

## 7

# NONLINEAR, SEMIPARAMETRIC AND NONPARAMETRIC REGRESSION MODELS<sup>1</sup>



## 7.1 INTRODUCTION

Up to this point, the focus has been on a **linear regression model**

$$y = x_1\beta_1 + x_2\beta_2 + \cdots + \varepsilon. \quad (7-1)$$

Chapters 2 to 5 developed the least squares method of estimating the parameters and obtained the statistical properties of the estimator that provided the tools we used for point and interval estimation, hypothesis testing, and prediction. The modifications suggested in Chapter 6 provided a somewhat more general form of the linear regression model,

$$y = f_1(\mathbf{x})\beta_1 + f_2(\mathbf{x})\beta_2 + \cdots + \varepsilon. \quad (7-2)$$

By the definition we want to use in this chapter, this model is still “linear,” because the parameters appear in a linear form. Section 7.2 of this chapter will examine the **nonlinear regression model** (which includes (7-1) and (7-2) as special cases),

$$y = h(x_1, x_2, \dots, x_P; \beta_1, \beta_2, \dots, \beta_K) + \varepsilon, \quad (7-3)$$

where the conditional mean function involves  $P$  variables and  $K$  parameters. This form of the model changes the conditional mean function from  $E[y|\mathbf{x}, \boldsymbol{\beta}] = \mathbf{x}'\boldsymbol{\beta}$  to  $E[y|\mathbf{x}] = h(\mathbf{x}, \boldsymbol{\beta})$  for more general functions. This allows a much wider range of functional forms than the linear model can accommodate.<sup>2</sup> This change in the model form will require us to develop an alternative method of estimation, **nonlinear least squares**. We will also examine more closely the interpretation of parameters in nonlinear models. In particular, since  $\partial E[y|\mathbf{x}]/\partial \mathbf{x}$  is no longer equal to  $\boldsymbol{\beta}$ , we will want to examine how  $\boldsymbol{\beta}$  should be interpreted.

Linear and nonlinear least squares are used to estimate the parameters of the **conditional mean function**,  $E[y|\mathbf{x}]$ . As we saw in Example 4.5, other relationships between  $y$  and  $\mathbf{x}$ , such as the **conditional median**, might be of interest. Section 7.3 revisits this idea with an examination of the conditional median function and the least absolute

<sup>1</sup>This chapter covers some fairly advanced features of regression modeling and numerical analysis. It may be bypassed in a first course without loss of continuity.

<sup>2</sup>A complete discussion of this subject can be found in Amemiya (1985). Other important references are Jennrich (1969), Malinvaud (1970), and especially Goldfeld and Quandt (1971, 1972). A very lengthy authoritative treatment is the text by Davidson and MacKinnon (1993).

## 182 PART I ♦ The Linear Regression Model

deviations estimator. This section will also relax the restriction that the model coefficients are always the same in the different parts of the distribution of  $y$  (given  $\mathbf{x}$ ). The LAD estimator estimates the parameters of the conditional median, that is, 50<sup>th</sup> percentile function. The **quantile regression model** allows the parameters of the regression to change as we analyze different parts of the conditional distribution.

The model forms considered thus far are semiparametric in nature, and less parametric as we move from Section 7.2 to 7.3. The **partially linear regression** examined in Section 7.4 extends (7-1) such that  $y = f(x) + \mathbf{z}'\boldsymbol{\beta} + \varepsilon$ . The endpoint of this progression is a model in which the relationship between  $y$  and  $x$  is not forced to conform to a particular parameterized function. Using largely graphical and kernel density methods, we consider in Section 7.5 how to analyze a **nonparametric regression** relationship that essentially imposes little more than  $E[y|\mathbf{x}] = h(\mathbf{x})$ .

## 7.2 NONLINEAR REGRESSION MODELS

The general form of the nonlinear regression model is

$$y_i = h(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i. \quad (7-4)$$

The linear model is obviously a special case. Moreover, some models which appear to be nonlinear, such as

$$y = e^{\beta_1} x_1^{\beta_2} x_2^{\beta_3} e^{\varepsilon},$$

become linear after a transformation, in this case after taking logarithms. In this chapter, we are interested in models for which there is no such transformation, such as the one in the following example.

### Example 7.1 CES Production Function

In Example 6.8, we examined a constant elasticity of substitution production function model:

$$\ln y = \ln \gamma - \frac{\nu}{\rho} \ln [\delta K^{-\rho} + (1 - \delta) L^{-\rho}] + \varepsilon. \quad (7-5)$$

No transformation reduces this equation to one that is linear in the parameters. In Example 6.5, a linear Taylor series approximation to this function around the point  $\rho = 0$  is used to produce an intrinsically linear equation that can be fit by least squares. Nonetheless, the underlying model in (7.5) is nonlinear in the sense that interests us in this chapter.

This and the next section will extend the assumptions of the linear regression model to accommodate nonlinear functional forms such as the one in Example 7.1. We will then develop the nonlinear least squares estimator, establish its statistical properties, and then consider how to use the estimator for hypothesis testing and analysis of the model predictions.

### 7.2.1 ASSUMPTIONS OF THE NONLINEAR REGRESSION MODEL

We shall require a somewhat more formal definition of a nonlinear regression model. Sufficient for our purposes will be the following, which include the linear model as the special case noted earlier. We assume that there is an underlying probability distribution, or data generating process (DGP) for the observable  $y_i$  and a true parameter vector,  $\boldsymbol{\beta}$ ,

which is a characteristic of that DGP. The following are the assumptions of the nonlinear regression model:

1. **Functional form:** The conditional mean function for  $y_i$  given  $\mathbf{x}_i$  is

$$E[y_i | \mathbf{x}_i] = h(\mathbf{x}_i, \boldsymbol{\beta}), \quad i = 1, \dots, n,$$

where  $h(\mathbf{x}_i, \boldsymbol{\beta})$  is a continuously differentiable function of  $\boldsymbol{\beta}$ .

2. **Identifiability of the model parameters:** The parameter vector in the model is identified (estimable) if there is no nonzero parameter  $\boldsymbol{\beta}^0 \neq \boldsymbol{\beta}$  such that  $h(\mathbf{x}_i, \boldsymbol{\beta}^0) = h(\mathbf{x}_i, \boldsymbol{\beta})$  for all  $\mathbf{x}_i$ . In the linear model, this was the full rank assumption, but the simple absence of “multicollinearity” among the variables in  $\mathbf{x}$  is not sufficient to produce this condition in the nonlinear regression model. Example 7.2 illustrates the problem.
3. **Zero mean of the disturbance:** It follows from Assumption 1 that we may write

$$y_i = h(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i.$$

where  $E[\varepsilon_i | h(\mathbf{x}_i, \boldsymbol{\beta})] = 0$ . This states that the disturbance at observation  $i$  is uncorrelated with the conditional mean function for all observations in the sample. This is not quite the same as assuming that the disturbances and the exogenous variables are uncorrelated, which is the familiar assumption, however.

4. **Homoscedasticity and nonautocorrelation:** As in the linear model, we assume conditional homoscedasticity,

$$E[\varepsilon_i^2 | h(\mathbf{x}_j, \boldsymbol{\beta}), j = 1, \dots, n] = \sigma^2, \quad \text{a finite constant,} \quad (7-6)$$

and nonautocorrelation

$$E[\varepsilon_i \varepsilon_j | h(\mathbf{x}_i, \boldsymbol{\beta}), h(\mathbf{x}_j, \boldsymbol{\beta}), j = 1, \dots, n] = 0 \quad \text{for all } j \neq i.$$

5. **Data generating process:** The data-generating process for  $\mathbf{x}_i$  is assumed to be a well-behaved population such that first and second moments of the data can be assumed to converge to fixed, finite population counterparts. The crucial assumption is that the process generating  $\mathbf{x}_i$  is strictly exogenous to that generating  $\varepsilon_i$ . The data on  $\mathbf{x}_i$  are assumed to be “well behaved.”
6. **Underlying probability model:** There is a well-defined probability distribution generating  $\varepsilon_i$ . At this point, we assume only that this process produces a sample of uncorrelated, identically (marginally) distributed random variables  $\varepsilon_i$  with mean zero and variance  $\sigma^2$  conditioned on  $h(\mathbf{x}_i, \boldsymbol{\beta})$ . Thus, at this point, our statement of the model is **semiparametric**. (See Section 12.3.) We will not be assuming any particular distribution for  $\varepsilon_i$ . The conditional moment assumptions in 3 and 4 will be sufficient for the results in this chapter. In Chapter 14, we will fully parameterize the model by assuming that the disturbances are normally distributed. This will allow us to be more specific about certain test statistics and, in addition, allow some generalizations of the regression model. The assumption is not necessary here.

**Example 7.2 Identification in a Translog Demand System**

Christensen, Jorgenson, and Lau (1975), proposed the translog **indirect utility function** for a consumer allocating a budget among  $K$  commodities:

$$\ln V = \beta_0 + \sum_{k=1}^K \beta_k \ln(p_k/M) + \sum_{k=1}^K \sum_{j=1}^K \gamma_{kj} \ln(p_k/M) \ln(p_j/M),$$

## 184 PART I ♦ The Linear Regression Model

where  $V$  is indirect utility,  $p_k$  is the price for the  $k$ th commodity, and  $M$  is income. Utility, direct or indirect, is unobservable, so the utility function is not usable as an empirical model. **Roy's identity** applied to this logarithmic function produces a budget share equation for the  $k$ th commodity that is of the form

$$S_k = -\frac{\partial \ln V / \partial \ln p_k}{\partial \ln V / \partial \ln M} = \frac{\beta_k + \sum_{j=1}^K \gamma_{kj} \ln(p_j/M)}{\beta_M + \sum_{j=1}^K \gamma_{Mj} \ln(p_j/M)} + \varepsilon, k = 1, \dots, K,$$

where  $\beta_M = \sum_k \beta_k$  and  $\gamma_{Mj} = \sum_k \gamma_{kj}$ . No transformation of the budget share equation produces a linear model. This is an intrinsically nonlinear regression model. (It is also one among a system of equations, an aspect we will ignore for the present.) Although the share equation is stated in terms of observable variables, it remains unusable as an empirical model because of an **identification problem**. If every parameter in the budget share is multiplied by the same constant, then the constant appearing in both numerator and denominator cancels out, and the same value of the function in the equation remains. The indeterminacy is resolved by imposing the normalization  $\beta_M = 1$ . Note that this sort of identification problem does not arise in the linear model.

### 7.2.2 THE NONLINEAR LEAST SQUARES ESTIMATOR

The nonlinear least squares estimator is defined as the minimizer of the sum of squares,

$$S(\boldsymbol{\beta}) = \frac{1}{2} \sum_{i=1}^n \varepsilon_i^2 = \frac{1}{2} \sum_{i=1}^n [y_i - h(\mathbf{x}_i, \boldsymbol{\beta})]^2. \quad (7-7)$$

The first order conditions for the minimization are

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n [y_i - h(\mathbf{x}_i, \boldsymbol{\beta})] \frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{0}. \quad (7-8)$$

In the linear model, the vector of partial derivatives will equal the regressors,  $\mathbf{x}_i$ . In what follows, we will identify the derivatives of the conditional mean function with respect to the parameters as the “pseudoregressors,”  $\mathbf{x}_i^0(\boldsymbol{\beta}) = \mathbf{x}_i^0$ . We find that the nonlinear least squares estimator is found as the solutions to

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n \mathbf{x}_i^0 \varepsilon_i = \mathbf{0}. \quad (7-9)$$

This is the nonlinear regression counterpart to the least squares normal equations in (3-5). Computation requires an iterative solution. (See Example 7.3.) The method is presented in Section 7.2.6.

Assumptions 1 and 3 imply that  $E[\varepsilon_i | h(\mathbf{x}_i, \boldsymbol{\beta})] = 0$ . In the linear model, it follows, *because of the linearity of the conditional mean*, that  $\varepsilon_i$  and  $\mathbf{x}_i$ , itself, are uncorrelated. However, *uncorrelatedness* of  $\varepsilon_i$  with a particular *nonlinear* function of  $\mathbf{x}_i$  (the regression function) does not necessarily imply uncorrelatedness with  $\mathbf{x}_i$ , itself, nor, for that matter, with other nonlinear functions of  $\mathbf{x}_i$ . On the other hand, the results we will obtain for the behavior of the estimator in this model are couched not in terms of  $\mathbf{x}_i$  but in terms of certain functions of  $\mathbf{x}_i$  (the derivatives of the regression function), so, in point of fact,  $E[\boldsymbol{\varepsilon} | \mathbf{X}] = \mathbf{0}$  is not even the assumption we need.

The foregoing is not a theoretical fine point. Dynamic models, which are very common in the contemporary literature, would greatly complicate this analysis. If it can be assumed that  $\varepsilon_i$  is strictly uncorrelated with any *prior information* in the model, including

## CHAPTER 7 ♦ Nonlinear, Semiparametric 185

previous disturbances, then perhaps a treatment analogous to that for the linear model would apply. But the convergence results needed to obtain the asymptotic properties of the estimator still have to be strengthened. The dynamic nonlinear regression model is beyond the reach of our treatment here. Strict independence of  $\varepsilon_i$  and  $\mathbf{x}_i$  would be sufficient for uncorrelatedness of  $\varepsilon_i$  and every function of  $\mathbf{x}_i$ , but, again, in a dynamic model, this assumption might be questionable. Some commentary on this aspect of the nonlinear regression model may be found in Davidson and MacKinnon (1993, 2004).

If the disturbances in the nonlinear model are normally distributed, then the log of the normal density for the  $i$ th observation will be

$$\ln f(y_i | \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) = -(1/2)\{\ln 2\pi + \ln \sigma^2 + [y_i - h(\mathbf{x}_i, \boldsymbol{\beta})]^2 / \sigma^2\}. \quad (7-10)$$

For this special case, we have from item D.2 in Theorem 14.2 (on maximum likelihood estimation), that the derivatives of the log density with respect to the parameters have mean zero. That is,

$$E \left[ \frac{\partial \ln f(y_i | \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta}} \right] = E \left[ \frac{1}{\sigma^2} \left( \frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right) \varepsilon_i \right] = \mathbf{0}, \quad (7-11)$$

so, in the normal case, the derivatives and the disturbances are uncorrelated. Whether this can be assumed to hold in other cases is going to be model specific, but under reasonable conditions, we would assume so. [See Ruud (2000, p. 540).]

In the context of the linear model, the **orthogonality condition**  $E[\mathbf{x}_i \varepsilon_i] = 0$  produces least squares as a **GMM estimator** for the model. (See Chapter 13.) The orthogonality condition is that the regressors and the disturbance in the model are uncorrelated. In this setting, the same condition applies to the first derivatives of the conditional mean function. The result in (7-11) produces a moment condition which will define the nonlinear least squares estimator as a GMM estimator.

**Example 7.3 First-Order Conditions for a Nonlinear Model**

The first-order conditions for estimating the parameters of the nonlinear regression model,

$$y_i = \beta_1 + \beta_2 e^{\beta_3 x_i} + \varepsilon_i,$$

by nonlinear least squares [see (7-13)] are

$$\begin{aligned} \frac{\partial S(\mathbf{b})}{\partial b_1} &= - \sum_{i=1}^n [y_i - b_1 - b_2 e^{b_3 x_i}] = 0, \\ \frac{\partial S(\mathbf{b})}{\partial b_2} &= - \sum_{i=1}^n [y_i - b_1 - b_2 e^{b_3 x_i}] e^{b_3 x_i} = 0, \\ \frac{\partial S(\mathbf{b})}{\partial b_3} &= - \sum_{i=1}^n [y_i - b_1 - b_2 e^{b_3 x_i}] b_2 x_i e^{b_3 x_i} = 0. \end{aligned}$$

These equations do not have an explicit solution.

Conceding the potential for ambiguity, we define a nonlinear regression model at this point as follows.

## 186 PART I ♦ The Linear Regression Model

**DEFINITION 7.1 Nonlinear Regression Model**

*A **nonlinear regression model** is one for which the first-order conditions for least squares estimation of the parameters are nonlinear functions of the parameters.*

Thus, nonlinearity is defined in terms of the techniques needed to estimate the parameters, not the shape of the regression function. Later we shall broaden our definition to include other techniques besides least squares.

