networks are often trained by a procedure called *back-propagation*. Back propagation is in fact the method of steepest descent, which can be very slow. Recommended methods include the *conjugate gradient* and *variable metric* methods. Reference 13.8 provides further details concerning back-propagation and other search procedures. ■

Example: Ischemic Heart Disease

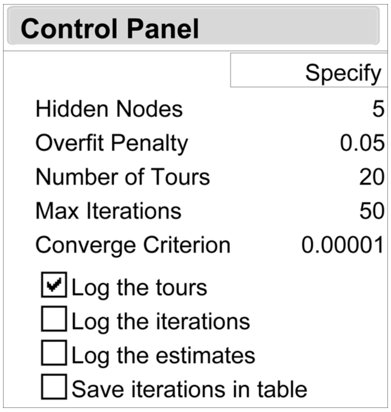
We illustrate the use of neural network model (13.44) and the penalized least squares fitting procedure using the Ischemic heart disease data set in Appendix C.9. These data were collected by a health insurance plan and provide information concerning 788 subscribers who made claims resulting from coronary heart disease. The response (Y) is the natural logarithm of the total cost of services provided and the predictors to be studied here are:

Predictor	Description
X_1:	Number of interventions, or procedures, carried out
X_2:	Number of tracked drugs used
X_3:	Number of comorbidities—other conditions present that complicate the treatment
X_4:	Number of complications—other conditions that arose during treatment due to heart disease

The first 400 observations are used to fit model (13.45) and the last $n^* = 388$ observations were held out for validation. (Note that the observations were originally sorted in a random order, so that the hold-out data set is a random sample.) We used JMP to fit and evaluate the neural network model.

Shown in Figure 13.9 is the JMP control panel, which allows the user to specify the various characteristics of the model and the fitting procedure. Here, we have chosen 5 hidden nodes, and we are using $\lambda = .05$ as the penalty weight. Also, we have chosen the default values for the number of tours (20), the maximum number of iterations for the search procedure

FIGURE 13.9 **JMP Control Panel for Neural Network Fit—Ischemic Heart Disease Example.**



Control Panel

	Specify
Hidden Nodes	5
Overfit Penalty	0.05
Number of Tours	20
Max Iterations	50
Converge Criterion	0.00001

☑ Log the tours
□ Log the iterations
□ Log the estimates
□ Save iterations in table

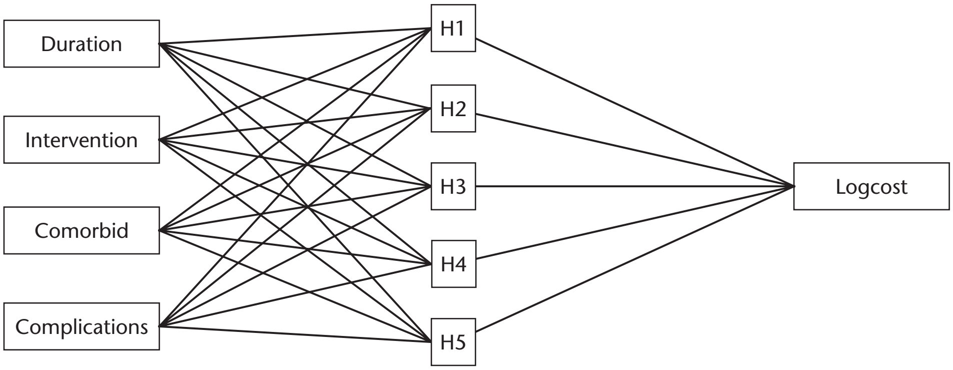
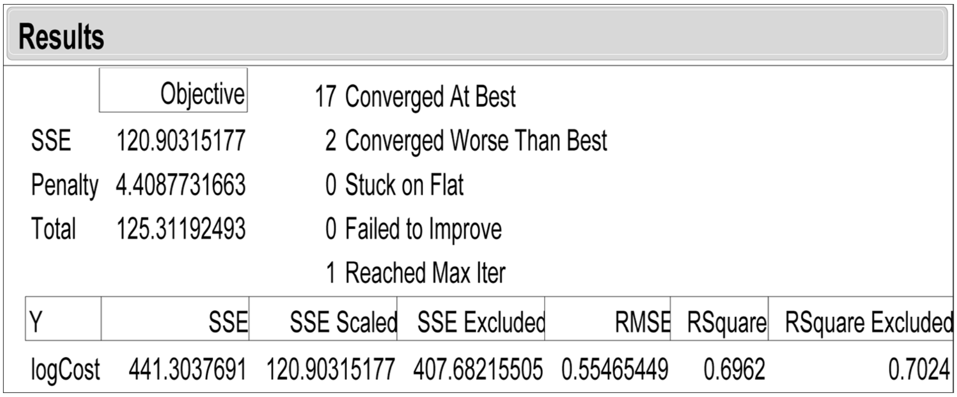


