7

NONLINEAR, SEMIPARAMETRIC, AND NONPARAMETRIC REGRESSION MODELS

7.1 INTRODUCTION

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

$$y = x_1\beta_1 + x_2\beta_2 + \cdots + \varepsilon. \tag{7-1}$$

Chapters 2 through 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,

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

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],

$$y = h(x_1, x_2, \ldots, x_P; \beta_1, \beta_2, \ldots, \beta_K) + \varepsilon, \tag{7-3}$$

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 $E[y|\mathbf{x}] = h(\mathbf{x}, \beta)$ for more general functions. This allows a much wider range of functional forms than the linear model can accommodate.[1] 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 4.3, 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 deviations estimator. This section will also relax the restriction that the model coefficients are always the same in the different parts of the distribution

[1] A complete discussion of this subject can be found in Amemiya (1985). Another authoritative treatment is the text by Davidson and MacKinnon (1993).

of y (given $\mathbf{x}$). The LAD estimator estimates the parameters of the conditional median, that is, the 50th 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 $y = f(z) + \mathbf{x}'\boldsymbol{\beta} + \varepsilon$. 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

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

The linear model is obviously a special case. Moreover, some models that 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 are no such transformations.

Example 7.1 CES Production Function

In Example 6.18, we examined a constant elasticity of substitution production function model,

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

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. The underlying model in (7-5) is nonlinear.

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, β,

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

NR1. Functional form: The conditional mean function for y_i given $\mathbf{x}_i$ is

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

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

NR2. 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. Full rank will be necessary, but it is not sufficient.

NR3. Zero conditional 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. We will want to assume that $\mathbf{x}$ is exogenous in this setting, so added to this assumption will be $\mathrm{E}[\varepsilon|\mathbf{x}] = 0$.

NR4. Homoscedasticity and nonautocorrelation: As in the linear model, we assume conditional homoscedasticity,

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

and nonautocorrelation,

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

This assumption parallels the specification of the linear model in Chapter 4. As before, we will want to relax these assumptions.

NR5. Data generating process: The DGP 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 ε_i. The data on $\mathbf{x}_i$ are assumed to be "well behaved."

NR6. Underlying probability model: There is a well-defined probability distribution generating ε_i. At this point, we assume only that this process produces a sample of uncorrelated, identically (marginally) distributed random variables ε_i with mean zero and variance σ^2 conditioned on $h(\mathbf{x}_i, \beta)$. Thus, at this point, our statement of the model is **semiparametric**. (See Section 12.3.) We will not be assuming any particular distribution for ε_i. The conditional moment assumptions in 3 and 4 will be sufficient for the results in this chapter.

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,

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

where V is indirect utility, p_k is the price for the kth 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 kth commodity that is of the form

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

where $\beta_M = \Sigma_k \beta_k$ and $\gamma_{Mj} = \Sigma_k \gamma_{kj}$. 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,

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

The first-order conditions for the minimization are

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

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 the solution to

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

This is the nonlinear regression counterpart to the least squares normal equations in (3-12). Computation requires an iterative solution. (See Example 7.3.) The method is presented in Section 7.2.6.

Assumptions NR1 and NR3 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 ε_i and $\mathbf{x}_i$, are uncorrelated. However, *uncorrelatedness* of ε_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 ε_i is strictly uncorrelated with *any prior information* in the model,

including previous disturbances, then 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 ε_i and $\mathbf{x}_i$ would be sufficient for uncorrelatedness of ε_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 ith observation will be

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

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,

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

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.[2]

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

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

These equations do not have an explicit solution.

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

[2]See Ruud (2000, p. 540).

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 these 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 [defined in (7-29)] when they are computed at the true parameter values. Therefore, for the nonlinear regression model, the analog to (4-19) is

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

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

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

This is the orthogonality condition noted earlier in (4-21). 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

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

With these in hand, the asymptotic properties of the nonlinear least squares estimator are essentially those we have already seen for the linear model, except that in this case we place the derivatives of the linearized function evaluated at $\boldsymbol{\beta}$, $\mathbf{X}^0$, in the role of the regressors.[3]

[3]See Amemiya (1985).