### 7.2.3 LARGE SAMPLE PROPERTIES OF THE NONLINEAR LEAST SQUARES ESTIMATOR

Numerous analytical results have been obtained for the nonlinear least squares estimator, such as consistency and asymptotic normality. We cannot be sure that nonlinear least squares is the most efficient estimator, except in the case of normally distributed disturbances. (This conclusion is the same one we drew for the linear model.) But, in the semiparametric setting of this chapter, we can ask whether this estimator is optimal in some sense given the information that we do have; the answer turns out to be yes. Some examples that follow will illustrate the points.

It is necessary to make some assumptions about the regressors. The precise requirements are discussed in some detail in Judge et al. (1985), Amemiya (1985), and Davidson and MacKinnon (2004). In the linear regression model, to obtain our asymptotic results, we assume that the sample moment matrix  $(1/n)\mathbf{X}'\mathbf{X}$  converges to a positive definite matrix  $\mathbf{Q}$ . By analogy, we impose the same condition on the derivatives of the regression function, which are called the **pseudoregressors** in the linearized model *when they are computed at the parameter values*. Therefore, for the nonlinear regression model, the analog to (4-21) is

$$\text{plim } \frac{1}{n} \mathbf{X}^0 \mathbf{X}^0 = \text{plim } \frac{1}{n} \sum_{i=1}^n \left( \frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_0} \right) \left( \frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_0} \right)' = \mathbf{Q}^0, \quad (7-12)$$

where  $\mathbf{Q}^0$  is a positive definite matrix. To establish consistency of  $\mathbf{b}$  in the linear model, we required  $\text{plim}(1/n)\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{0}$ . We will use the counterpart to this for the pseudoregressors:

$$\text{plim } \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i^0 \varepsilon_i = \mathbf{0}.$$

This is the orthogonality condition noted earlier in (4-24). In particular, note that orthogonality of the disturbances and the data is not the same condition. Finally, asymptotic normality can be established under general conditions if

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{x}_i^0 \varepsilon_i \xrightarrow{d} N[\mathbf{0}, \sigma^2 \mathbf{Q}^0].$$

With these in hand, the asymptotic properties of the nonlinear least squares estimator have been derived. They are, in fact, essentially those we have already seen for the

## CHAPTER 7 ♦ Nonlinear, Semiparametric 187

linear model, except that in this case we place the derivatives of the linearized function evaluated at  $\beta$ ,  $\mathbf{X}^0$  in the role of the regressors. [See Amemiya (1985).]

The nonlinear least squares criterion function is

$$S(\mathbf{b}) = \frac{1}{2} \sum_{i=1}^n [y_i - h(\mathbf{x}_i, \mathbf{b})]^2 = \frac{1}{2} \sum_{i=1}^n e_i^2, \quad (7-13)$$

where we have inserted what will be the solution value,  $\mathbf{b}$ . The values of the parameters that minimize (one half of) the sum of squared deviations are the nonlinear least squares estimators. The first-order conditions for a minimum are

$$\mathbf{g}(\mathbf{b}) = - \sum_{i=1}^n [y_i - h(\mathbf{x}_i, \mathbf{b})] \frac{\partial h(\mathbf{x}_i, \mathbf{b})}{\partial \mathbf{b}} = \mathbf{0}. \quad (7-14)$$

In the linear model of Chapter 3, this produces a set of linear equations, the normal equations (3-4). But in this more general case, (7-14) is a set of nonlinear equations that do not have an explicit solution. Note that  $\sigma^2$  is not relevant to the solution [nor was it in (3-4)]. At the solution,

$$\mathbf{g}(\mathbf{b}) = -\mathbf{X}^{0r} \mathbf{e} = \mathbf{0},$$

which is the same as (3-12) for the linear model.

Given our assumptions, we have the following general results:

**THEOREM 7.1 Consistency of the Nonlinear Least Squares Estimator**

*If the following assumptions hold;*

- a. *The parameter space containing  $\beta$  is compact (has no gaps or nonconcave regions),*
- b. *For any vector  $\beta^0$  in that parameter space,  $\text{plim } (1/n)S(\beta^0) = q(\beta^0)$ , a continuous and differentiable function,*
- c.  *$q(\beta^0)$  has a unique minimum at the true parameter vector,  $\beta$ ,*

*then, the nonlinear least squares estimator defined by (7-13) and (7-14) is consistent. We will sketch the proof, then consider why the theorem and the proof differ as they do from the apparently simpler counterpart for the linear model. The proof, notwithstanding the underlying subtleties of the assumptions, is straightforward. The estimator, say,  $\mathbf{b}^0$ , minimizes  $(1/n)S(\beta^0)$ . If  $(1/n)S(\beta^0)$  is minimized for every  $n$ , then it is minimized by  $\mathbf{b}^0$  as  $n$  increases without bound. We also assumed that the minimizer of  $q(\beta^0)$  is uniquely  $\beta$ . If the minimum value of  $\text{plim } (1/n)S(\beta^0)$  equals the probability limit of the minimized value of the sum of squares, the theorem is proved. This equality is produced by the continuity in assumption b.*

In the linear model, consistency of the least squares estimator could be established based on  $\text{plim}(1/n)\mathbf{X}'\mathbf{X} = \mathbf{Q}$  and  $\text{plim}(1/n)\mathbf{X}'\mathbf{e} = \mathbf{0}$ . To follow that approach here, we would use the linearized model and take essentially the same result. The loose

## 188 PART I ♦ The Linear Regression Model

end in that argument would be that the linearized model is not the true model, and there remains an approximation. For this line of reasoning to be valid, it must also be either assumed or shown that  $\text{plim}(1/n)\mathbf{X}^{0r}\boldsymbol{\delta} = \mathbf{0}$  where  $\delta_i = h(\mathbf{x}_i, \boldsymbol{\beta})$  minus the Taylor series approximation. An argument to this effect appears in Mittelhammer et al. (2000, pp. 190–191).

Note that no mention has been made of unbiasedness. The linear least squares estimator in the linear regression model is essentially alone in the estimators considered in this book. It is generally not possible to establish unbiasedness for any other estimator. As we saw earlier, unbiasedness is of fairly limited virtue in any event—we found, for example, that the property would not differentiate an estimator based on a sample of 10 observations from one based on 10,000. Outside the linear case, consistency is the primary requirement of an estimator. Once this is established, we consider questions of efficiency and, in most cases, whether we can rely on asymptotic normality as a basis for statistical inference.

**THEOREM 7.2 Asymptotic Normality of the Nonlinear Least Squares Estimator**

*If the pseudoregressors defined in (7-12) are “well behaved,” then*

$$\mathbf{b} \overset{a}{\sim} N \left[ \boldsymbol{\beta}, \frac{\sigma^2}{n} (\mathbf{Q}^0)^{-1} \right],$$

*where*

$$\mathbf{Q}^0 = \text{plim} \frac{1}{n} \mathbf{X}^{0r} \mathbf{X}^0.$$

*The sample estimator of the asymptotic covariance matrix is*

$$\text{Est. Asy. Var}[\mathbf{b}] = \hat{\sigma}^2 (\mathbf{X}^{0r} \mathbf{X}^0)^{-1}. \quad (7-15)$$

Asymptotic efficiency of the nonlinear least squares estimator is difficult to establish without a distributional assumption. There is an indirect approach that is one possibility. The assumption of the orthogonality of the pseudoregressors and the true disturbances implies that the nonlinear least squares estimator is a GMM estimator in this context. With the assumptions of homoscedasticity and nonautocorrelation, the optimal weighting matrix is the one that we used, which is to say that in the class of GMM estimators for this model, nonlinear least squares uses the optimal weighting matrix. As such, it is asymptotically efficient in the class of GMM estimators.

The requirement that the matrix in (7-12) converges to a positive definite matrix implies that the columns of the regressor matrix  $\mathbf{X}^0$  must be linearly independent. This **identification condition** is analogous to the requirement that the independent variables in the linear model be linearly independent. Nonlinear regression models usually involve several independent variables, and at first blush, it might seem sufficient to examine the data directly if one is concerned with multicollinearity. However, this situation is not the case. Example 7.4 gives an application.



## CHAPTER 7 ♦ Nonlinear, Semiparametric 189

A consistent estimator of  $\sigma^2$  is based on the residuals:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n [y_i - h(\mathbf{x}_i, \mathbf{b})]^2. \quad (7-16)$$

A degrees of freedom correction,  $1/(n - K)$ , where  $K$  is the number of elements in  $\boldsymbol{\beta}$ , is not strictly necessary here, because all results are asymptotic in any event. Davidson and MacKinnon (2004) argue that on average, (7-16) will underestimate  $\sigma^2$ , and one should use the degrees of freedom correction. Most software in current use for this model does, but analysts will want to verify which is the case for the program they are using. With this in hand, the estimator of the asymptotic covariance matrix for the nonlinear least squares estimator is given in (7-15).

Once the nonlinear least squares estimates are in hand, inference and hypothesis tests can proceed in the same fashion as prescribed in Chapter 5. A minor problem can arise in evaluating the fit of the regression in that the familiar measure,

$$R^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (7-17)$$

is no longer guaranteed to be in the range of 0 to 1. It does, however, provide a useful descriptive measure.

#### 7.2.4 HYPOTHESIS TESTING AND PARAMETRIC RESTRICTIONS

In most cases, the sorts of hypotheses one would test in this context will involve fairly simple linear restrictions. The tests can be carried out using the familiar formulas discussed in Chapter 5 and the asymptotic covariance matrix presented earlier. For more involved hypotheses and for nonlinear restrictions, the procedures are a bit less clear-cut. Two principal testing procedures were discussed in Section 5.4: the Wald test, which relies on the consistency and asymptotic normality of the estimator, and the  $F$  test, which is appropriate in finite (all) samples, that relies on normally distributed disturbances. In the nonlinear case, we rely on large-sample results, so the Wald statistic will be the primary inference tool. An analog to the  $F$  statistic based on the fit of the regression will also be developed later. Finally, **Lagrange multiplier tests** for the general case can be constructed. Since we have not assumed normality of the disturbances (yet), we will postpone treatment of the likelihood ratio statistic until we revisit this model in Chapter 14.

The hypothesis to be tested is

$$H_0: \mathbf{r}(\boldsymbol{\beta}) = \mathbf{q}, \quad (7-18)$$

where  $\mathbf{r}(\boldsymbol{\beta})$  is a column vector of  $J$  continuous functions of the elements of  $\boldsymbol{\beta}$ . These restrictions may be linear or nonlinear. It is necessary, however, that they be **overidentifying restrictions**. Thus, in formal terms, if the original parameter vector has  $K$  free elements, then the hypothesis  $\mathbf{r}(\boldsymbol{\beta}) - \mathbf{q}$  must impose at least one functional relationship on the parameters. If there is more than one restriction, then they must be functionally independent. These two conditions imply that the  $J \times K$  **Jacobian**,

$$\mathbf{R}(\boldsymbol{\beta}) = \frac{\partial \mathbf{r}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}, \quad (7-19)$$

## 190 PART I ♦ The Linear Regression Model

must have full row rank and that  $J$ , the number of restrictions, must be strictly less than  $K$ . This situation is analogous to the linear model, in which  $\mathbf{R}(\boldsymbol{\beta})$  would be the matrix of coefficients in the restrictions. (See, as well, Section 5.4, where the methods examined here are applied to the linear model.)

Let  $\mathbf{b}$  be the unrestricted, nonlinear least squares estimator, and let  $\mathbf{b}_*$  be the estimator obtained when the constraints of the hypothesis are imposed.<sup>3</sup> Which test statistic one uses depends on how difficult the computations are. Unlike the linear model, the various testing procedures vary in complexity. For instance, in our example, the Lagrange multiplier is by far the simplest to compute. Of the four methods we will consider, only this test does not require us to compute a nonlinear regression.

The nonlinear analog to the familiar  $F$  statistic based on the fit of the regression (i.e., the sum of squared residuals) would be

$$F[J, n - K] = \frac{[S(\mathbf{b}_*) - S(\mathbf{b})]/J}{S(\mathbf{b})/(n - K)}. \quad (7-20)$$

This equation has the appearance of our earlier  $F$  ratio in (5-29). In the nonlinear setting, however, neither the numerator nor the denominator has exactly the necessary chi-squared distribution, so the  $F$  distribution is only approximate. Note that this  $F$  statistic requires that both the restricted and unrestricted models be estimated.

The Wald test is based on the distance between  $\mathbf{r}(\mathbf{b})$  and  $\mathbf{q}$ . If the unrestricted estimates fail to satisfy the restrictions, then doubt is cast on the validity of the restrictions. The statistic is

$$\begin{aligned} W &= [\mathbf{r}(\mathbf{b}) - \mathbf{q}]' \{ \text{Est. Asy. Var}[\mathbf{r}(\mathbf{b}) - \mathbf{q}] \}^{-1} [\mathbf{r}(\mathbf{b}) - \mathbf{q}] \\ &= [\mathbf{r}(\mathbf{b}) - \mathbf{q}]' \{ \mathbf{R}(\mathbf{b}) \hat{\mathbf{V}} \mathbf{R}'(\mathbf{b}) \}^{-1} [\mathbf{r}(\mathbf{b}) - \mathbf{q}], \end{aligned} \quad (7-21)$$

where

$$\hat{\mathbf{V}} = \text{Est. Asy. Var}[\mathbf{b}],$$

and  $\mathbf{R}(\mathbf{b})$  is evaluated at  $\mathbf{b}$ , the estimate of  $\boldsymbol{\beta}$ .

Under the null hypothesis, this statistic has a limiting chi-squared distribution with  $J$  degrees of freedom. If the restrictions are correct, the Wald statistic and  $J$  times the  $F$  statistic are asymptotically equivalent. The Wald statistic can be based on the estimated covariance matrix obtained earlier using the unrestricted estimates, which may provide a large savings in computing effort if the restrictions are nonlinear. It should be noted that the small-sample behavior of  $W$  can be erratic, and the more conservative  $F$  statistic may be preferable if the sample is not large.

The caveat about Wald statistics that applied in the linear case applies here as well. Because it is a pure significance test that does not involve the alternative hypothesis, the Wald statistic is not invariant to how the hypothesis is framed. In cases in which there are more than one equivalent ways to specify  $\mathbf{r}(\boldsymbol{\beta}) = \mathbf{q}$ ,  $W$  can give different answers depending on which is chosen.

The Lagrange multiplier test is based on the decrease in the sum of squared residuals that would result if the restrictions in the restricted model were released. The formalities of the test are given in Section 14.6.3. For the nonlinear regression model,

<sup>3</sup>This computational problem may be extremely difficult in its own right, especially if the constraints are nonlinear. We assume that the estimator has been obtained by whatever means are necessary.

## CHAPTER 7 ♦ Nonlinear, Semiparametric 191

the test has a particularly appealing form.<sup>4</sup> Let  $\mathbf{e}_*$  be the vector of residuals  $y_i - h(\mathbf{x}_i, \mathbf{b}_*)$  computed using the restricted estimates. Recall that we defined  $\mathbf{X}^0$  as an  $n \times K$  matrix of derivatives computed at a particular parameter vector in (7-29). Let  $\mathbf{X}_*^0$  be this matrix *computed at the restricted estimates*. Then the Lagrange multiplier statistic for the nonlinear regression model is

$$\text{LM} = \frac{\mathbf{e}_*' \mathbf{X}_*^0 [\mathbf{X}_*^0{}' \mathbf{X}_*^0]^{-1} \mathbf{X}_*^0{}' \mathbf{e}_*}{\mathbf{e}_*' \mathbf{e}_* / n}. \quad (7-22)$$

Under  $H_0$ , this statistic has a limiting chi-squared distribution with  $J$  degrees of freedom. What is especially appealing about this approach is that it requires only the restricted estimates. This method may provide some savings in computing effort if, as in our example, the restrictions result in a linear model. Note, also, that the Lagrange multiplier statistic is  $n$  times the uncentered  $R^2$  in the regression of  $\mathbf{e}_*$  on  $\mathbf{X}_*^0$ . Many Lagrange multiplier statistics are computed in this fashion.