FIGURE 13.10 **JMP Neural Network Diagram—Ischemic Heart Disease Example.**



```
Results

          Objective      17 Converged At Best
SSE       120.90315177    2 Converged Worse Than Best
Penalty   4.4087731663    0 Stuck on Flat
Total     125.31192493    0 Failed to Improve
                          1 Reached Max Iter

Y                SSE  SSE Scaled  SSE Excluded       RMSE  RSquare  RSquare Excluded
logCost  441.3037691 120.90315177 407.68215505 0.55465449   0.6962            0.7024
```

FIGURE 13.11 **JMP Results for Neural Network Fit—Ischemic Heart Disease Example.**

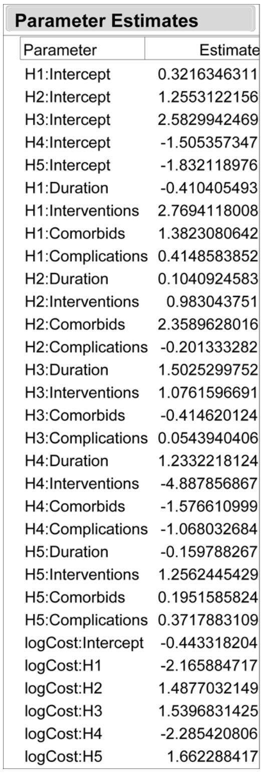
(50) and the convergence criterion (.00001). By checking the "log the tours" box, we will be keeping a record of the results of each of the 20 tours. A JMP network representation of model (13.45) is shown in Figure 13.10. Note that this representation excludes the constant nodes X_0 and H_0. In our notation, there are $m = 6$ hidden nodes and $p = 5$ predictor nodes, and it is necessary to estimate $m + p(m - 1) = 6 + 5(6 - 1) = 31$ parameters.

The results of the best fit, after 20 attempts or tours, is shown in Figure 13.11. The penalized least squares criterion value is 125.31. *SSE* for the scaled response is 120.90. JMP indicates that the corresponding *SSE* for the unscaled (original) responses is 441.30. The total prediction error for the validation (excluded) data, is given here by:

$$SSE_{VAL} = \sum_{i=401}^{788} (Y_i - \hat{Y}_i)^2 = 407.68$$

The mean squared prediction error (9.20) is obtained as $MSPR = SSE_{VAL}/n^* = 407.68/388 = 1.05$. JMP also gives R^2 for the training data (.6962), and for the validation data

FIGURE 13.12 **JMP Parameter Estimates for Neural Network Fit—Ischemic Heart Disease Example.**



Parameter Estimates

Parameter	Estimate
H1:Intercept	0.3216346311
H2:Intercept	1.2553122156
H3:Intercept	2.5829942469
H4:Intercept	-1.505357347
H5:Intercept	-1.832118976
H1:Duration	-0.410405493
H1:Interventions	2.7694118008
H1:Comorbids	1.3823080642
H1:Complications	0.4148583852
H2:Duration	0.1040924583
H2:Interventions	0.983043751
H2:Comorbids	2.3589628016
H2:Complications	-0.201333282
H3:Duration	1.5025299752
H3:Interventions	1.0761596691
H3:Comorbids	-0.414620124
H3:Complications	0.0543940406
H4:Duration	1.2332218124
H4:Interventions	-4.887856867
H4:Comorbids	-1.576610999
H4:Complications	-1.068032684
H5:Duration	-0.159788267
H5:Interventions	1.2562445429
H5:Comorbids	0.1951585824
H5:Complications	0.3717883109
logCost:Intercept	-0.443318204
logCost:H1	-2.165884717
logCost:H2	1.4877032149
logCost:H3	1.5396831425
logCost:H4	-2.285420806
logCost:H5	1.662288417

(.7024). This latter diagnostic was obtained using:

$$R^2_{VAL} = 1 - \frac{SSE_{VAL}}{SST_{VAL}}$$

where SST_{VAL} is the total sum of squares for the validation data. Because these R^2 values are approximately equal, we conclude that the use of weight penalty $\lambda = .05$ led to a good balance between underfitting and overfitting.