The nonlinear least squares criterion function is

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

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

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

In the linear model of Chapter 3, this produces a set of linear normal equations, (3-12). In this more general case, (7-14) is a set of nonlinear equations that do not have an explicit solution. Note that σ^2 is not relevant to the solution. At the solution,

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

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 $\boldsymbol{\beta}$ is compact (has no gaps or nonconcave regions).*
- ***b.*** *For any vector $\boldsymbol{\beta}^0$ in that parameter space, plim $(1/n)S(\boldsymbol{\beta}^0) = q(\boldsymbol{\beta}^0)$, a continuous and differentiable function.*
- ***c.*** *If $q(\boldsymbol{\beta}^0)$ has a unique minimum at the true parameter vector, $\boldsymbol{\beta}$, 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 $(1/n)S(\boldsymbol{\beta}^0)$. If $(1/n)S(\boldsymbol{\beta}^0)$ 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 $q(\boldsymbol{\beta}^0)$ is uniquely $\boldsymbol{\beta}$. If the minimum value of plim $(1/n)S(\boldsymbol{\beta}^0)$ 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 $\text{plim}(1/n)\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{0}$. To follow that approach here, we would use the linearized model and take essentially the same result. The loose 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 $\text{plim}(1/n)\mathbf{X}^{0\prime}\boldsymbol{\delta} = \mathbf{0}$ where $\delta_i = h(\mathbf{x}_i, \boldsymbol{\beta})$ minus the Taylor series approximation.[4]

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

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

The sample estimator of the asymptotic covariance matrix is

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

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.

[4]An argument to this effect appears in Mittelhammer et al. (2000, pp. 190–191).

A consistent estimator of σ^2 is based on the residuals,

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

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 σ^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 this is the case for the program they are using. With this in mind, 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,

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

is no longer guaranteed to be in the range of 0 to 1. It does, however, provide a useful descriptive measure. An intuitively appealing measure of the fit of the model to the data will be the squared correlation between the fitted and actual values, $h(\mathbf{x}_i,\mathbf{b})$ and y_i. This will differ from R^2, partly because the mean prediction will not equal the mean of the observed values.

7.2.4 ROBUST COVARIANCE MATRIX ESTIMATION

Theorem 7.2 relies on assumption NR4, homoscedasticity and nonautocorrelation. We considered two generalizations in the linear case, heteroscedasticity and autocorrelation due to clustering in the sample. The counterparts for the nonlinear case would be based on the linearized model,

$$\begin{aligned} y_i &= x_i^{0\prime}\boldsymbol{\beta} + [h(\mathbf{x}_i, \boldsymbol{\beta}) - \mathbf{x}_i^{0\prime}\boldsymbol{\beta}] + \varepsilon_i \\ &= \mathbf{x}_i^{0\prime}\boldsymbol{\beta} + u_i. \end{aligned}$$

The counterpart to (4-37) that accommodates unspecified heteroscedasticity would then be

$$Est.Asy.Var[\mathbf{b}] = (\mathbf{X}^{0\prime}\mathbf{X}^0)^{-1}\left[\sum_{i=1}^{n}\mathbf{x}_i^0\mathbf{x}_i^{0\prime}(y_i - h(\mathbf{x}_i, \mathbf{b}))^2\right](\mathbf{X}^{0\prime}\mathbf{X}^0)^{-1}.$$

Likewise, to allow for clustering, the computation would be analogous to (4-41) and (4-42),

$$Est.Asy.Var[\mathbf{b}] = \frac{C}{C-1}(\mathbf{X}^{0\prime}\mathbf{X}^0)^{-1}\left[\sum_{c=1}^{C}\left\{\sum_{i=1}^{N_c}\mathbf{x}_i^0 e_i\right\}\left\{\sum_{i=1}^{N_c}\mathbf{x}_i^0 e_i\right\}'\right](\mathbf{X}^{0\prime}\mathbf{X}^0)^{-1}.$$

Note that the residuals are computed as $e_i = y_i - h(\mathbf{x}_i, \mathbf{b})$ using the conditional mean function, not the linearized regression.

7.2.5 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.