## 7.2.5 APPLICATIONS

This section will present three applications of estimation and inference for nonlinear regression models. Example 7.4 illustrates a nonlinear consumption function that extends Examples 1.2 and 2.1. The model provides a simple demonstration of estimation and hypothesis testing for a nonlinear model. Example 7.5 analyzes the Box–Cox transformation. This specification is used to provide a more general functional form than the linear regression—it has the linear and loglinear models as special cases. Finally, Example 7.6 is a lengthy examination of an exponential regression model. In this application, we will explore some of the implications of nonlinear modeling, specifically “interaction effects.” We examined interaction effects in Section 6.3.3 in a model of the form

$$y = \beta_1 + \beta_2 x + \beta_3 z + \beta_4 xz + \varepsilon.$$

In this case, the interaction effect is  $\partial^2 E[y|x, z] / \partial x \partial z = \beta_4$ . There is no interaction effect if  $\beta_4$  equals zero. Example 7.6 considers the (perhaps unintended) implication of the nonlinear model that when  $E[y|x, z] = h(x, z, \boldsymbol{\beta})$ , there is an interaction effect even if the model is

$$h(x, z, \boldsymbol{\beta}) = h(\beta_1 + \beta_2 x + \beta_3 z).$$

**Example 7.4 Analysis of a Nonlinear Consumption Function**

The linear consumption function analyzed at the beginning of Chapter 2 is a restricted version of the more general consumption function

$$C = \alpha + \beta Y^\gamma + \varepsilon,$$

in which  $\gamma$  equals 1. With this restriction, the model is linear. If  $\gamma$  is free to vary, however, then this version becomes a nonlinear regression. Quarterly data on consumption, real disposable income, and several other variables for the U.S. economy for 1950 to 2000 are listed in Appendix Table F5.2. We will use these to fit the nonlinear consumption function. (Details of the computation of the estimates are given in Section 7.2.6 in Example 7.8.) The restricted linear and unrestricted nonlinear least squares regression results are shown in Table 7.1.

The procedures outlined earlier are used to obtain the asymptotic standard errors and an estimate of  $\sigma^2$ . (To make this comparable to  $s^2$  in the linear model, the value includes the degrees of freedom correction.)

<sup>4</sup>This test is derived in Judge et al. (1985). A lengthy discussion appears in Mittelhammer et al. (2000).

## 192 PART I ♦ The Linear Regression Model

TABLE 7.1 Estimated Consumption Functions

| Parameter               | Linear Model  |                | Nonlinear Model |                |
|-------------------------|---------------|----------------|-----------------|----------------|
|                         | Estimate      | Standard Error | Estimate        | Standard Error |
| $\alpha$                | -80.3547      | 14.3059        | 458.7990        | 22.5014        |
| $\beta$                 | 0.9217        | 0.003872       | 0.10085         | 0.01091        |
| $\gamma$                | 1.0000        | —              | 1.24483         | 0.01205        |
| $\mathbf{e}'\mathbf{e}$ | 1,536,321.881 |                | 504,403.1725    |                |
| $\sigma$                | 87.20983      |                | 50.0946         |                |
| $R^2$                   | 0.996448      |                | 0.998834        |                |
| $\text{Var}[b]$         | —             |                | 0.000119037     |                |
| $\text{Var}[c]$         | —             |                | 0.00014532      |                |
| $\text{Cov}[b, c]$      | —             |                | -0.000131491    |                |

In the preceding example, there is no question of collinearity in the data matrix  $\mathbf{X} = [\mathbf{i}, \mathbf{y}]$ ; the variation in  $Y$  is obvious on inspection. But, at the final parameter estimates, the  $R^2$  in the regression is 0.998834 and the correlation between the two pseudoregressors  $x_2^0 = Y^\gamma$  and  $x_3^0 = \beta Y^\gamma \ln Y$  is 0.999752. The condition number for the normalized matrix of sums of squares and cross products is 208.306. (The condition number is computed by computing the square root of the ratio of the largest to smallest characteristic root of  $\mathbf{X}_0' \mathbf{X}_0 \mathbf{D}^{-1}$  where  $x_1^0 = 1$  and  $\mathbf{D}$  is the diagonal matrix containing the square roots of  $\mathbf{x}_k^0' \mathbf{x}_k^0$  on the diagonal.) Recall that 20 was the benchmark for a problematic data set. By the standards discussed in Section 4.7.1 and A.6.6, the collinearity problem in this “data set” is severe. In fact, it appears not to be a problem at all.

For hypothesis testing and confidence intervals, the familiar procedures can be used, with the proviso that all results are only asymptotic. As such, for testing a restriction, the chi-squared statistic rather than the  $F$  ratio is likely to be more appropriate. For example, for testing the hypothesis that  $\gamma$  is different from 1, an asymptotic  $t$  test, based on the standard normal distribution, is carried out, using

$$z = \frac{1.24483 - 1}{0.01205} = 20.3178.$$

This result is larger than the critical values of 1.96 for the 5 percent significance level, and thus reject the linear model in favor of the nonlinear regression. The three procedures for hypotheses produce the same conclusion.

- The  $F$  statistic is

$$F[1.204 - 3] = \frac{(1,536,321.881 - 504,403.17)/1}{504,403.17/(204 - 3)} = 411.29.$$

The critical value from the tables is 3.84, so the hypothesis is rejected.

- The Wald statistic is based on the distance of  $\hat{\gamma}$  from 1 and is simply the square of the asymptotic  $t$  ratio we computed earlier:

$$W = \frac{(1.24483 - 1)^2}{0.01205^2} = 412.805.$$

The critical value from the chi-squared table is 3.84.

- For the Lagrange multiplier statistic, the elements in  $\mathbf{x}_i^*$  are

$$\mathbf{x}_i^* = [1, Y^\gamma, \beta Y^\gamma \ln Y].$$

To compute this at the restricted estimates, we use the ordinary least squares estimates for  $\alpha$  and  $\beta$  and 1 for  $\gamma$  so that

$$\mathbf{x}_i^* = [1, Y, \beta Y \ln Y].$$

## CHAPTER 7 ♦ Nonlinear, Semiparametric 193

The residuals are the least squares residuals computed from the linear regression. Inserting the values given earlier, we have

$$LM = \frac{996,103.9}{(1,536,321.881/204)} = 132.267.$$

As expected, this statistic is also larger than the critical value from the chi-squared table.

We are also interested in the marginal propensity to consume. In this expanded model,  $H_0 : \gamma = 1$  is a test that the marginal propensity to consume is constant, not that it is 1. (That would be a joint test of both  $\gamma = 1$  and  $\beta = 1$ .) In this model, the marginal propensity to consume is

$$MPC = dC/dY = \beta\gamma Y^{\gamma-1},$$

which varies with  $Y$ . To test the hypothesis that this value is 1, we require a particular value of  $Y$ . Because it is the most recent value, we choose  $DPI/2000.4 = 6634.9$ . At this value, the MPC is estimated as 0.86971. We estimate its standard error using the delta method, with the square root of

$$\begin{aligned} & \begin{bmatrix} \partial MPC/\partial b & \partial MPC/\partial c \end{bmatrix} \begin{bmatrix} \text{Var}[b] & \text{Cov}[b, c] \\ \text{Cov}[b, c] & \text{Var}[c] \end{bmatrix} \begin{bmatrix} \partial MPC/\partial b \\ \partial MPC/\partial c \end{bmatrix} \\ &= \begin{bmatrix} cY^{c-1} & bY^{c-1}(1 + c \ln Y) \end{bmatrix} \begin{bmatrix} 0.000119037 & -0.000131491 \\ -0.000131491 & 0.00014532 \end{bmatrix} \begin{bmatrix} cY^{c-1} \\ bY^{c-1}(1 + c \ln Y) \end{bmatrix} \\ &= 0.00007469, \end{aligned}$$

which gives a standard error of 0.0086423. For testing the hypothesis that the MPC is equal to 1.0 in 2000.4 we would refer  $z = (1.08264 - 1)/0.0086423 = -9.56299$  to the standard normal table. This difference is certainly statistically significant, so we would reject the hypothesis.

### Example 7.5 The Box–Cox Transformation

The **Box–Cox transformation** [Box and Cox (1964), Zarembka (1974)] is used as a device for generalizing the linear model. The transformation is

$$x^{(\lambda)} = (x^\lambda - 1)/\lambda.$$

Special cases of interest are  $\lambda = 1$ , which produces a linear transformation,  $x^{(1)} = x - 1$ , and  $\lambda = 0$ . When  $\lambda$  equals zero, the transformation is, by L'Hôpital's rule,

$$\lim_{\lambda \rightarrow 0} \frac{x^\lambda - 1}{\lambda} = \lim_{\lambda \rightarrow 0} \frac{d(x^\lambda - 1)/d\lambda}{1} = \lim_{\lambda \rightarrow 0} x^\lambda \times \ln x = \ln x.$$

The regression analysis can be done *conditionally* on  $\lambda$ . For a given value of  $\lambda$ , the model,

$$y = \alpha + \sum_{k=2}^K \beta_k x_k^{(\lambda)} + \varepsilon, \quad (7-23)$$

is a linear regression that can be estimated by least squares. However, if  $\lambda$  in (7-23) is taken to be an unknown parameter, then the regression becomes nonlinear in the parameters.

In principle, each regressor could be transformed by a different value of  $\lambda$ , but, in most applications, this level of generality becomes excessively cumbersome, and  $\lambda$  is assumed to be the same for all the variables in the model.<sup>5</sup> To be defined for all values of  $\lambda$ ,  $x$  must be strictly positive. In most applications, some of the regressors—for example, a dummy variable—will not be transformed. For such a variable, say  $v_k$ ,  $v_k^{(\lambda)} = v_k$ , and the relevant derivatives in (7-24) will be zero. It is also possible to transform  $y$ , say, by  $y^{(\theta)}$ . Transformation of the dependent variable, however, amounts to a specification of the whole model, not just

<sup>5</sup>See, for example, Seaks and Layson (1983).

## 194 PART I ♦ The Linear Regression Model

the functional form of the conditional mean. For example,  $\theta = 1$  implies a linear equation while  $\theta = 0$  implies a logarithmic equation.

In some applications, the motivation for the transformation is to program around zero values in a loglinear model. Caves, Christensen, and Trethaway (1980) analyzed the costs of production for railroads providing freight and passenger service. Continuing a long line of literature on the costs of production in regulated industries, a translog cost function (see Section 10.4.2) would be a natural choice for modeling this multiple-output technology. Several of the firms in the study, however, produced no passenger service, which would preclude the use of the translog model. (This model would require the log of zero.) An alternative is the Box–Cox transformation, which is computable for zero output levels. A question does arise in this context (and other similar ones) as to whether zero outputs should be treated the same as nonzero outputs or whether an output of zero represents a discrete corporate decision distinct from other variations in the output levels. In addition, as can be seen in (7-24), this solution is only partial. The zero values of the regressors preclude computation of appropriate standard errors.

Nonlinear least squares is straightforward. In most instances, we can expect to find the least squares value of  $\lambda$  between  $-2$  and  $2$ . Typically, then,  $\lambda$  is estimated by scanning this range for the value that minimizes the sum of squares. Note what happens if there are zeros for  $x$  in the sample. Then, a constraint must still be placed on  $\lambda$  in their model, as  $0^{(\lambda)}$  is defined only if  $\lambda$  is strictly positive. A positive value of  $\lambda$  is not assured. Once the optimal value of  $\lambda$  is located, the least squares estimates, the mean squared residual, and this value of  $\lambda$  constitute the nonlinear least squares estimates of the parameters.

After determining the optimal value of  $\lambda$ , it is sometimes treated as if it were a known value in the least squares results. But  $\hat{\lambda}$  is an estimate of an unknown parameter. It is not hard to show that the least squares standard errors will always underestimate the correct asymptotic standard errors.<sup>6</sup> To get the appropriate values, we need the derivatives of the right-hand side of (7-23) with respect to  $\alpha$ ,  $\beta$ , and  $\lambda$ . The pseudoregressors are

$$\begin{aligned}\frac{\partial h(\cdot)}{\partial \alpha} &= 1, \\ \frac{\partial h(\cdot)}{\partial \beta_k} &= x_k^{(\lambda)}, \\ \frac{\partial h(\cdot)}{\partial \lambda} &= \sum_{k=1}^K \beta_k \frac{\partial x_k^{(\lambda)}}{\partial \lambda} = \sum_{k=1}^K \beta_k \left[ \frac{1}{\lambda} (x_k^\lambda \ln x_k - x_k^{(\lambda)}) \right].\end{aligned}\tag{7-24}$$

We can now use (7-15) and (7-16) to estimate the asymptotic covariance matrix of the parameter estimates. Note that  $\ln x_k$  appears in  $\partial h(\cdot)/\partial \lambda$ . If  $x_k = 0$ , then this matrix cannot be computed. This was the point noted earlier.

It is important to remember that the coefficients in a nonlinear model are not equal to the slope (or the elasticities) with respect to the variables. For the particular Box–Cox model in (7-23),

$$\frac{\partial E[\ln y | \mathbf{x}]}{\partial \ln x_k} = x_k \frac{\partial E[\ln y | \mathbf{x}]}{\partial x_k} = \beta_k x_k^\lambda = \eta_k.$$

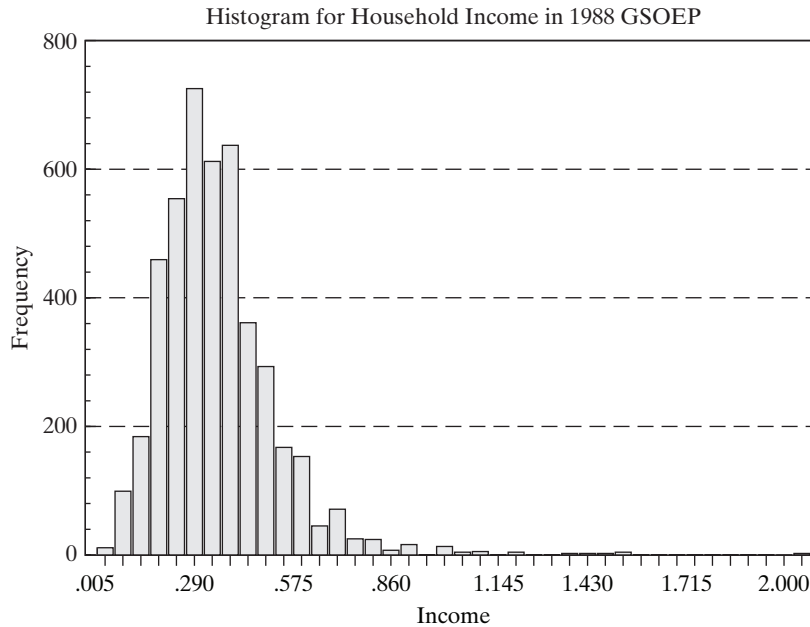
Standard errors for these estimates can be obtained using the **delta method**. The derivatives are  $\partial \eta / \partial \beta_k = x_k^\lambda = \eta_k / \beta_k$  and  $\partial \eta / \partial \lambda = \eta \ln x_k$ . Collecting terms, we obtain

$$\text{Asy.Var}[\hat{\eta}_k] = (\eta_k / \beta_k)^2 \{ \text{Asy.Var}[\hat{\beta}_k] + (\beta \ln x_k)^2 \text{Asy.Var}[\hat{\lambda}] + (2\beta \ln x_k) \text{Asy.Cov}[\hat{\beta}_k, \hat{\lambda}] \}$$

The application in Example 7.4 is a Box–Cox model of the sort discussed here. We can rewrite (7-23) as

$$\begin{aligned}y &= (\alpha - 1/\lambda) + (\beta/\lambda) X^\lambda + \varepsilon \\ &= \alpha^* + \beta^* X^\lambda + \varepsilon.\end{aligned}$$

<sup>6</sup>See Fomby, Hill, and Johnson (1984, pp. 426–431).



**FIGURE 7.1** Histogram for Income.

This shows that an alternative way to handle the Box–Cox regression model is to transform the model into a nonlinear regression and then use the Gauss–Newton regression (see Section 7.2.6) to estimate the parameters. The original parameters of the model can be recovered by  $\lambda = \gamma$ ,  $\alpha = \alpha^* + 1/\gamma$  and  $\beta = \gamma\beta^*$ .