Figure 13.12 shows the 31 parameter estimates produced by JMP and the corresponding parameters. We display these values only for completeness–we make no attempt at interpretation. As noted earlier, our interest is centered on the prediction of future responses.

For comparison, two least squares regressions of Y on the four predictors X_1, X_2, X_3, and X_4 were also carried out. The first was based on a first-order model consisting of the four predictors and an intercept term; the second was based on a full second-order model consisting of an intercept plus the four linear terms, the four quadratic terms, and the six cross-products among the four predictors. The results for these two multiple regression models and the neural network model are summarized in the Table 13.6.

From the results, we see that the neural network model's ability to predict holdout responses is superior to the first-order multiple regression and slightly better that the second-order multiple regression model. *MSPR* for the neural network is 1.05, whereas this statistic for the first and second-order multiple regression models is 1.28 and 1.09, respectively.

TABLE 13.6 **Comparisons of Results for Neural Network Model with Multiple Linear Regression Model—Ischemic Heart Disease Example.**

	Neural Network	Multiple Linear Regression	
		First-Order	Second-Order
Number of Parameters	31	5	15
MSE	1.20	1.74	1.34
MSPR	1.05	1.28	1.09

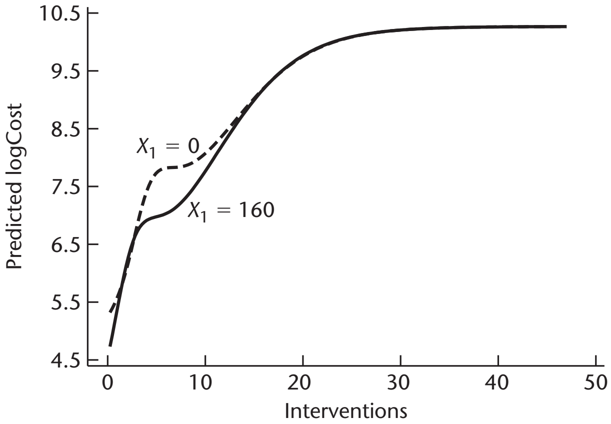


FIGURE 13.13 **Conditional Effects Plot—Ischemic Heart Disease Example.**

Model Interpretation and Prediction

While individual parameters and derived predictors are usually not interpretable, some understanding of the effects of individual predictors can be realized through the use of conditional effects plots. For example, Figure 13.13 shows for the ischemic heart data example, plots of predicted response as a function the number of interventions (X_2) for duration (X_1) equal to 0 and 160. The remaining predictors, comorbidities ($X_3 = 3.55$) and complications ($X_4 = 0.05$), are fixed at their averages for values in the training set. The plot indicates that the natural logarithm of cost increases rapidly as the number of interventions increases from 0 to 25, and then reaches a plateau and is stable as the number of interventions increases from 25 to 50. The duration variable seems to have very little effect, except possibly when interventions are between 5 and 10.

We have noted that neural network models can be very effective tools for prediction when large data sets are available. As always, it is important that the uncertainty in any prediction be quantified. Methods for producing approximate confidence intervals for estimation and prediction have been developed and some packages such as JMP now provide these intervals. Details are provided in Reference 13.9.

Some Final Comments on Neural Network Modeling

In recent years, neural networks have found widespread application in many fields. Indeed, they have become one of the standard tools in the field of data mining, and their use continues to grow. This is due largely to the widespread availability of powerful computers that permit the fitting of complex models having dozens, hundreds, and even thousands, of parameters.

A vocabulary has developed that is unique to the field of neural networks. The table below (adapted from Ref. 13.10) lists a number of terms that are commonly used by statisticians and their neural network equivalents:

Statistical Term	Neural Network Term
coefficient	weight
predictor	input
response	output
observation	exemplar
parameter estimation	training or learning
steepest descent	back-propagation
intercept	bias term
derived predictor	hidden node
penalty function	weight decay

There are a number of advantages to the neural network modeling approach. These include:

1. Model (13.45) is extremely flexible, and can be used to represent a wide range of response surface shapes. For example, with sufficient data, curvatures, interactions, plateaus, and step functions can be effectively modeled.
2. Standard regression assumptions, such as the requirements that the true residuals are mutually independent, normally distributed, and have constant variance, are not required for neural network modeling.
3. Outliers in the response and predictors can still have a detrimental effect on the fit of the model, but the use of the bounded logistic activation function tends to limit the influence of individual cases in comparison with standard regression approaches.

Of course, there are disadvantages associated with the use of neural networks. Model parameters are generally uninterpretable, and the method depends on the availability of large data sets. Diagnostics, such as lack of fit tests, identification of influential observations and outliers, and significance testing for the effects of the various predictors, are currently not generally available.

Cited References

13.1. Hartley, H. O. “The Modified Gauss-Newton Method for the Fitting of Non-linear Regression Functions by Least Squares,” *Technometrics* 3 (1961), pp. 269–80.

13.2. Gallant, A. R. *Nonlinear Statistical Models*. New York: John Wiley & Sons, 1987.

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

13.4. Bates, D. M., and D. G. Watts. *Nonlinear Regression Analysis and Its Applications*. New York: John Wiley & Sons, 1988.

13.5. Box, M. J. “Bias in Nonlinear Estimation,” *Journal of the Royal Statistical Society B* 33 (1971), pp. 171–201.

13.6. Hougaard, P. “The Appropriateness of the Asymptotic Distribution in a Nonlinear Regression Model in Relation to Curvature,” *Journal of the Royal Statistical Society B* 47 (1985), pp. 103–14.

13.7. Ratkowsky, D. A. *Nonlinear Regression Modeling*. New York: Marcel Dekker, 1983.

13.8. Hastie, T., Tibshirani, R., and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, 2001.

13.9. DeVeaux, R. D., Schumi, J., Schweinsberg, J., and L. H. Ungar. “Prediction Intervals for Neural Networks via Nonlinear Regression,” *Technometrics* 40 (1998), pp. 273–82.

13.10. DeVeaux, R. D., and L. H. Ungar. “A Brief Introduction to Neural Networks,” www.williams.edu/mathematics/rdeveaux/pubs.html (1996).

Problems

*13.1. For each of the following response functions, indicate whether it is a linear response function, an intrinsically linear response function, or a nonlinear response function. In the case of an intrinsically linear response function, state how it can be linearized by a suitable transformation:

a. $f(\mathbf{X}, \boldsymbol{\gamma}) = \exp(\gamma_0 + \gamma_1 X)$

b. $f(\mathbf{X}, \boldsymbol{\gamma}) = \gamma_0 + \gamma_1(\gamma_2)^{X_1} - \gamma_3 X_2$

c. $f(\mathbf{X}, \boldsymbol{\gamma}) = \gamma_0 + \dfrac{\gamma_1}{\gamma_0} X$

13.2. For each of the following response functions, indicate whether it is a linear response function, an intrinsically linear response function, or a nonlinear response function. In the case of an intrinsically linear response function, state how it can be linearized by a suitable transformation:

a. $f(\mathbf{X}, \boldsymbol{\gamma}) = \exp(\gamma_0 + \gamma_1 \log_e X)$

b. $f(\mathbf{X}, \boldsymbol{\gamma}) = \gamma_0(X_1)^{\gamma_1}(X_2)^{\gamma_2}$

c. $f(\mathbf{X}, \boldsymbol{\gamma}) = \gamma_0 - \gamma_1(\gamma_2)^{X}$

*13.3. a. Plot the logistic response function:

$$f(\mathbf{X}, \boldsymbol{\gamma}) = \frac{300}{1 + (30)\exp(-1.5X)} \qquad X \geq 0$$

b. What is the asymptote of this response function? For what value of X does the response function reach 90 percent of its asymptote?

13.4. a. Plot the exponential response function:

$$f(\mathbf{X}, \boldsymbol{\gamma}) = 49 - (30)\exp(-1.1X) \qquad X \geq 0$$

b. What is the asymptote of this response function? For what value of X does the response function reach 95 percent of its asymptote?

*13.5. **Home computers.** A computer manufacturer hired a market research firm to investigate the relationship between the likelihood a family will purchase a home computer and the price of the home computer. The data that follow are based on replicate surveys done in two similar cities. One thousand heads of households in each city were randomly selected and asked if they would be likely to purchase a home computer at a given price. Eight prices (X, in dollars) were studied, and 100 heads of households in each city were randomly assigned to a given price. The proportion likely to purchase at a given price is denoted by Y.