**Example 7.6 Interaction Effects in a Loglinear Model for Income**

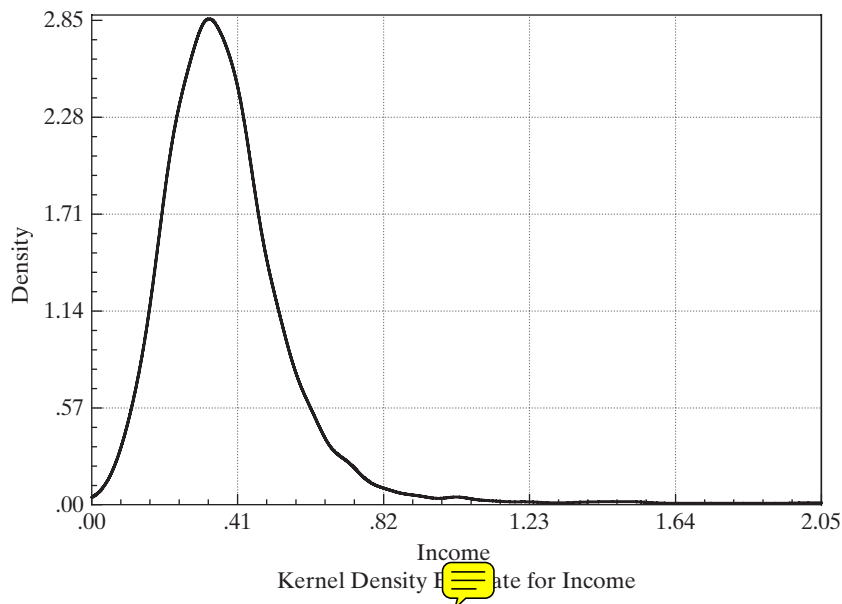
A recent study in health economics is “Incentive Effects in the Demand for Health Care: A Bivariate Panel Count Data Estimation” by Riphahn, Wambach, and Million (2003). The authors were interested in counts of physician visits and hospital visits and in the impact that the presence of private insurance had on the utilization counts of interest, that is, whether the data contain evidence of moral hazard. The sample used is an unbalanced panel of 7,293 households, the German Socioeconomic Panel (GSOEP) data set.<sup>7</sup> Among the variables reported in the panel are household income, with numerous other sociodemographic variables such as age, gender, and education. For this example, we will model the distribution of income using the last wave of the data set (1988), a cross section with 4,483 observations. Two of the individuals in this sample reported zero income, which is incompatible with the underlying models suggested in the development below. Deleting these two observations leaves a sample of 4,481 observations. Figures 7.1 and 7.2 display a histogram and a kernel density estimator for the household income variable for these observations.

We will fit an exponential regression model to the income variable, with

$$\begin{aligned} \text{Income} = & \exp(\beta_1 + \beta_2 \text{Age} + \beta_3 \text{Age}^2 + \beta_4 \text{Education} + \beta_5 \text{Female} \\ & + \beta_6 \text{Female} \times \text{Education} + \beta_7 \text{Age} \times \text{Education}) + \varepsilon. \end{aligned}$$

<sup>7</sup>The data are published on the *Journal of Applied Econometrics* data archive web site, at <http://qed.econ.queensu.ca/jae/2003-v18.4/riphahn-wambach-million/>. The variables in the data file are listed in Appendix Table F7.1. The number of observations in each year varies from one to seven with a total number of 27,326 observations. We will use these data in several examples here and later in the book.

196 PART I ♦ The Linear Regression Model



**FIGURE 7.2** Kernel Density Estimator for Income.

Table 7.2 provides descriptive statistics for the variables used in this application.

**Loglinear models** play a prominent role in statistics. Many derive from a density function of the form  $f(y|\mathbf{x}) = p[y|\alpha^0 + \mathbf{x}'\boldsymbol{\beta}, \theta]$ , where  $\alpha^0$  is a constant term and  $\theta$  is an additional parameter, and

$$E[y|\mathbf{x}] = g(\theta) \exp(\alpha^0 + \mathbf{x}'\boldsymbol{\beta}),$$

(hence the name “loglinear models”). Examples include the Weibull, gamma, lognormal, and exponential models for continuous variables and the Poisson and negative binomial models for counts. We can write  $E[y|\mathbf{x}]$  as  $\exp[\ln g(\theta) + \alpha^0 + \mathbf{x}'\boldsymbol{\beta}]$ , and then absorb  $\ln g(\theta)$  in the constant term in  $\ln E[y|\mathbf{x}] = \alpha + \mathbf{x}'\boldsymbol{\beta}$ . The lognormal distribution (see Section B.4.4) is often used to model incomes. For the lognormal random variable,

$$p[y|\alpha^0 + \mathbf{x}'\boldsymbol{\beta}, \theta] = \frac{\exp[-\frac{1}{2}(\ln y - \alpha^0 - \mathbf{x}'\boldsymbol{\beta})^2/\theta^2]}{\theta y \sqrt{2\pi}}, y > 0,$$

$$E[y|\mathbf{x}] = \exp(\alpha^0 + \mathbf{x}'\boldsymbol{\beta} + \theta^2/2) = \exp(\alpha + \mathbf{x}'\boldsymbol{\beta}).$$

**TABLE 7.2** Descriptive Statistics for Variables Used in Nonlinear Regression

| Variable | Mean    | Std.Dev. | Min   | Maximum |
|----------|---------|----------|-------|---------|
| INCOME   | .348896 | .164054  | .0050 | 2       |
| AGE      | 43.4452 | 11.2879  | 25.00 | 64      |
| EDUC     | 11.4167 | 2.36615  | 7.000 | 18      |
| FEMALE   | .484267 | .499808  | .0000 | 1       |



## CHAPTER 7 ♦ Nonlinear, Semiparametric 197

The exponential regression model is also consistent with a gamma distribution. The density of a gamma distributed random variable is

$$p[y|\alpha^0 + \mathbf{x}'\boldsymbol{\beta}, \theta] = \frac{\lambda^\theta \exp(-\lambda y) y^{\theta-1}}{\Gamma(\theta)}, y > 0, \theta > 0, \lambda = \exp(-\alpha^0 - \mathbf{x}'\boldsymbol{\beta}),$$

$$E[y|\mathbf{x}] = \theta/\lambda = \theta \exp(\alpha^0 + \mathbf{x}'\boldsymbol{\beta}) = \exp(\ln \theta + \alpha^0 + \mathbf{x}'\boldsymbol{\beta}) = \exp(\alpha + \mathbf{x}'\boldsymbol{\beta}).$$

The parameter  $\theta$  determines the shape of the distribution. When  $\theta > 2$ , the gamma density has the shape of a chi-squared variable (which is a special case). Finally, the Weibull model has a similar form,

$$p[y|\alpha^0 + \mathbf{x}'\boldsymbol{\beta}, \theta] = \theta \lambda^\theta \exp[-(\lambda y)^\theta] y^{\theta-1}, y \geq 0, \theta > 0, \lambda = \exp(-\alpha^0 - \mathbf{x}'\boldsymbol{\beta}),$$

$$E[y|\mathbf{x}] = \Gamma(1 + 1/\theta) \exp(\alpha^0 + \mathbf{x}'\boldsymbol{\beta}) = \exp[\ln \Gamma(1 + 1/\theta) + \alpha^0 + \mathbf{x}'\boldsymbol{\beta}] = \exp(\alpha + \mathbf{x}'\boldsymbol{\beta}).$$

In all cases, the maximum likelihood estimator is the most efficient estimator of the parameters. (Maximum likelihood estimation of the parameters of this model is considered in Chapter 14.) However, nonlinear least squares estimation of the model

$$E[y|\mathbf{x}] = \exp(\alpha + \mathbf{x}'\boldsymbol{\beta}) + \varepsilon$$

has a virtue in that the nonlinear least squares estimator will be consistent even if the distributional assumption is incorrect—it is *robust* to this type of misspecification since it does not make explicit use of a distributional assumption.

Table 7.3 presents the nonlinear least squares regression results. Superficially, the pattern of signs and significance might be expected—with the exception of the dummy variable for female. However, two issues complicate the interpretation of the coefficients in this model. First, the model is nonlinear, so the coefficients do not give the magnitudes of the interesting effects in the equation. In particular, for this model,

$$\partial E[y|\mathbf{x}]/\partial x_k = \exp(\alpha + \mathbf{x}'\boldsymbol{\beta}) \times \partial(\alpha + \mathbf{x}'\boldsymbol{\beta})/\partial x_k.$$

Second, as we have constructed our model, the second part of the derivative is not equal to the coefficient, because the variables appear either in a quadratic term or as a product with some other variable. Moreover, for the dummy variable, *Female*, we would want to compute the partial effect using

$$\Delta E[y|\mathbf{x}]/\Delta \text{Female} = E[y|\mathbf{x}, \text{Female} = 1] - E[y|\mathbf{x}, \text{Female} = 0]$$

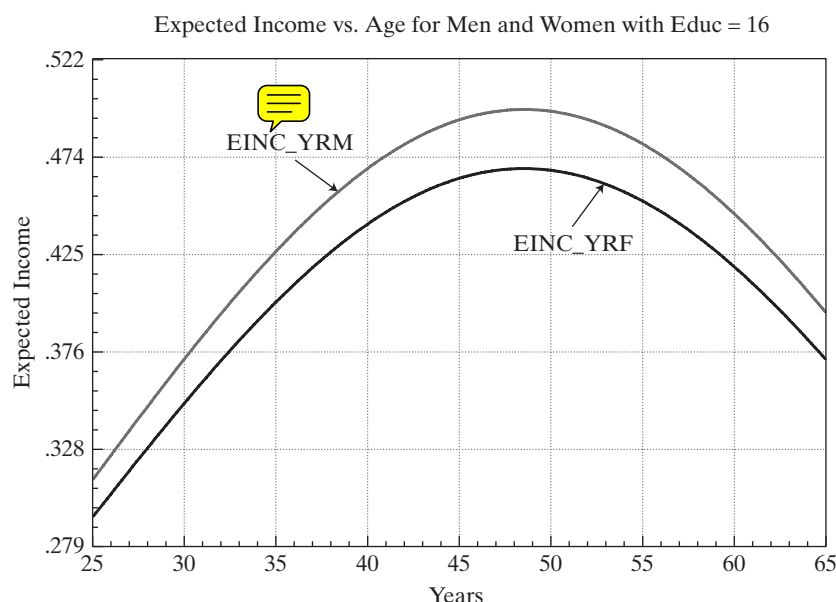
A third consideration is how to compute the partial effects, as sample averages or at the means of the variables. For example,

$$\partial E[y|\mathbf{x}]/\partial \text{Age} = E[y|\mathbf{x}] \times (\beta_2 + 2\beta_3 \text{Age} + \beta_7 \text{Educ}).$$

**TABLE 7.3** Estimated Regression Equations

| Variable         | Nonlinear Least Squares |            |        | Linear Least Squares |            |        |
|------------------|-------------------------|------------|--------|----------------------|------------|--------|
|                  | Estimate                | Std. Error | t      | Estimate             | Std. Error | t      |
| Constant         | −2.58070                | .17455     | 14.78  | −.13050              | .06261     | −2.08  |
| Age              | .06020                  | .00615     | 9.79   | .01791               | .00214     | 8.37   |
| Age <sup>2</sup> | −.00084                 | .00006082  | −13.83 | −.00027              | .00001985  | −13.51 |
| Education        | −.00616                 | .01095     | −.56   | −.00281              | .00418     | −.67   |
| Female           | .17497                  | .05986     | 2.92   | .07955               | .02339     | 3.40   |
| Female × Educ    | −.01476                 | .00493     | −2.99  | −.00685              | .00202     | −3.39  |
| Age × Educ       | .00134                  | .00024     | 5.59   | .00055               | .00009394  | 5.88   |
| e'e              |                         | 106.09825  |        |                      | 106.24323  |        |
| s                |                         | .15387     |        |                      | .15410     |        |
| R <sup>2</sup>   |                         | .12005     |        |                      | .11880     |        |

## 198 PART I ♦ The Linear Regression Model

**FIGURE 7.3** Expected Incomes.

The average value of *Age* in the sample is 43.4452 and the average *Education* is 11.4167. The partial effect of a year of education is estimated to be 0.000948 if it is computed by computing the partial effect for each individual and averaging the result. It is 0.000925 if it is computed by computing the conditional mean and the linear term at the averages of the three variables. The partial effect is difficult to interpret without information about the scale of the income variable. Since the average income in the data is about 0.35, these partial effects suggest that an additional year of education is associated with a change in expected income of about 2.6 percent (i.e.,  $0.009/0.35$ ).

The rough calculation of partial effects with respect to *Age* does not reveal the model implications about the relationship between age and expected income. Note, for example, that the coefficient on *Age* is positive while the coefficient on  $Age^2$  is negative. This implies (neglecting the interaction term at the end), that the *Age – Income* relationship implied by the model is parabolic. The partial effect is positive at some low values and negative at higher values. To explore this, we have computed the expected *Income* using the model separately for men and women, both with assumed college education ( $Educ = 16$ ) and for the range of ages in the sample, 25 to 64. Figure 7.3 shows the result of this calculation. The upper curve is for men ( $Female = 0$ ) and the lower one is for women. The parabolic shape is as expected; what the figure reveals is the relatively strong effect—*ceteris paribus*, incomes are predicted to rise by about 80 percent between ages 25 and 48. (There is an important aspect of this computation that the model builder would want to develop in the analysis. It remains to be argued whether this parabolic relationship describes the trajectory of expected income for an individual as they age, or the average incomes of different cohorts at a particular moment in time (1988). The latter would seem to be the more appropriate conclusion at this point, though one might be tempted to infer the former.)

The figure reveals a second implication of the estimated model that would not be obvious from the regression results. The coefficient on the dummy variable for *Female* is positive, highly significant, and, in isolation, by far the largest effect in the model. This might lead the analyst to conclude that on average, expected incomes in these data are higher for women than men. But, Figure 7.3 shows precisely the opposite. The difference is accounted

## CHAPTER 7 ♦ Nonlinear, Semiparametric 199

for by the interaction term, *Female*  $\times$  *Education*. The negative sign on the latter coefficient is suggestive. But, the total effect would remain ambiguous without the sort of secondary analysis suggested by the figure.

Finally, in addition to the quadratic term in age, the model contains an interaction term, *Age*  $\times$  *Education*. The coefficient is positive and highly significant. But, it is far from obvious how this should be interpreted. In a linear model,

$$\begin{aligned} \text{Income} = & \beta_1 + \beta_2 \text{Age} + \beta_3 \text{Age}^2 + \beta_4 \text{Education} + \beta_5 \text{Female} \\ & + \beta_6 \text{Female} \times \text{Education} + \beta_7 \text{Age} \times \text{Education} + \varepsilon, \end{aligned}$$

we would find that  $\beta_7 = \partial^2 E[\text{Income}|x]/\partial \text{Age} \partial \text{Education}$ . That is, the “interaction effect” is the change in the partial effect of *Age* associated with a change in *Education* (or vice versa). Of course, if  $\beta_7$  equals zero, that is, if there is no product term in the model, then there is no interaction effect—the second derivative equals zero. However, this simple interpretation usually does not apply in nonlinear models (i.e., in any nonlinear model). Consider our exponential regression, and suppose that in fact,  $\beta_7$  is indeed zero. For convenience, let  $\mu(x)$  equal the conditional mean function. Then, the partial effect with respect to *Age* is

$$\partial \mu(x) / \partial \text{Age} = \mu(x) \times (\beta_2 + 2\beta_3 \text{Age})$$

and

$$\partial^2 \mu(x) / \partial \text{Age} \partial \text{Educ} = \mu(x) \times (\beta_2 + 2\beta_3 \text{Age})(\beta_4 + \beta_6 \text{Female}), \quad (7-25)$$

which is nonzero even if there is no “interaction term” in the model. The interaction effect in the model that we estimated, which includes the product term, is

$$\partial^2 E[y|x] / \partial \text{Age} \partial \text{Educ} = \mu(x) \times [\beta_7 + (\beta_2 + 2\beta_3 \text{Age} + \beta_7 \text{Educ})(\beta_4 + \beta_6 \text{Female} + \beta_7 \text{Age})]. \quad (7-26)$$

At least some of what is being called the interaction effect in this model is attributable entirely to the fact the model is nonlinear. To isolate the “functional form effect” from the true “interaction effect,” we might subtract (7-25) from (7-26) and then reassemble the components:

$$\begin{aligned} \partial^2 \mu(x) / \partial \text{Age} \partial \text{Educ} = & \mu(x) [(\beta_2 + 2\beta_3 \text{Age})(\beta_4 + \beta_6 \text{Female})] \\ & + \mu(x) \beta_7 [1 + \text{Age}(\beta_2 + 2\beta_3) + \text{Educ}(\beta_4 + \beta_6 \text{Female}) + \text{Educ} \times \text{Age}(\beta_7)]. \quad (7-27) \end{aligned}$$

It is clear that the coefficient on the product term bears essentially no relationship to the quantity of interest (assuming it is the change in the partial effects that is of interest). On the other hand, the second term is nonzero if and only if  $\beta_7$  is nonzero. One might, therefore, identify the second part with the “interaction effect” in the model. Whether a behavioral interpretation could be attached to this is questionable, however. Moreover, that would leave unexplained the functional form effect. The point of this exercise is to suggest that one should proceed with some caution in interpreting interaction effects in nonlinear models. This sort of analysis has a focal point in the literature in Ai and Norton (2004). A number of comments and extensions of the result are to be found, including Greene (2010).