City A

***i*:**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
X_i:	200	400	800	1200	1600	2000	3000	4000
Y_i:	.65	.46	.34	.26	.17	.15	.06	.04

City B

***i*:**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
X_i:	200	400	800	1200	1600	2000	3000	4000
Y_i:	.63	.50	.30	.24	.19	.12	.08	.05

No location effect is expected and the data are to be treated as independent replicates at each of the 8 prices. The following exponential model with independent normal error terms is deemed to be appropriate:

$$Y_i = \gamma_0 + \gamma_2 \exp(-\gamma_1 X_i) + \varepsilon_i$$

a. To obtain initial estimates of γ_0, γ_1, and γ_2, note that $f(\mathbf{X}, \boldsymbol{\gamma})$ approaches a lower asymptote γ_0 as X increases without bound. Hence, let $g_0^{(0)} = 0$ and observe that when we ignore the error term, a logarithmic transformation then yields $Y_i' = \beta_0 + \beta_1 X_i$, where $Y_i' = \log_e Y_i$, $\beta_0 = \log_e \gamma_2$, and $\beta_1 = -\gamma_1$. Therefore, fit a linear regression function based on the transformed data and use as initial estimates $g_0^{(0)} = 0$, $g_1^{(0)} = -b_1$, and $g_2^{(0)} = \exp(b_0)$.

b. Using the starting values obtained in part (a), find the least squares estimates of the parameters γ_0, γ_1, and γ_2.

*13.6. Refer to **Home computers** Problem 13.5.

a. Plot the estimated nonlinear regression function and the data. Does the fit appear to be adequate?

b. Obtain the residuals and plot them against the fitted values and against X on separate graphs. Also obtain a normal probability plot. Does the model appear to be adequate?

*13.7. Refer to **Home computers** Problem 13.5. Assume that large-sample inferences are appropriate here. Conduct a formal approximate test for lack of fit of the nonlinear regression function; use $\alpha = .01$. State the alternatives, decision rule, and conclusion.

*13.8. Refer to **Home computers** Problem 13.5. Assume that the fitted model is appropriate and that large-sample inferences can be employed. Obtain approximate joint confidence intervals for the parameters γ_0, γ_1, and γ_2, using the Bonferroni procedure and a 90 percent family confidence coefficient.

*13.9. Refer to **Home computers** Problem 13.5. A question has been raised whether the two cities are similar enough so that the data can be considered to be replicates. Adding a location effect parameter analogous to (13.38) to the model proposed in Problem 13.5 yields the four-parameter nonlinear regression model:

$$Y_i = \gamma_0 + \gamma_3 X_{i2} + \gamma_2 \exp(-\gamma_1 X_{i1}) + \varepsilon_i$$

where:

$$X_2 = \begin{cases} 0 & \text{if city A} \\ 1 & \text{if city B} \end{cases}$$

a. Using the same starting values as those obtained in Problem 13.5a and $g_3^{(0)} = 0$, find the least squares estimates of the parameters γ_0, γ_1, γ_2, and γ_3.

b. Assume that large-sample inferences can be employed reasonably here. Obtain an approximate 95 percent confidence interval for γ_3. What does this interval indicate about city differences? Is this result consistent with your conclusion in Problem 13.7? Does it have to be? Discuss.

13.10. **Enzyme kinetics.** In an enzyme kinetics study the velocity of a reaction (Y) is expected to be related to the concentration (X) as follows:

$$Y_i = \frac{\gamma_0 X_i}{\gamma_1 + X_i} + \varepsilon_i$$

Eighteen concentrations have been studied and the results follow:

i:	1	2	3	...	16	17	18
X_i:	1	1.5	2	...	30	35	40
Y_i:	2.1	2.5	4.9	...	19.7	21.3	21.6

a. To obtain starting values for γ_0 and γ_1, observe that when the error term is ignored we have $Y_i' = \beta_0 + \beta_1 X_i'$, where $Y_i' = 1/Y_i$, $\beta_0 = 1/\gamma_0$, $\beta_1 = \gamma_1/\gamma_0$, and $X_i' = 1/X_i$. Therefore fit a linear regression function to the transformed data to obtain initial estimates $g_0^{(0)} = 1/b_0$ and $g_1^{(0)} = b_1/b_0$.

b. Using the starting values obtained in part (a), find the least squares estimates of the parameters γ_0 and γ_1.

13.11. Refer to **Enzyme kinetics** Problem 13.10.

a. Plot the estimated nonlinear regression function and the data. Does the fit appear to be adequate?

b. Obtain the residuals and plot them against the fitted values and against X on separate graphs. Also obtain a normal probability plot. What do your plots show?

c. Can you conduct an approximate formal lack of fit test here? Explain.

d. Given that only 18 trials can be made, what are some advantages and disadvantages of considering fewer concentration levels but with some replications, as compared to considering 18 different concentration levels as was done here?

13.12. Refer to **Enzyme kinetics** Problem 13.10. Assume that the fitted model is appropriate and that large-sample inferences can be employed here. (1) Obtain an approximate 95 percent confidence interval for γ_0. (2) Test whether or not $\gamma_1 = 20$; use $\alpha = .05$. State the alternatives, decision rule, and conclusion.

*13.13. **Drug responsiveness.** A pharmacologist modeled the responsiveness to a drug using the following nonlinear regression model:

$$Y_i = \gamma_0 - \frac{\gamma_0}{1 + \left(\dfrac{X_i}{\gamma_2}\right)^{\gamma_1}} + \varepsilon_i$$

X denotes the dose level, in coded form, and Y the responsiveness expressed as a percent of the maximum possible responsiveness. In the model, γ_0 is the expected response at saturation, γ_2 is the concentration that produces a half-maximal response, and γ_1 is related to the slope. The data for 19 cases at 13 dose levels follow:

i:	1	2	3	...	17	18	19
X_i:	1	2	3	...	7	8	9
Y_i:	.5	2.3	3.4	...	94.8	96.2	96.4

Obtain least squares estimates of the parameters γ_0, γ_1, and γ_2, using starting values $g_0^{(0)} = 100$, $g_1^{(0)} = 5$, and $g_2^{(0)} = 4.8$.

*13.14. Refer to **Drug responsiveness** Problem 13.13.

a. Plot the estimated nonlinear regression function and the data. Does the fit appear to be adequate?

b. Obtain the residuals and plot them against the fitted values and against X on separate graphs. Also obtain a normal probability plot. What do your plots show about the adequacy of the regression model?

*13.15. Refer to **Drug responsiveness** Problem 13.13. Assume that large-sample inferences are appropriate here. Conduct a formal approximate test for lack of fit of the nonlinear regression function; use $\alpha = .01$. State the alternatives, decision rule, and conclusion.

*13.16. Refer to **Drug responsiveness** Problem 13.13. Assume that the fitted model is appropriate and that large-sample inferences can be employed here. Obtain approximate joint confidence intervals for the parameters γ_0, γ_1, and γ_2 using the Bonferroni procedure with a 91 percent family confidence coefficient. Interpret your results.

13.17. **Process yield.** The yield (Y) of a chemical process depends on the temperature (X_1) and pressure (X_2). The following nonlinear regression model is expected to be applicable:

$$Y_i = \gamma_0(X_{i1})^{\gamma_1}(X_{i2})^{\gamma_2} + \varepsilon_i$$

Prior to beginning full-scale production, 18 tests were undertaken to study the process yield for various temperature and pressure combinations. The results follow.

i:	1	2	3	...	16	17	18
X_{i1}:	1	10	100	...	1	10	100
X_{i2}:	1	1	1	...	100	100	100
Y_i:	12	32	103	...	43	128	398

a. To obtain starting values for γ_0, γ_1, and γ_2, note that when we ignore the random error term, a logarithmic transformation yields $Y_i' = \beta_0 + \beta_1 X_{i1}' + \beta_1 X_{i2}'$, where $Y_i' = \log_{10} Y_i$, $\beta_0 = \log_{10}\gamma_0$, $\beta_1 = \gamma_1$, $X_{i1}' = \log_{10} X_{i1}$, $\beta_2 = \gamma_2$, and $X_{i2}' = \log_{10} X_{i2}$. Fit a first-order multiple regression model to the transformed data, and use as starting values $g_0^{(0)} = \text{antilog}_{10}\, b_0$, $g_1^{(0)} = b_1$, and $g_2^{(0)} = b_2$.

b. Using the starting values obtained in part (a), find the least squares estimates of the parameters γ_0, γ_1, and γ_2.