We make one final observation about the nonlinear regression. In a loglinear, single-index function model such as the one analyzed here, one might, “for comparison purposes,” compute simple linear least squares results. The coefficients in the right-hand side of Table 7.3 suggest superficially that nonlinear least squares and least squares are computing completely different relationships. To uncover the similarity (if there is one), it is useful to consider the partial effects rather than the coefficients. We found, for example, the partial effect of education in the nonlinear model, using the means of the variables, is 0.000925. Although the linear least squares coefficients are very different, if the partial effect for education is computed for the linear equation, we find  $-0.00281 - 0.00685(.5) + 0.00055(43.4452) = 0.01766$ , where we have used 0.5 for *Female*. Dividing by 0.35, we obtain 0.0504, which is at least close to its counterpart in the nonlinear model. As a general result, at least approximately, the linear least squares coefficients are making this approximation.

## 200 PART I ♦ The Linear Regression Model

### 7.2.6 COMPUTING THE NONLINEAR LEAST SQUARES ESTIMATOR

Minimizing the sum of squared residuals for a nonlinear regression is a standard problem in nonlinear optimization that can be solved by a number of methods. (See Section E.3.) The method of Gauss–Newton is often used. This algorithm (and most of the sampling theory results for the asymptotic properties of the estimator) is based on a linear Taylor series approximation to the nonlinear regression function. The iterative estimator is computed by transforming the optimization to a series of linear least squares regressions.

The nonlinear regression model is  $y = h(\mathbf{x}, \boldsymbol{\beta}) + \varepsilon$ . (To save some notation, we have dropped the observation subscript). The procedure is based on a linear Taylor series approximation to  $h(\mathbf{x}, \boldsymbol{\beta})$  at a particular value for the parameter vector,  $\boldsymbol{\beta}^0$ :

$$h(\mathbf{x}, \boldsymbol{\beta}) \approx h(\mathbf{x}, \boldsymbol{\beta}^0) + \sum_{k=1}^K \frac{\partial h(\mathbf{x}, \boldsymbol{\beta}^0)}{\partial \beta_k^0} (\beta_k - \beta_k^0). \quad (7-28)$$

This form of the equation is called the **linearized regression model**. By collecting terms, we obtain

$$h(\mathbf{x}, \boldsymbol{\beta}) \approx \left[ h(\mathbf{x}, \boldsymbol{\beta}^0) - \sum_{k=1}^K \beta_k^0 \left( \frac{\partial h(\mathbf{x}, \boldsymbol{\beta}^0)}{\partial \beta_k^0} \right) \right] + \sum_{k=1}^K \beta_k \left( \frac{\partial h(\mathbf{x}, \boldsymbol{\beta}^0)}{\partial \beta_k^0} \right). \quad (7-29)$$

Let  $x_k^0$  equal the  $k$ th partial derivative,<sup>8</sup>  $\partial h(\mathbf{x}, \boldsymbol{\beta}^0) / \partial \beta_k^0$ . For a given value of  $\boldsymbol{\beta}^0$ ,  $x_k^0$  is a function only of the data, not of the unknown parameters. We now have

$$h(\mathbf{x}, \boldsymbol{\beta}) \approx \left[ h^0 - \sum_{k=1}^K x_k^0 \beta_k^0 \right] + \sum_{k=1}^K x_k^0 \beta_k,$$

which may be written

$$h(\mathbf{x}, \boldsymbol{\beta}) \approx h^0 - \mathbf{x}^{0r} \boldsymbol{\beta}^0 + \mathbf{x}^{0r} \boldsymbol{\beta},$$

which implies that

$$y \approx h^0 - \mathbf{x}^{0r} \boldsymbol{\beta}^0 + \mathbf{x}^{0r} \boldsymbol{\beta} + \varepsilon.$$

By placing the known terms on the left-hand side of the equation, we obtain a linear equation:

$$y^0 = y - h^0 + \mathbf{x}^{0r} \boldsymbol{\beta}^0 = \mathbf{x}^{0r} \boldsymbol{\beta} + \varepsilon^0. \quad (7-30)$$

Note that  $\varepsilon^0$  contains both the true disturbance,  $\varepsilon$ , and the error in the first-order Taylor series approximation to the true regression, shown in (7-29). That is,

$$\varepsilon^0 = \varepsilon + \left[ h(\mathbf{x}, \boldsymbol{\beta}) - \left\{ h^0 - \sum_{k=1}^K x_k^0 \beta_k^0 + \sum_{k=1}^K x_k^0 \beta_k \right\} \right]. \quad (7-31)$$

Because all the errors are accounted for, (7-30) is an equality, not an approximation. With a value of  $\boldsymbol{\beta}^0$  in hand, we can compute  $y^0$  and  $\mathbf{x}^0$  and then estimate the parameters of (7-30) by linear least squares. (Whether this estimator is consistent or not remains to be seen.)

<sup>8</sup>You should verify that for the linear regression model, these derivatives are the independent variables.

## CHAPTER 7 ♦ Nonlinear, Semiparametric 201

**Example 7.7 Linearized Regression**

For the model in Example 7.3, the regressors in the linearized equation would be

$$\begin{aligned}x_1^0 &= \frac{\partial h(\cdot)}{\partial \beta_1^0} = 1, \\x_2^0 &= \frac{\partial h(\cdot)}{\partial \beta_2^0} = e^{\beta_3^0 x}, \\x_3^0 &= \frac{\partial h(\cdot)}{\partial \beta_3^0} = \beta_2^0 x e^{\beta_3^0 x}.\end{aligned}$$

With a set of values of the parameters  $\beta^0$ ,

$$y^0 = y - h(x, \beta_1^0, \beta_2^0, \beta_3^0) + \beta_1^0 x_1^0 + \beta_2^0 x_2^0 + \beta_3^0 x_3^0$$

can be linearly regressed on the three variables previously defined to estimate  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$ .

The linearized regression model shown in (7-30) can be estimated by linear least squares. Once a parameter vector is obtained, it can play the role of a new  $\beta^0$ , and the computation can be done again. The **iteration** can continue until the difference between successive parameter vectors is small enough to assume convergence. One of the main virtues of this method is that at the last iteration the estimate of  $(\mathbf{Q}^0)^{-1}$  will, apart from the scale factor  $\hat{\sigma}^2/n$ , provide the correct estimate of the asymptotic covariance matrix for the parameter estimator.

This iterative solution to the minimization problem is

$$\begin{aligned}\mathbf{b}_{t+1} &= \left[ \sum_{i=1}^n \mathbf{x}_i^0 \mathbf{x}_i^{0'} \right]^{-1} \left[ \sum_{i=1}^n \mathbf{x}_i^0 (y_i - h_i^0 + \mathbf{x}_i^{0'} \mathbf{b}_t) \right] \\&= \mathbf{b}_t + \left[ \sum_{i=1}^n \mathbf{x}_i^0 \mathbf{x}_i^{0'} \right]^{-1} \left[ \sum_{i=1}^n \mathbf{x}_i^0 (y_i - h_i^0) \right] \\&= \mathbf{b}_t + (\mathbf{X}^{0'} \mathbf{X}^0)^{-1} \mathbf{X}^{0'} \mathbf{e}^0 \\&= \mathbf{b}_t + \Delta_t,\end{aligned} \tag{7-32}$$


where all terms on the right-hand side are evaluated at  $\mathbf{b}_t$  and  $\mathbf{e}^0$  is the vector of nonlinear least squares residuals. This algorithm has some intuitive appeal as well. For each iteration, we update the previous parameter estimates by regressing the nonlinear least squares residuals on the derivatives of the regression functions. The process will have converged (i.e., the update will be  $\mathbf{0}$ ) when  $\mathbf{X}^{0'} \mathbf{e}^0$  is close enough to  $\mathbf{0}$ . This derivative has a direct counterpart in the normal equations for the linear model,  $\mathbf{X}' \mathbf{e} = \mathbf{0}$ .

As usual, when using a digital computer, we will not achieve exact convergence with  $\mathbf{X}^{0'} \mathbf{e}^0$  exactly equal to zero. A useful, scale-free counterpart to the convergence criterion discussed in Section E.3.6 is  $\delta = \mathbf{e}^{0'} \mathbf{X}^0 (\mathbf{X}^{0'} \mathbf{X}^0)^{-1} \mathbf{X}^{0'} \mathbf{e}^0$ . [See (7-22).] We note, finally, that iteration of the linearized regression, although a very effective algorithm for many problems, does not always work. As does Newton's method, this algorithm sometimes "jumps off" to a wildly errant second iterate, after which it may be impossible to compute the residuals for the next iteration. The choice of starting values for the iterations can be crucial. There is art as well as science in the computation of nonlinear least squares estimates. [See McCullough and Vinod (1999).] In the absence of information about starting values, a workable strategy is to try the Gauss-Newton iteration first. If it

## 202 PART I ♦ The Linear Regression Model

fails, go back to the initial starting values and try one of the more general algorithms, such as BFGS, treating minimization of the sum of squares as an otherwise ordinary optimization problem.

### Example 7.8 Nonlinear Least Squares

Example 7.4 considered analysis of a nonlinear consumption function 

$$C = \alpha + \beta Y^\gamma + \varepsilon.$$

The linearized regression model is

$$C - (\alpha^0 + \beta^0 Y^{\gamma^0}) + (\alpha^0 1 + \beta^0 Y^{\gamma^0} + \gamma^0 \beta^0 Y^{\gamma^0} \ln Y) = \alpha + \beta(Y^{\gamma^0}) + \gamma(\beta^0 Y^{\gamma^0} \ln Y) + \varepsilon^0.$$

Combining terms, we find that the nonlinear least squares procedure reduces to iterated regression of

$$C^0 = C + \gamma^0 \beta^0 Y^{\gamma^0} \ln Y$$

on

$$\mathbf{x}^0 = \begin{bmatrix} \frac{\partial h(\cdot)}{\partial \alpha} & \frac{\partial h(\cdot)}{\partial \beta} & \frac{\partial h(\cdot)}{\partial \gamma} \end{bmatrix}' = \begin{bmatrix} 1 \\ Y^{\gamma^0} \\ \beta^0 Y^{\gamma^0} \ln Y \end{bmatrix}.$$

Finding the **starting values** for a nonlinear procedure can be difficult. Simply trying a convenient set of values can be unproductive. Unfortunately, there are no good rules for starting values, except that they should be as close to the final values as possible (not particularly helpful). When it is possible, an initial consistent estimator of  $\beta$  will be a good starting value. In many cases, however, the only consistent estimator available is the one we are trying to compute by least squares. For better or worse, trial and error is the most frequently used procedure. For the present model, a natural set of values can be obtained because a simple linear model is a special case. Thus, we can start  $\alpha$  and  $\beta$  at the linear least squares values that would result in the special case of  $\gamma = 1$  and use 1 for the starting value for  $\gamma$ . The **iterations** are begun at the least squares estimates for  $\alpha$  and  $\beta$  and 1 for  $\gamma$ .

The solution is reached in eight iterations, after which any further iteration is merely “fine tuning” the hidden digits (i.e., those that the analyst would not be reporting to their reader. “Gradient” is the scale-free convergence measure,  $\delta$ , noted earlier.) Note that the coefficient vector takes a very errant step after the first iteration—the sum of squares becomes huge—but the iterations settle down after that and converge routinely.

Begin NLSQ iterations. Linearized regression.

Iteration = 1; Sum of squares = 1536321.88; Gradient = 996103.930

Iteration = 2; Sum of squares = 0.184780956E+12; Gradient = 0.184780452E+12 ( $\times 10^{12}$ )

Iteration = 3; Sum of squares = 20406917.6; Gradient = 19902415.7

Iteration = 4; Sum of squares = 581703.598; Gradient = 77299.6342

Iteration = 5; Sum of squares = 504403.969; Gradient = 0.752189847

Iteration = 6; Sum of squares = 504403.216; Gradient = 0.526642396E-04

Iteration = 7; Sum of squares = 504403.216; Gradient = 0.511324981E-07

Iteration = 8; Sum of squares = 504403.216; Gradient = 0.606793426E-10

## 7.3 MEDIAN AND QUANTILE REGRESSION

We maintain the essential assumptions of the linear regression model,

$$y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$$

where  $E[\varepsilon|\mathbf{x}] = 0$  and  $E[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$ . If  $\varepsilon|\mathbf{x}$  is normally distributed, so that the distribution of  $\varepsilon|\mathbf{x}$  is also symmetric, then the median,  $\text{Med}[\varepsilon|\mathbf{x}]$ , is also zero and  $\text{Med}[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$ .

## CHAPTER 7 ♦ Nonlinear, Semiparametric 203

Under these assumptions, least squares remains a natural choice for estimation of  $\beta$ . But, as we explored in Example 4.5, **least absolute deviations** (LAD) is a possible alternative that might even be preferable in a small sample. Suppose, however, that we depart from the second assumption directly. That is, the statement of the model is

$$\text{Med}[y|\mathbf{x}] = \mathbf{x}'\beta.$$

This result suggests a motivation for LAD in its own right, rather than as a robust (to outliers) alternative to least squares.<sup>9</sup> The conditional median of  $y_i|\mathbf{x}_i$  might be an interesting function. More generally, other quantiles of the distribution of  $y_i|\mathbf{x}_i$  might also be of interest. For example, we might be interested in examining the various quantiles of the distribution of income or spending. Quantile regression (rather than least squares) is used for this purpose. The (linear) quantile regression model can be defined as

$$Q[y|\mathbf{x}, q] = \mathbf{x}'\beta_q \text{ such that } \text{Prob}[y \leq \mathbf{x}'\beta_q|\mathbf{x}] = q, 0 < q < 1. \quad (7-33)$$

The **median regression** would be defined for  $q = \frac{1}{2}$ . Other focal points are the lower and upper quartiles,  $q = \frac{1}{4}$  and  $q = \frac{3}{4}$ , respectively. We will develop the median regression in detail in Section 7.3.1, once again largely as an alternative estimator in the linear regression setting.

The quantile regression model is a richer specification than the linear model that we have studied thus far, because the coefficients in (7-33) are indexed by  $q$ . The model is nonparametric—it requires a much less detailed specification of the distribution of  $y|\mathbf{x}$ . In the simplest linear model with fixed coefficient vector,  $\beta$ , the quantiles of  $y|\mathbf{x}$  would be defined by variation of the constant term. The implication of the model is shown in Figure 7.4. For a fixed  $\beta$  and conditioned on  $x$ , the value of  $\alpha_q + \beta x$  such that  $\text{Prob}(y < \alpha_q + \beta x)$  is shown for  $q = 0.15, 0.5$ , and  $0.9$  in Figure 7.4. There is a value of  $\alpha_q$  for each quantile. In Section 7.3.2, we will examine the more general specification of the quantile regression model in which the entire coefficient vector plays the role of  $\alpha_q$  in Figure 7.4.