13.18. Refer to **Process yield** Problem 13.17.

a. Plot the estimated nonlinear regression function and the data. Does the fit appear to be adequate?

b. Obtain the residuals and plot them against $\hat{Y}$, X_1, and X_2 on separate graphs. Also obtain a normal probability plot. What do your plots show about the adequacy of the model?

13.19. Refer to **Process yield** Problem 13.17. Assume that large-sample inferences are appropriate here. Conduct a formal approximate test for lack of fit of the nonlinear regression function; use $\alpha = .05$. State the alternatives, decision rule, and conclusion.

13.20. Refer to **Process yield** Problem 13.17. Assume that the fitted model is appropriate and that large-sample inferences are applicable here.

a. Test the hypotheses H_0: $\gamma_1 = \gamma_2$ against H_a: $\gamma_1 \neq \gamma_2$ using $\alpha = .05$. State the alternatives, decision rule, and conclusion.

b. Obtain approximate joint confidence intervals for the parameters γ_1 and γ_2, using the Bonferroni procedure and a 95 percent family confidence coefficient.

c. What do you conclude about the parameters γ_1 and γ_2 based on the results in parts (a) and (b)?

Exercises

13.21. (Calculus needed.) Refer to **Home computers** Problem 13.5.

a. Obtain the least squares normal equations and show that they are nonlinear in the estimated regression coefficients g_0, g_1, and g_2.

b. State the likelihood function for the nonlinear regression model, assuming that the error terms are independent $N(0, \sigma^2)$.

13.22. (Calculus needed.) Refer to **Enzyme kinetics** Problem 13.10.

a. Obtain the least squares normal equations and show that they are nonlinear in the estimated regression coefficients g_0 and g_1.

b. State the likelihood function for the nonlinear regression model, assuming that the error terms are independent $N(0, \sigma^2)$.

13.23. (Calculus needed.) Refer to **Process yield** Problem 13.17.

a. Obtain the least squares normal equations and show that they are nonlinear in the estimated regression coefficients g_0, g_1, and g_2.

b. State the likelihood function for the nonlinear regression model, assuming that the error terms are independent $N(0, \sigma^2)$.

13.24. Refer to **Drug responsiveness** Problem 13.13.

a. Assuming that $E\{\varepsilon_i\} = 0$, show that:

$$E\{Y\} = \gamma_0\left(\frac{A}{1 + A}\right)$$

where:

$$A = \exp[\gamma_1(\log_e X - \log_e \gamma_2)] = \exp(\beta_0 + \beta_1 X')$$

and $\beta_0 = -\gamma_1 \log_e \gamma_2$, $\beta_1 = \gamma_1$, and $X' = \log_e X$.

b. Assuming γ_0 is known, show that:

$$\frac{E\{Y'\}}{1 - E\{Y'\}} = \exp(\beta_0 + \beta_1 X')$$

where $Y' = Y/\gamma_0$.

c. What transformation do these results suggest for obtaining a simple linear regression function in the transformed variables?

d. How can starting values for finding the least squares estimates of the nonlinear regression parameters be obtained from the estimates of the linear regression coefficients?

Projects

13.25. Refer to **Enzyme kinetics** Problem 13.10. Starting values for finding the least squares estimates of the nonlinear regression model parameters are to be obtained by a grid search. The following bounds for the two parameters have been specified:

$$5 \leq \gamma_0 \leq 65$$
$$5 \leq \gamma_1 \leq 65$$

Obtain 49 grid points by using all possible combinations of the boundary values and five other equally spaced points for each parameter range. Evaluate the least squares criterion (13.15) for each grid point and identify the point providing the best fit. Does this point give reasonable starting values here?

13.26. Refer to **Process yield** Problem 13.17. Starting values for finding the least squares estimates of the nonlinear regression model parameters are to be obtained by a grid search. The following bounds for the parameters have been postulated:

$$1 \leq \gamma_0 \leq 21$$
$$.2 \leq \gamma_1 \leq .8$$
$$.1 \leq \gamma_2 \leq .7$$

Obtain 27 grid points by using all possible combinations of the boundary values and the midpoint for each of the parameter ranges. Evaluate the least squares criterion (13.15) for each grid point and identify the point providing the best fit. Does this point give reasonable starting values here?