### 7.3.1 LEAST ABSOLUTE DEVIATIONS ESTIMATION

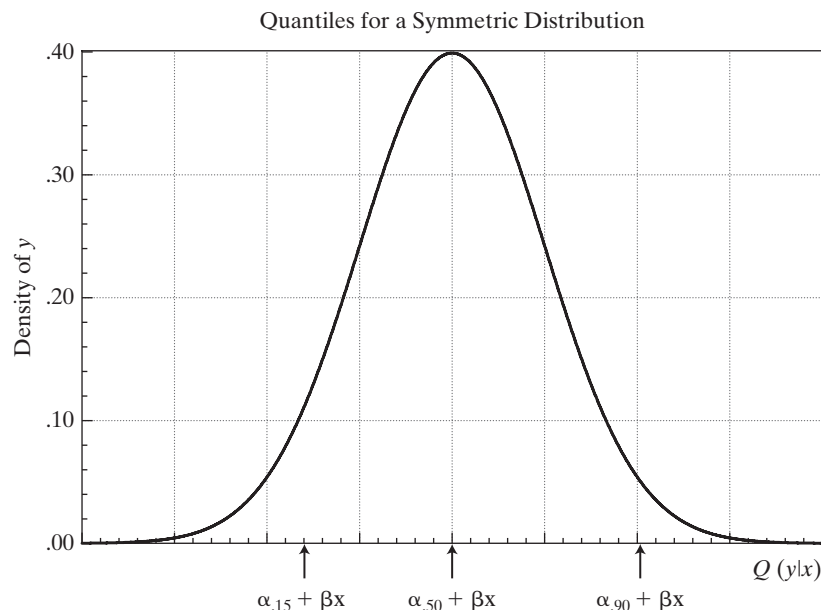
Least squares can be severely distorted by outlying observations. Recent applications in microeconomics and financial economics involving thick-tailed disturbance distributions, for example, are particularly likely to be affected by precisely these sorts of observations. (Of course, in those applications in finance involving hundreds of thousands of observations, which are becoming commonplace, this discussion is moot.) These applications have led to the proposal of “robust” estimators that are unaffected by outlying observations.<sup>10</sup> In this section, we will examine one of these, the least absolute deviations, or LAD estimator.

That least squares gives such large weight to large deviations from the regression causes the results to be particularly sensitive to small numbers of atypical data points when the sample size is small or moderate. The least absolute deviations (LAD) estimator has been suggested as an alternative that remedies (at least to some degree) the

<sup>9</sup>In Example 4.5, we considered the possibility that in small samples with possibly thick-tailed disturbance distributions, the LAD estimator might have smaller variance than least squares.

<sup>10</sup>For some applications, see Taylor (1974), Amemiya (1985, pp. 70–80), Andrews (1974), Koenker and Bassett (1978), and a survey written at a very accessible level by Birkes and Dodge (1993). A somewhat more rigorous treatment is given by Hardle (1990).

## 204 PART I ♦ The Linear Regression Model

**FIGURE 7.4** Quantile Regression Model.

problem. The LAD estimator is the solution to the optimization problem,

$$\text{Min}_{\mathbf{b}_0} \sum_{i=1}^n |y_i - \mathbf{x}'_i \mathbf{b}_0|.$$

The LAD estimator's history predates least squares (which itself was proposed over 200 years ago). It has seen little use in econometrics, primarily for the same reason that Gauss's method (LS) supplanted LAD at its origination; LS is vastly easier to compute. Moreover, in a more modern vein, its statistical properties are more firmly established than LAD's and samples are usually large enough that the small sample advantage of LAD is not needed.

The LAD estimator is a special case of the quantile regression:

$$\text{Prob}[y_i \leq \mathbf{x}'_i \boldsymbol{\beta}_q] = q.$$

The LAD estimator estimates the *median regression*. That is, it is the solution to the quantile regression when  $q = 0.5$ . Koenker and Bassett (1978, 1982), Huber (1967), and Rogers (1993) have analyzed this regression.<sup>11</sup> Their results suggest an estimator for the asymptotic covariance matrix of the quantile regression estimator,

$$\text{Est. Asy. Var}[\mathbf{b}_q] = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{D}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1},$$

<sup>11</sup>Powell (1984) has extended the LAD estimator to produce a robust estimator for the case in which observations on the dependent variable are censored, that is, when negative values of  $y_i$  are recorded as zero. See Example 14.7 for discussion and Melenberg and van Soest (1996) for an application. For some related results on other semiparametric approaches to regression, see Butler et al. (1990) and McDonald and White (1993).



## CHAPTER 7 ♦ Nonlinear, Semiparametric 205

where  $\mathbf{D}$  is a diagonal matrix containing weights

$$d_i = \left[ \frac{q}{f(0)} \right]^2 \text{ if } y_i - \mathbf{x}'_i \boldsymbol{\beta} \text{ is positive and } \left[ \frac{1-q}{f(0)} \right]^2 \text{ otherwise,}$$

and  $f(0)$  is the true density of the disturbances evaluated at 0.<sup>12</sup> [It remains to obtain an estimate of  $f(0)$ .] There is a useful symmetry in this result. Suppose that the true density were normal with variance  $\sigma^2$ . Then the preceding would reduce to  $\sigma^2(\pi/2)(\mathbf{X}'\mathbf{X})^{-1}$ , which is the result we used in Example 4.5. For more general cases, some other empirical estimate of  $f(0)$  is going to be required. Nonparametric methods of density estimation are available [see Section 12.4 and, e.g., Johnston and DiNardo (1997, pp. 370–375)]. But for the small sample situations in which techniques such as this are most desirable (our application below involves 25 observations), nonparametric kernel density estimation of a single ordinate is optimistic; these are, after all, asymptotic results. But asymptotically, as suggested by Example 4.5, the results begin overwhelmingly to favor least squares. For better or worse, a convenient estimator would be a **kernel density estimator** as described in Section 12.4.1. Looking ahead, the computation would be

$$\hat{f}(0) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h} K \left[ \frac{e_i}{h} \right]$$

where  $h$  is the **bandwidth** (to be discussed shortly),  $K[\cdot]$  is a weighting, or kernel function and  $e_i, i = 1, \dots, n$  is the set of residuals. There are no hard and fast rules for choosing  $h$ ; one popular choice is that used by Stata (2006),  $h = .9s/n^{1/5}$ . The kernel function is likewise discretionary, though it rarely matters much which one chooses; the logit kernel (see Table 12.2) is a common choice.

The **bootstrap** method of inferring statistical properties is well suited for this application. Since the efficacy of the bootstrap has been established for this purpose, the search for a formula for standard errors of the LAD estimator is not really necessary. The bootstrap estimator for the asymptotic covariance matrix can be computed as follows:

$$\text{Est. Var}[\mathbf{b}_{LAD}] = \frac{1}{R} \sum_{r=1}^R (\mathbf{b}_{LAD}(r) - \mathbf{b}_{LAD})(\mathbf{b}_{LAD}(r) - \mathbf{b}_{LAD})',$$

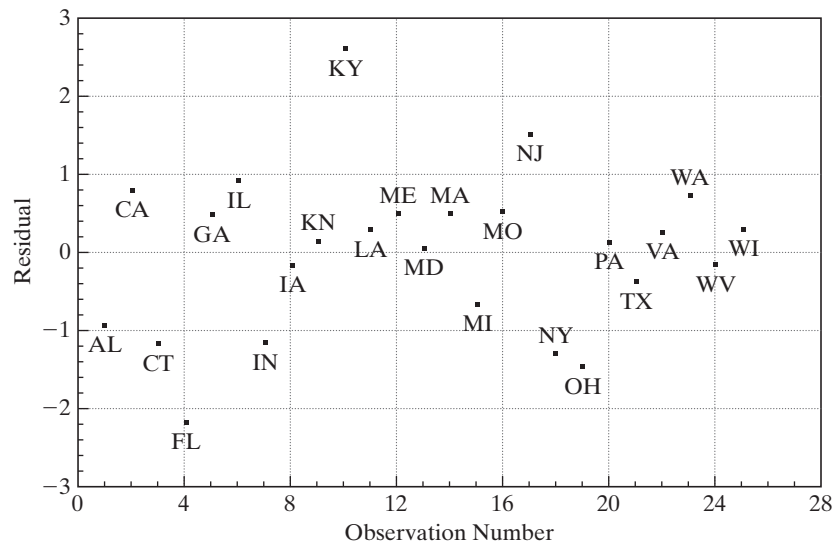
where  $\mathbf{b}_{LAD}$  is the LAD estimator and  $\mathbf{b}_{LAD}(r)$  is the  $r$ th LAD estimate of  $\boldsymbol{\beta}$  based on a sample of  $n$  observations, drawn with replacement, from the original data set.

**Example 7.9 LAD Estimation of a Cobb–Douglas Production Function**

Zellner and Revankar (1970) proposed a generalization of the Cobb–Douglas production function that allows economies of scale to vary with output. Their statewide data on  $Y$  = value added (output),  $K$  = capital,  $L$  = labor, and  $N$  = the number of establishments in the transportation industry are given in Appendix Table F7.2. For this application, estimates of the

<sup>12</sup>Koenker suggests that for independent and identically distributed observations, one should replace  $d_i$  with the constant  $a = q(1-q)/[f(F^{-1}(q))]^2 = [.50/f(0)]^2$  for the median (LAD) estimator. This reduces the expression to the true asymptotic covariance matrix,  $a(\mathbf{X}'\mathbf{X})^{-1}$ . The one given is a sample estimator which will behave the same in large samples. (Personal communication to the author.)

## 206 PART I ♦ The Linear Regression Model

**FIGURE 7.5** Standardized Residuals for Production Function.**TABLE 7.4** LS and LAD Estimates of a Production Function

| <i>Least Squares</i> |                 |                 |                | <i>LAD</i>      |                   |                |                       |                |
|----------------------|-----------------|-----------------|----------------|-----------------|-------------------|----------------|-----------------------|----------------|
| <i>Coefficient</i>   | <i>Estimate</i> | <i>Standard</i> |                | <i>Estimate</i> | <i>Bootstrap</i>  |                | <i>Kernel Density</i> |                |
|                      |                 | <i>Error</i>    | <i>t Ratio</i> |                 | <i>Std. Error</i> | <i>t Ratio</i> | <i>Std. Error</i>     | <i>t Ratio</i> |
| Constant             | 2.293           | 0.107           | 21.396         | 2.275           | 0.202             | 11.246         | 0.183                 | 12.374         |
| $\beta_k$            | 0.279           | 0.081           | 3.458          | 0.261           | 0.124             | 2.099          | 0.138                 | 1.881          |
| $\beta_l$            | 0.927           | 0.098           | 9.431          | 0.927           | 0.121             | 7.637          | 0.169                 | 5.498          |
| $\Sigma e^2$         | 0.7814          |                 |                | 0.7984          |                   |                |                       |                |
| $\Sigma  e $         | 3.3652          |                 |                | 3.2541          |                   |                |                       |                |

Cobb–Douglas production function,

$$\ln(Y_i/N_i) = \beta_1 + \beta_2 \ln(K_i/N_i) + \beta_3 \ln(L_i/N_i) + \varepsilon_i,$$

are obtained by least squares and LAD. The standardized least squares residuals shown in Figure 7.5 suggest that two observations (Florida and Kentucky) are outliers by the usual construction. The least squares coefficient vectors with and without these two observations are (2.293, 0.279, 0.927) and (2.205, 0.261, 0.879), respectively, which bears out the suggestion that these two points do exert considerable influence. Table 7.4 presents the LAD estimates of the same parameters, with standard errors based on 500 bootstrap replications. The LAD estimates with and without these two observations are identical, so only the former are presented. Using the simple approximation of multiplying the corresponding OLS standard error by  $(\pi/2)^{1/2} = 1.2533$  produces a surprisingly close estimate of the bootstrap estimated standard errors for the two slope parameters (0.102, 0.123) compared with the bootstrap estimates of (0.124, 0.121). The second set of estimated standard errors are based on Koenker's suggested estimator,  $.25/\hat{f}^2(0) = 0.25/1.5467^2 = 0.104502$ . The bandwidth and kernel function are those suggested earlier. The results are surprisingly consistent given the small sample size.

## 7.3.2 QUANTILE REGRESSION MODELS

The quantile regression model is

$$Q[y|\mathbf{x}, q] = \mathbf{x}'\boldsymbol{\beta}_q \text{ such that } \text{Prob}[y \leq \mathbf{x}'\boldsymbol{\beta}_q|\mathbf{x}] = q, 0 < q < 1.$$

This is essentially a nonparametric specification. No assumption is made about the distribution of  $y|\mathbf{x}$  or about its conditional variance. The fact that  $q$  can vary continuously (strictly) between zero and one means that there are an infinite number of possible “parameter vectors.” It seems reasonable to view the coefficients, which we might write  $\boldsymbol{\beta}(q)$  less as fixed “parameters,” as we do in the linear regression model, than loosely as *features* of the distribution of  $y|\mathbf{x}$ . For example, it is not likely to be meaningful to view  $\boldsymbol{\beta}(.49)$  to be discretely different from  $\boldsymbol{\beta}(.50)$  or to compute precisely a particular difference such as  $\boldsymbol{\beta}(.5) - \boldsymbol{\beta}(.3)$ . On the other hand, the qualitative difference, or possibly the lack of a difference, between  $\boldsymbol{\beta}(.3)$  and  $\boldsymbol{\beta}(.5)$  as displayed in our following example, may well be an interesting characteristic of the sample.

The estimator,  $\mathbf{b}_q$  of  $\boldsymbol{\beta}_q$  for a specific quantile is computed by minimizing the function

$$\begin{aligned} F_n(\boldsymbol{\beta}_q|\mathbf{y}, \mathbf{X}) &= \sum_{i: y_i \geq \mathbf{x}_i'\boldsymbol{\beta}_q}^n q|y_i - \mathbf{x}_i'\boldsymbol{\beta}_q| + \sum_{i: y_i < \mathbf{x}_i'\boldsymbol{\beta}_q}^n (1-q)|y_i - \mathbf{x}_i'\boldsymbol{\beta}_q| \\ &= \sum_{i=1}^n g(y_i - \mathbf{x}_i'\boldsymbol{\beta}_q|q) \end{aligned}$$

where

$$g(e_{i,q}|q) = \begin{cases} qe_{i,q} & \text{if } e_{i,q} \geq 0 \\ (1-q)e_{i,q} & \text{if } e_{i,q} < 0 \end{cases}, e_{i,q} = y_i - \mathbf{x}_i'\boldsymbol{\beta}_q.$$

When  $q = 0.5$ , the estimator is the least absolute deviations estimator we examined in Example 4.5 and Section 7.3.1. Solving the minimization problem requires an iterative estimator. It can be set up as a linear programming problem.<sup>13</sup> [See Keonker and D’Oray (1987).]

We cannot use the methods of Chapter 4 to determine the asymptotic covariance matrix of the estimator. But, the fact that the estimator is obtained by minimizing a sum does lead to a set of results similar to those we obtained in Section 4.4 for least squares. [See Buchinsky (1998).] Assuming that the regressors are “well behaved,” the quantile regression estimator of  $\boldsymbol{\beta}_q$  is consistent and asymptotically normally distributed with asymptotic covariance matrix

$$\text{Asy. Var.}[b_q] = \frac{1}{n} \mathbf{H}^{-1} \mathbf{G} \mathbf{H}^{-1}$$

where

$$\mathbf{H} = \text{plim} \frac{1}{n} \sum_{i=1}^n f_q(0|\mathbf{x}_i) \mathbf{x}_i \mathbf{x}_i'$$

<sup>13</sup>Quantile regression is supported as a built in procedure in contemporary software such as Stata, SAS, and NLOGIT.

## 208 PART I ♦ The Linear Regression Model

and

$$\mathbf{G} = \text{plim} \frac{q(1-q)}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i'.$$

This is the result we had earlier for the LAD estimator, now with quantile  $q$  instead of 0.5. As before, computation is complicated by the need to compute the density of  $\varepsilon_q$  at zero. This will require either an approximation of uncertain quality or a specification of the particular density, which we have hoped to avoid. The usual approach, as before, is to use bootstrapping.

**Example 7.10 Income Elasticity of Credit Card Expenditure**

Greene (1992, 2007) analyzed the default behavior and monthly expenditure behavior of a large sample (13,444 observations) of credit card users. Among the results of interest in the study was an estimate of the income elasticity of the monthly expenditure. A conventional regression approach might be based on

$$Q[\ln \text{Spending} | \mathbf{x}, q] = \beta_{1,q} + \beta_{2,q} \ln \text{Income} + \beta_{3,q} \text{Age} + \beta_{4,q} \text{Dependents}$$

The data in Appendix Table F7.3 contain these and numerous other covariates that might explain spending; we have chosen these three for this example only. The 13,444 observations in the data set are based on credit card applications. Of the full sample, 10,499 applications were approved and the next 12 months of spending and default behavior were observed.<sup>14</sup> Spending is the average monthly expenditure in the 12 months after the account was initiated. Average monthly income and number of household dependents are among the demographic data in the application. Table 7.5 presents least squares estimates of the coefficients of the conditional mean function as well as full results for several quantiles.<sup>15</sup> Standard errors are shown for the least squares and median ( $1 = 0.5$ ) results. The results for the other quantiles are essentially the same. The least squares estimate of 1.08344 is slightly and significantly greater than one—the estimated standard error is 0.03212 so the  $t$  statistic is  $(1.08344 - 1)/.03212 = 2.60$ . This suggests an aspect of consumption behavior that might not be surprising. However, the very large amount of variation over the range of quantiles might not have been expected. We might guess that at the highest levels of spending for any income level, there is (comparably so) some saturation in the response of spending to changes in income.

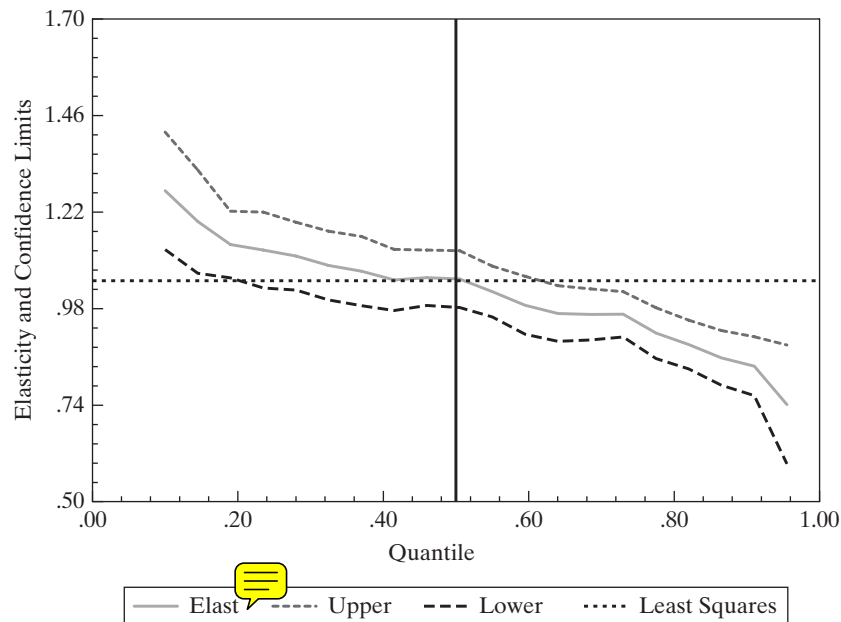
Figure 7.6 displays the estimates of the income elasticity of expenditure for the range of quantiles from 0.1 to 0.9, with the least squares estimate which would correspond to the fixed value at all quantiles shown in the center of the figure. Confidence limits shown in the figure are based on the asymptotic normality of the estimator. They are computed as the estimated income elasticity plus and minus 1.96 times the estimated standard error. Figure 7.7 shows the implied quantile regressions for  $q = .1, .3, .5, .7$ , and  $.9$ . The relatively large increase from the .1 quantile to the .3 suggests some skewness in the spending distribution. In broad

<sup>14</sup>The expenditure data are taken from the credit card records while the income and demographic data are taken from the applications. While it might be tempting to use, for example, Powell's (1986a,b) censored quantile regression estimator to accommodate this large cluster of zeros for the dependent variable, this approach would misspecify the model—the “zeros” represent nonexistent observations, not missing ones. A more detailed approach—the one used in the 1992 study—would model separately the presence or absence of the observation on spending, then model spending conditionally on acceptance of the application. We will revisit this issue in Chapter 17 in the context of the sample selection model. The income data are censored at 100,000 and 220 of the observations have expenditures that are filled with \$1 or less. We have not “cleaned” the data set for these aspects. The full 10,499 observations have been used as they are in the original data set.

<sup>15</sup>We would note, if (7-33) is the statement of the model, then it does not follow that that the conditional mean function is a linear regression. That would be an additional assumption.

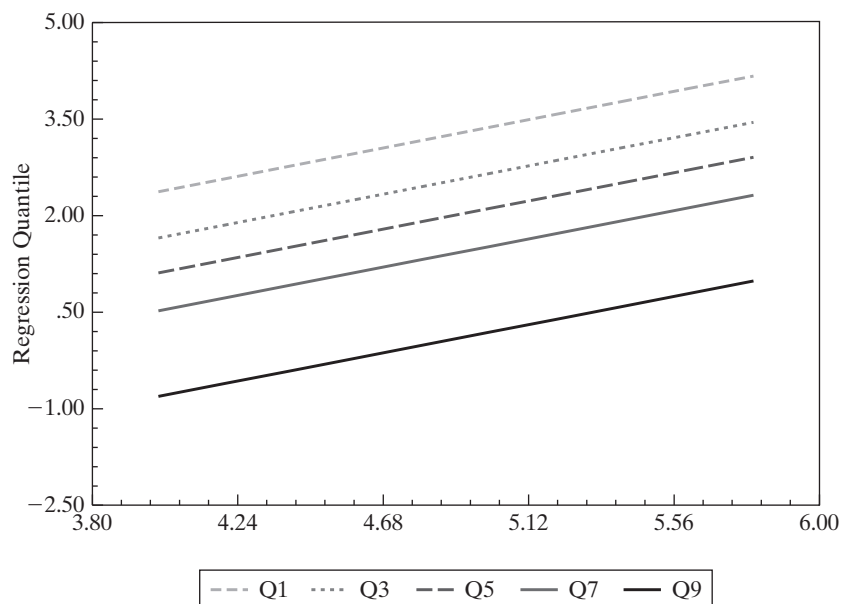
**TABLE 7.5** Estimated Quantile Regression Models

| <i>Quantile</i> | <i>Estimated Parameters</i> |                  |            |                   |
|-----------------|-----------------------------|------------------|------------|-------------------|
|                 | <i>Constant</i>             | <i>ln Income</i> | <i>Age</i> | <i>Dependents</i> |
| 0.1             | -6.73560                    | 1.40306          | -.03081    | -.04297           |
| 0.2             | -4.31504                    | 1.16919          | -.02460    | -.04630           |
| 0.3             | -3.62455                    | 1.12240          | -.02133    | -.04788           |
| 0.4             | -2.98830                    | 1.07109          | -.01859    | -.04731           |
| (Median) 0.5    | -2.80376                    | 1.07493          | -.01699    | -.04995           |
| Std.Error       | (.24564)                    | (.03223)         | (.00157)   | (.01080)          |
| <i>t</i>        | -11.41                      | 33.35            | -10.79     | -4.63             |
| Least Squares   | -3.05581                    | 1.08344          | -.01736    | -.04461           |
| Std.Error       | (.23970)                    | (.03212)         | (.00135)   | (.01092)          |
| <i>t</i>        | -12.75                      | 33.73            | -12.88     | -4.08             |
| 0.6             | -2.05467                    | 1.00302          | -.01478    | -.04609           |
| 0.7             | -1.63875                    | .97101           | -.01190    | -.03803           |
| 0.8             | -.94031                     | .91377           | -.01126    | -.02245           |
| 0.9             | -.05218                     | .83936           | -.00891    | -.02009           |

**FIGURE 7.6** Estimates of Income Elasticity of Expenditure.

terms, the results do seem to be largely consistent with our earlier result of the quantiles largely being differentiated by shifts in the constant term, in spite of the seemingly large change in the coefficient on  $\ln$  Income in the results.

## 210 PART I ♦ The Linear Regression Model



**FIGURE 7.7** Quantile Regressions for Ln Spending.

## 7.4 PARTIALLY LINEAR REGRESSION

The proper functional form in the linear regression is an important specification issue. We examined this in detail in Chapter 6. Some approaches, including the use of dummy variables, logs, quadratics, and so on, were considered as means of capturing nonlinearity. The translog model in particular (Example 2.4) is a well-known approach to approximating an unknown nonlinear function. Even with these approaches, the researcher might still be interested in relaxing the assumption of functional form in the model. The partially linear model [analyzed in detail by Yatchew (1998, 2000) and Härdle, Liang, and Gao (2000)] is another approach. Consider a regression model in which one variable,  $x$ , is of particular interest, and the functional form with respect to  $x$  is problematic. Write the model as

$$y_i = f(x_i) + \mathbf{z}_i' \boldsymbol{\beta} + \varepsilon_i,$$

where the data are assumed to be well behaved and, save for the functional form, the assumptions of the classical model are met. The function  $f(x_i)$  remains unspecified. As stated, estimation by least squares is not feasible until  $f(x_i)$  is specified. Suppose the data were such that they consisted of pairs of observations  $(y_{j1}, y_{j2})$ ,  $j = 1, \dots, n/2$ , in which  $x_{j1} = x_{j2}$  within every pair. If so, then estimation of  $\boldsymbol{\beta}$  could be based on the simple transformed model

$$y_{j2} - y_{j1} = (\mathbf{z}_{j2} - \mathbf{z}_{j1})' \boldsymbol{\beta} + (\varepsilon_{j2} - \varepsilon_{j1}), \quad j = 1, \dots, n/2.$$

As long as observations are independent, the constructed disturbances,  $v_i$  still have zero mean, variance now  $2\sigma^2$ , and remain uncorrelated across pairs, so a classical model applies and least squares is actually optimal. Indeed, with the estimate of  $\boldsymbol{\beta}$ , say,  $\hat{\boldsymbol{\beta}}_d$  in

## CHAPTER 7 ♦ Nonlinear, Semiparametric 211

hand, a noisy estimate of  $f(x_i)$  could be estimated with  $y_i - \mathbf{z}'_i \hat{\beta}_d$  (the estimate contains the estimation error as well as  $\varepsilon_i$ ).<sup>16</sup>

The problem, of course, is that the enabling assumption is heroic. Data would not behave in that fashion unless they were generated experimentally. The logic of the partially linear regression estimator is based on this observation nonetheless. Suppose that the observations are sorted so that  $x_1 < x_2 < \dots < x_n$ . Suppose, as well, that this variable is well behaved in the sense that as the sample size increases, this sorted data vector more tightly and uniformly fills the space within which  $x_i$  is assumed to vary. Then, intuitively, the difference is “almost” right, and becomes better as the sample size grows. [Yatchew (1997, 1998) goes more deeply into the underlying theory.] A theory is also developed for a better differencing of groups of two or more observations. The transformed observation is  $y_{d,i} = \sum_{m=0}^M d_m y_{i-m}$  where  $\sum_{m=0}^M d_m = 0$  and  $\sum_{m=0}^M d_m^2 = 1$ . (The data are not separated into nonoverlapping groups for this transformation—we merely used that device to motivate the technique.) The pair of weights for  $M = 1$  is obviously  $\pm\sqrt{0.5}$ —this is just a scaling of the simple difference, 1,  $-1$ . Yatchew [1998, p. 697] tabulates “optimal” differencing weights for  $M = 1, \dots, 10$ . The values for  $M = 2$  are (0.8090,  $-0.500$ ,  $-0.3090$ ) and for  $M = 3$  are (0.8582,  $-0.3832$ ,  $-0.2809$ ,  $-0.1942$ ). This estimator is shown to be consistent, asymptotically normally distributed, and have asymptotic covariance matrix<sup>17</sup>

$$\text{Asy. Var}[\hat{\beta}_d] = \left(1 + \frac{1}{2M}\right) \frac{\sigma_v^2}{n} E_x[\text{Var}[\mathbf{z} | x]].$$

The matrix can be estimated using the sums of squares and cross products of the differenced data. The residual variance is likewise computed with

$$\hat{\sigma}_v^2 = \frac{\sum_{i=M+1}^n (y_{d,i} - \mathbf{z}'_{d,i} \hat{\beta}_d)^2}{n - M}.$$

Yatchew suggests that the partial residuals,  $y_{d,i} - \mathbf{z}'_{d,i} \hat{\beta}_d$  be smoothed with a kernel density estimator to provide an improved estimator of  $f(x_i)$ . Manzan and Zeron (2010) present an application of this model to the U.S. gasoline market.

### Example 7.11 Partially Linear Translog Cost Function

Yatchew (1998, 2000) applied this technique to an analysis of scale effects in the costs of electricity supply. The cost function, following Nerlove (1963) and Christensen and Greene (1976), was specified to be a translog model (see Example 2.4 and Section 10.5.2) involving labor and capital input prices, other characteristics of the utility, and the variable of interest, the number of customers in the system,  $C$ . We will carry out a similar analysis using Christensen and Greene’s 1970 electricity supply data. The data are given in Appendix Table F4.4. (See Section 10.5.1 for description of the data.) There are 158 observations in the data set, but the last 35 are holding companies which are comprised of combinations of the others. In addition, there are several extremely small New England utilities whose costs are clearly unrepresentative of the best practice in the industry. We have done the analysis using firms 6–123 in the data set. Variables in the data set include  $Q$  = output,  $C$  = total cost, and  $PK$ ,  $PL$ , and  $PF$  = unit cost measures for capital, labor, and fuel, respectively. The parametric model

<sup>16</sup>See Estes and Honoré (1995) who suggest this approach (with simple differencing of the data).

<sup>17</sup>Yatchew (2000, p. 191) denotes this covariance matrix  $E[\text{Cov}[\mathbf{z} | x]]$ .

## 212 PART I ♦ The Linear Regression Model

specified is a restricted version of the Christensen and Greene model,

$$\ln c = \beta_1 k + \beta_2 l + \beta_3 q + \beta_4 (q^2/2) + \beta_5 + \varepsilon,$$

where  $c = \ln[C/(Q \times \text{price})]$ ,  $k = \ln(PK/PF)$ ,  $l = \ln(PL/PF)$ , and  $q = \ln Q$ . The partially linear model substitutes  $f(Q)$  for the last three terms. The division by  $PF$  ensures that average cost is homogeneous of degree one in the prices, a theoretical necessity. The estimated equations, with estimated standard errors, are shown here.

$$\begin{array}{llllll} \text{(parametric)} & c = & -7.32 & + & 0.069k & + & 0.241l - 0.569q + 0.057q^2/2 + \varepsilon, \\ & & (0.333) & & (0.065) & & (0.069) & (0.042) & (0.006) & s = 0.13949 \end{array}$$

$$\begin{array}{llllll} \text{(partially linear)} & c_d = & 0.108k_d & + & 0.163l_d & + & f(q) + v \\ & & (0.076) & & (0.081) & & & & & s = 0.16529 \end{array}$$

## 7.5 NONPARAMETRIC REGRESSION

The regression function of a variable  $y$  on a single variable  $x$  is specified as

$$y = \mu(x) + \varepsilon.$$

No assumptions about distribution, homoscedasticity, serial correlation or, most importantly, functional form are made at the outset;  $\mu(x)$  may be quite nonlinear. Because this is the conditional mean, the only substantive restriction would be that deviations from the conditional mean function are not a function of (correlated with)  $x$ . We have already considered several possible strategies for allowing the conditional mean to be nonlinear, including spline functions, polynomials, logs, dummy variables, and so on. But, each of these is a “global” specification. The functional form is still the same for all values of  $x$ . Here, we are interested in methods that do not assume any particular functional form.

The simplest case to analyze would be one in which several (different) observations on  $y_i$  were made with each specific value of  $x_i$ . Then, the conditional mean function could be estimated naturally using the simple group means. The approach has two shortcomings, however. Simply connecting the points of means,  $(x_i, \bar{y} | x_i)$  does not produce a smooth function. The method would still be assuming something specific about the function between the points, which we seek to avoid. Second, this sort of data arrangement is unlikely to arise except in an experimental situation. Given that data are not likely to be grouped, another possibility is a piecewise regression in which we define “neighborhoods” of points around each  $x$  of interest and fit a separate linear or quadratic regression in each neighborhood. This returns us to the problem of continuity that we noted earlier, but the method of splines, discussed in Section 6.3.1, is actually designed specifically for this purpose. Still, unless the number of neighborhoods is quite large, such a function is still likely to be crude.

Smoothing techniques are designed to allow construction of an estimator of the conditional mean function without making strong assumptions about the behavior of the function between the points. They retain the usefulness of the **nearest neighbor** concept but use more elaborate schemes to produce smooth, well-behaved functions. The general class may be defined by a conditional mean estimating function

$$\hat{\mu}(x^*) = \sum_{i=1}^n w_i(x^* | x_1, x_2, \dots, x_n) y_i = \sum_{i=1}^n w_i(x^* | \mathbf{x}) y_i,$$



## CHAPTER 7 ♦ Nonlinear, Semiparametric 213

where the weights sum to 1. The linear least squares regression line is such an estimator. The predictor is

$$\hat{\mu}(x^*) = a + bx^*.$$

where  $a$  and  $b$  are the least squares constant and slope. For this function, you can show that

$$w_i(x^*|\mathbf{x}) = \frac{1}{n} + \frac{x^*(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

The problem with this particular weighting function, which we seek to avoid here, is that it allows every  $x_i$  to be in the neighborhood of  $x^*$ , but it does not reduce the weight of any  $x_i$  when it is far from  $x^*$ . A number of **smoothing functions** have been suggested that are designed to produce a better behaved regression function. [See Cleveland (1979) and Schimek (2000).] We will consider two.