13.27. Refer to **Home computers** Problem 13.5.

a. To check on the appropriateness of large-sample inferences here, generate 1,000 bootstrap samples of size 16 using the fixed X sampling procedure. For each bootstrap sample, obtain the least squares estimates g_0^*, g_1^*, and g_2^*.

b. Plot histograms of the bootstrap sampling distributions of g_0^*, g_1^*, and g_2^*. Do these distributions appear to be approximately normal?

c. Compute the means and standard deviations of the bootstrap sampling distributions for g_0^*, g_1^*, and g_2^*. Are the bootstrap means and standard deviations close to the final least squares estimates?

d. Obtain a confidence interval for γ_1 using the reflection method in (11.59) and confidence coefficient .9667. How does this interval compare with the one obtained in Problem 13.8 by the large-sample inference method?

e. What are the implications of your findings in parts (b), (c), and (d) about the appropriateness of large-sample inferences here? Discuss.

13.28. Refer to **Enzyme kinetics** Problem 13.10.

a. To check on the appropriateness of large-sample inferences here, generate 1,000 bootstrap samples of size 18 using the fixed X sampling procedure. For each bootstrap sample, obtain the least squares estimates g_0^* and g_1^*.

b. Plot histograms of the bootstrap sampling distributions of g_0^* and g_1^*. Do these distributions appear to be approximately normal?

c. Compute the means and standard deviations of the bootstrap sampling distributions for g_0^* and g_1^*. Are the bootstrap means and standard deviations close to the final least squares estimates?

d. Obtain a confidence interval for γ_0 using the reflection method in (11.59) and confidence coefficient .95. How does this interval compare with the one obtained in Problem 13.12 by the large-sample inference method?

e. What are the implications of your findings in parts (b), (c), and (d) about the appropriateness of large-sample inferences here? Discuss.

13.29. Refer to **Drug responsiveness** Problem 13.13.

a. To check on the appropriateness of large-sample inferences here, generate 1,000 bootstrap samples of size 19 using the fixed X sampling procedure. For each bootstrap sample, obtain the least squares estimates g_0^*, g_1^*, and g_2^*.

b. Plot histograms of the bootstrap sampling distributions of g_0^*, g_1^*, and g_2^*. Do these distributions appear to be approximately normal?

c. Compute the means and standard deviations of the bootstrap sampling distributions for g_0^*, g_1^*, and g_2^*. Are the bootstrap means and standard deviations close to the final least squares estimates?

d. Obtain a confidence interval for γ_2 using the reflection method in (11.59) and confidence coefficient .97. How does this interval compare with the one obtained in Problem 13.16 by the large-sample inference method?

e. What are the implications of your findings in parts (b), (c), and (d) about the appropriateness of large-sample inferences here? Discuss.

13.30. Refer to **Process yield** Problem 13.17.

a. To check on the appropriateness of large-sample inferences here, generate 1,000 bootstrap samples of size 18 using the fixed X sampling procedure. For each bootstrap sample, obtain the least squares estimates g_0^*, g_1^*, and g_2^*.

b. Plot histograms of the bootstrap sampling distributions of g_0^*, g_1^*, and g_2^*. Do these distributions appear to be approximately normal?

c. Compute the means and standard deviations of the bootstrap sampling distributions for g_0^*, g_1^*, and g_2^*. Are the bootstrap means and standard deviations close to the final least squares estimates?

d. Obtain a confidence interval for γ_1 using the reflection method in (11.59) and confidence coefficient .975. How does this interval compare with the one obtained in Problem 13.20b by the large-sample inference method?

e. What are the implications of your findings in parts (b), (c), and (d) about the appropriateness of large-sample inferences here? Discuss.

Case Studies

13.31. Refer to the **Prostate cancer** data set in Appendix C.5 and Case Study 9.30. Select a random sample of 65 observations to use as the model-building data set.

a. Develop a neural network model for predicting PSA. Justify your choice of number of hidden nodes and penalty function weight and interpret your model.

b. Assess your model's ability to predict and discuss its usefulness to the oncologists.

c. Compare the performance of your neural network model with that of the best regression model obtained in Case Study 9.30. Which model is more easily interpreted and why?

13.32. Refer to the **Real estate sales** data set in Appendix C.7 and Case Study 9.31. Select a random sample of 300 observations to use as the model-building data set.

a. Develop a neural network model for predicting sales price. Justify your choice of number of hidden nodes and penalty function weight and interpret your model.

b. Assess your model's ability to predict and discuss its usefulness as a tool for predicting sales prices.

c. Compare the performance of your neural network model with that of the best regression model obtained in Case Study 9.31. Which model is more easily interpreted and why?