The locally weighted smoothed regression estimator (“loess” or “lowess” depending on your source) is based on explicitly defining a neighborhood of points that is close to  $x^*$ . This requires the choice of a bandwidth,  $h$ . The **neighborhood** is the set of points for which  $|x^* - x_i|$  is small. For example, the set of points that are within the range  $x^* \pm h/2$  might constitute the neighborhood. The choice of bandwidth is crucial, as we will explore in the following example, and is also a challenge. There is no single best choice. A common choice is **Silverman’s** (1986) **rule of thumb**,

$$h_{Silverman} = \frac{.9[\min(s, IQR)]}{1.349 n^{0.2}}$$

where  $s$  is the sample standard deviation and  $IQR$  is the interquartile range (.75 quantile minus .25 quantile). A suitable weight is then required. Cleveland (1979) recommends the tricube weight,

$$T_i(x^*|\mathbf{x}, h) = \left[ 1 - \left( \frac{|x_i - x^*|}{h} \right)^3 \right]^3.$$

Combining terms, then the weight for the loess smoother is

$$w_i(x^*|\mathbf{x}, h) = 1(x_i \text{ in the neighborhood}) \times T_i(x^*|\mathbf{x}, h).$$

The bandwidth is essential in the results. A wider neighborhood will produce a smoother function, but the wider neighborhood will track the data less closely than a narrower one. A second possibility, similar to the least squares approach, is to allow the neighborhood to be all points but make the weighting function decline smoothly with the distance between  $x^*$  and any  $x_i$ . A variety of **kernel functions** are used for this purpose. Two common choices are the **logistic kernel**,

$$K(x^*|x_i, h) = \Lambda(v_i)[1 - \Lambda(v_i)] \text{ where } \Lambda(v_i) = \exp(v_i)/[1 + \exp(v_i)], v_i = (x_i - x^*)/h,$$

## 214 PART I ♦ The Linear Regression Model

and the Epanechnikov kernel,

$$K(x^*|x_i, h) = 0.75(1 - 0.2 v_i^2)/\sqrt{5} \text{ if } |v_i| \leq 5 \text{ and } 0 \text{ otherwise.}$$

This produces the kernel weighted regression estimator,

$$\hat{\mu}(x^*|\mathbf{x}, h) = \frac{\sum_{i=1}^n \frac{1}{k} K\left[\frac{x_i - x^*}{h}\right] y_i}{\sum_{i=1}^n \frac{1}{k} K\left[\frac{x_i - x^*}{h}\right]},$$

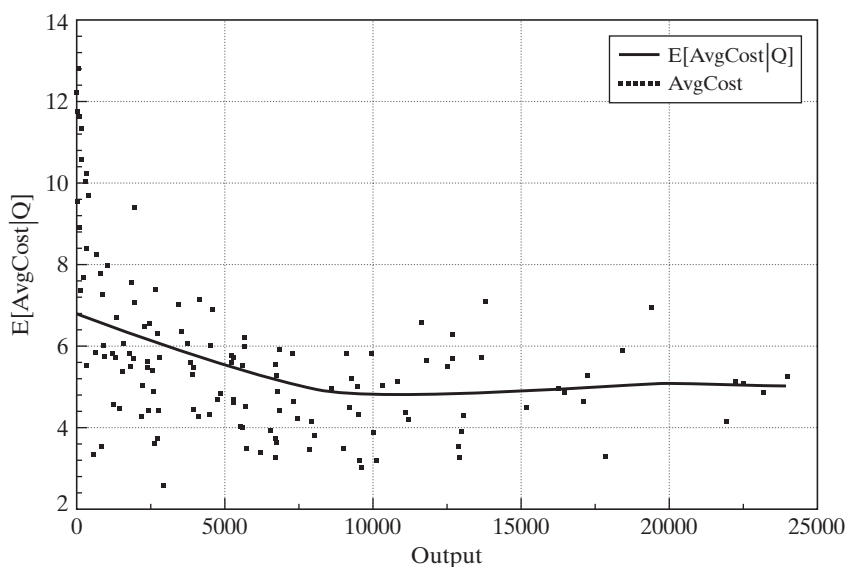
which has become a standard tool in nonparametric analysis.

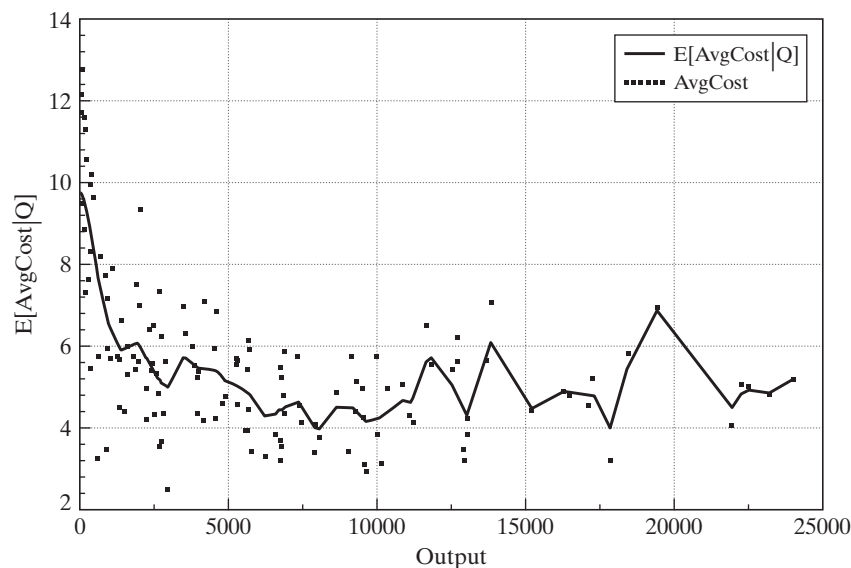
### Example 7.12 A Nonparametric Average Cost Function

In Example 7.11, we fit a partially linear regression for the relationship between average cost and output for electricity supply. Figures 7.8 and 7.9 show the less ambitious nonparametric regressions of average cost on output. The overall picture is the same as in the earlier example. The kernel function is the logistic density in both cases. The function in Figure 7.8 uses a bandwidth of 2,000. Because this is a fairly large proportion of the range of variation of output, the function is quite smooth. The regression in Figure 7.9 uses a bandwidth of only 200. The function tracks the data better, but at an obvious cost. The example demonstrates what we and others have noted often. The choice of bandwidth in this exercise is crucial.

Data smoothing is essentially data driven. As with most nonparametric techniques, inference is not part of the analysis—this body of results is largely descriptive. As can be seen in the example, nonparametric regression can reveal interesting characteristics of the data set. For the econometrician, however, there are a few drawbacks. There is no danger of misspecifying the conditional mean function, for example. But, the great

**FIGURE 7.8** Nonparametric Cost Function.





**FIGURE 7.9** Nonparametric Cost Function.

generality of the approach limits the ability to test one's specification or the underlying theory. [See, for example, Blundell, Browning, and Crawford's (2003) extensive study of British expenditure patterns.] Most relationships are more complicated than a simple conditional mean of one variable. In the Example 7.12, some of the variation in average cost relates to differences in factor prices (particularly fuel) and in load factors. Extensions of the fully nonparametric regression to more than one variable is feasible, but very cumbersome. [See Härdle (1990) and Li and Racine (2007).] A promising approach is the partially linear model considered earlier.

## 7.6 SUMMARY AND CONCLUSIONS

In this chapter, we extended the regression model to a form that allows nonlinearity in the parameters in the regression function. The results for interpretation, estimation, and hypothesis testing are quite similar to those for the linear model. The two crucial differences between the two models are, first, the more involved estimation procedures needed for the nonlinear model and, second, the ambiguity of the interpretation of the coefficients in the nonlinear model (because the derivatives of the regression are often nonconstant, in contrast to those in the linear model).

### Key Terms and Concepts

- Bandwidth
- Bootstrap
- Box-Cox transformation
- Conditional mean function
- Conditional median
- Delta method
- GMM estimator
- Identification condition
- Identification problem
- Incidental parameters problem
- Index function model



## 216 PART I ♦ The Linear Regression Model

- Indirect utility function
- Interaction term
- Iteration
- Jacobian
- Kernel density estimator
- Kernel functions
- Least absolute deviations (LAD)
- Linear regression model
- Linearized regression model
- Lagrange multiplier test
- Logistic kernel
- Logit model
- Loglinear model
- Median regression
- Nearest neighbor
- Neighborhood
- Nonlinear least squares
- Nonlinear regression model
- Nonparametric estimators
- Nonparametric regression
- Normalization
- Orthogonality condition
- Overidentifying restrictions
- Partially linear model
- Pseudoregressors
- Quantile regression
- Roy's identity
- Semiparametric
- Semiparametric estimation
- Silverman's rule of thumb
- Smoothing function
- Starting values
- Two-step estimation
- Wald test

### Exercises

- Describe how to obtain nonlinear least squares estimates of the parameters of the model  $y = \alpha x^\beta + \varepsilon$ .
- Verify the following differential equation, which applies to the Box–Cox transformation:

$$\frac{d^i x^{(\lambda)}}{d\lambda^i} = \left(\frac{1}{\lambda}\right) \left[ x^\lambda (\ln x)^i - \frac{i d^{i-1} x^{(\lambda)}}{d\lambda^{i-1}} \right]. \quad (7-34)$$

Show that the limiting sequence for  $\lambda = 0$  is

$$\lim_{\lambda \rightarrow 0} \frac{d^i x^{(\lambda)}}{d\lambda^i} = \frac{(\ln x)^{i+1}}{i+1}. \quad (7-35)$$

These results can be used to great advantage in deriving the actual second derivatives of the log-likelihood function for the Box–Cox model.

### Applications



- The data in Appendix table F5.3 present 27 statewide observations on value added (output), labor input (labor), and capital stock (capital) for SIC 33 (primary metals). We are interested in determining whether a linear or loglinear production model is more appropriate for these data. Use MacKinnon, White, and Davidson's (1983)  $P_F$  test to determine whether a linear or loglinear production model is preferred.
- Using the Box–Cox transformation, we may specify an alternative to the Cobb–Douglas model as

$$\ln Y = \alpha + \beta_K \frac{(K^\lambda - 1)}{\lambda} + \beta_L \frac{(L^\lambda - 1)}{\lambda} + \varepsilon.$$

Using Zellner and Revankar's data in Appendix Table , estimate  $\alpha$ ,  $\beta_K$ ,  $\beta_L$ , and  $\lambda$  by using the scanning method suggested in Section 11.5.2. (Do not forget to scale  $Y$ ,  $K$ , and  $L$  by the number of establishments.) Use (7-24), (7-15), and (7-16) to compute the appropriate asymptotic standard errors for your estimates. Compute the two output elasticities,  $\partial \ln Y / \partial \ln K$  and  $\partial \ln Y / \partial \ln L$ , at the sample means of  $K$  and  $L$ . (Hint:  $\partial \ln Y / \partial \ln K = K \partial \ln Y / \partial K$ .)

## CHAPTER 7 ♦ Nonlinear, Semiparametric 217

3. For the model in Application 2, test the hypothesis that  $\lambda = 0$  using a Wald test and a Lagrange multiplier test. Note that the restricted model is the Cobb–Douglas loglinear model. The LM test statistic is shown in (7-22). To carry out the test, you will need to compute the elements of the fourth column of  $\mathbf{X}^0$ , the pseudoregressor corresponding to  $\lambda$  is  $\partial E[y | x] / \partial \lambda | \lambda = 0$ . Result (7-35) will be useful.
4. The National Institute of Standards and Technology (NIST) has created a web site that contains a variety of estimation problems, with data sets, designed to test the accuracy of computer programs. (The URL is <http://www.itl.nist.gov/div898/strd/>.) One of the five suites of test problems is a set of 27 nonlinear least squares problems, divided into three groups: easy, moderate, and difficult. We have chosen one of them for this application. You might wish to try the others (perhaps see if the software you are using can solve the problems). This is the Misralc problem (<http://www.itl.nist.gov/div898/strd/nls/data/misralc.shtml>). The nonlinear regression model is

$$y_i = h(x_i, \beta) + \varepsilon_i$$

$$= \beta_1 \left( 1 - \frac{1}{\sqrt{1 + 2\beta_2 x_i}} \right) + \varepsilon_i.$$

The data are as follows:

| Y     | X     |
|-------|-------|
| 10.07 | 77.6  |
| 14.73 | 114.9 |
| 17.94 | 141.1 |
| 23.93 | 190.8 |
| 29.61 | 239.9 |
| 35.18 | 289.0 |
| 40.02 | 332.8 |
| 44.82 | 378.4 |
| 50.76 | 434.8 |
| 55.05 | 477.3 |
| 61.01 | 536.8 |
| 66.40 | 593.1 |
| 75.47 | 689.1 |
| 81.78 | 760.0 |

For each problem posed, NIST also provides the “certified solution,” (i.e., the right answer). For the Misralc problem, the solutions are as follows:

|                                       | <i>Estimate</i>    | <i>Estimated Standard Error</i> |
|---------------------------------------|--------------------|---------------------------------|
| $\beta_1$                             | 6.3642725809E + 02 | 4.6638326572E + 00              |
| $\beta_2$                             | 2.0813627256E – 04 | 1.7728423155E – 06              |
| $\mathbf{e}'\mathbf{e}$               |                    | 4.0966836971E – 02              |
| $s^2 = \mathbf{e}'\mathbf{e}/(n - K)$ |                    | 5.8428615257E – 02              |

Finally, NIST provides two sets of starting values for the iterations, generally one set that is “far” from the solution and a second that is “close” from the solution. For this problem, the starting values provided are  $\beta^1 = (500, 0.0001)$  and  $\beta^2 = (600, 0.0002)$ . The exercise here is to reproduce the NIST results with your software. [For a detailed

## 218 PART I ♦ The Linear Regression Model

analysis of the NIST nonlinear least squares benchmarks with several well-known computer programs, see McCullough (1999).]

5. In Example 7.1, the CES function is suggested as a model for production;

$$\ln y = \ln \gamma - \frac{\nu}{\rho} \ln [\delta K^{-\rho} + (1 - \delta)L^{-\rho}] + \varepsilon. \quad (7-36)$$

Example 6.8 suggested an indirect method of estimating the parameters of this model. The function is linearized around  $\rho = 0$ , which produces an intrinsically linear approximation to the function,

$$\ln y = \beta_1 + \beta_2 \ln K + \beta_3 \ln L + \beta_4 [1/2(\ln K - \ln L)^2] + \varepsilon$$

where  $\beta_1 = \ln \gamma$ ,  $\beta_2 = \nu\delta$ ,  $\beta_3 = \nu(1 - \delta)$  and  $\beta_4 = \rho\nu\delta(1 - \delta)$ . The approximation can be estimated by linear least squares. Estimates of the structural parameters are found by inverting the preceding four equations. An estimator of the asymptotic covariance matrix is suggested using the delta method. The parameters of (7-36) can also be estimated directly using nonlinear least squares and the results given earlier in this chapter.

Christensen and Greene's (1976) data on U.S. electricity generation are given in Appendix Table F4.4. The data file contains 158 observations. Using the first 123, fit the CES production function, using capital and fuel as the two factors of production rather than capital and labor. Compare the results obtained by the two approaches, and comment on why the differences (which are substantial) arise.

The following exercises require specialized software. The relevant techniques are available in several packages that might be in use, such as SAS, Stata, or LIMDEP. The exercises are suggested as departure points for explorations using a few of the many estimation techniques listed in this chapter.

6. Using the gasoline market data in Appendix Table F2.2, use the partially linear regression method in Section 7.4 to fit an equation of the form

$$\ln(G/Pop) = \beta_1 \ln(Income) + \beta_2 \ln P_{new\ cars} + \beta_3 \ln P_{used\ cars} + g(\ln P_{gasoline}) + \varepsilon.$$

7. To continue the analysis in Question 6, consider a nonparametric regression of  $G/Pop$  on the price. Using the nonparametric estimation method in Section 7.5, fit the nonparametric estimator using a range of bandwidth values to explore the effect of bandwidth.