The hypothesis to be tested is

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

where $\mathbf{c}(\boldsymbol{\beta})$ 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**. In formal terms, if the original parameter vector has K free elements, then the hypothesis $\mathbf{c}(\boldsymbol{\beta}) - \mathbf{q}$ 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**,

$$\mathbf{R}(\boldsymbol{\beta}) = \partial \mathbf{c}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}', \tag{7-19}$$

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 5.5, 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.[5] 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 statistic is by far the simplest to compute. Of the 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

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

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.

[5] This computational problem may be extremely difficult in its own right, especially if the constraints are nonlinear. We assume that the estimates have been obtained by whatever means are necessary.

The Wald test is based on the distance between $\mathbf{c}(\mathbf{b})$ 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

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

where

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

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 is more than one equivalent way to specify $\mathbf{c}(\boldsymbol{\beta}) = \mathbf{q}$, 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. For the nonlinear regression model, the test has a particularly appealing form.[6] 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

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

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.

7.2.6 APPLICATIONS

This section will present two 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

[6] This test is derived in Judge et al. (1985). Discussion appears in Mittelhammer et al. (2000).

than the linear regression—it has the linear and loglinear models as special cases. Finally, Example 7.6 in the next section 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 6.5.2 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 β_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 model analyzed at the beginning of Chapter 2 is a restricted version of the more general function

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

in which γ equals 1. With this restriction, the model is linear. If γ 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. 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 σ^2. (To make this comparable to s^2 in the linear model, the value includes the degrees of freedom correction.)

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 $\mathbf{D}^{-1}\mathbf{X}_0'\mathbf{X}_0\mathbf{D}^{-1}$ where $x_1^0 = 1$ and $\mathbf{D}$ is the diagonal matrix containing the square roots of $\mathbf{x}_k^{0\prime}\mathbf{x}_k^0$ on the diagonal.) Recall that 20 was the benchmark for a problematic data set. By the standards discussed in

TABLE 7.1 Estimated Consumption Functions

	Linear Model		*Nonlinear Model*	
Parameter	***Estimate***	***Standard Error***	***Estimate***	***Standard Error***
α	−80.3547	14.3059	458.7990	22.5014
β	0.9217	0.003872	0.10085	0.01091
γ	1.0000	–	1.24483	0.01205
$\mathbf{e}'\mathbf{e}$	1,536,321.881		504,403.1725	
σ	87.20983		50.0946	
R^2	0.996448		0.998834	
Est.Var[b]	–		0.000119037	
Est.Var[c]	–		0.00014532	
Est.Cov[b,c]	–		−0.000131491	

Sections 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 γ 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 value of 1.96 for the 5% significance level, and we thus reject the linear model in favor of the nonlinear regression. The three procedures for testing hypotheses produce the same conclusion.

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

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

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

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 α and β and 1 for γ so that $\mathbf{x}_i{}^* = [1, Y, \beta Y \ln Y]$. The residuals are the least squares residuals computed from the linear regression.

Example 7.5 The Box–Cox Transformation

The **Box–Cox transformation** is used as a device for generalizing the linear model.[7] 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 λ equals zero, the transformation is, by L'Hôpital's rule,

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

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

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

is a linear regression that can be estimated by least squares. However, if λ 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 λ, but, in most applications, this level of generality becomes excessively cumbersome, and λ is assumed to be the same for all the variables in the model.[8] To be defined for all values of λ, x must be strictly positive. In most applications, some of the regressors—for example, a dummy

[7]Box and Cox (1964); Zarembka (1974).

[8]See, for example, Seaks and Layson (1983).

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

Nonlinear least squares is straightforward. In most instances, we can expect to find the least squares value of λ between -2 and 2. Typically, then, λ is estimated by scanning this range for the value that minimizes the sum of squares. Once the optimal value of λ is located, the least squares estimates, the mean squared residual, and this value of λ constitute the nonlinear least squares estimates of the parameters. The optimal value of $\hat{\lambda}$ is an estimate of an unknown parameter. The least squares standard errors will always underestimate the correct asymptotic standard errors if $\hat{\lambda}$ is treated as if it were a known constant.[9] To get the appropriate values, we need the pseudoregressors,

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

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 $\partial h(.)/\partial \lambda$. If $x_k = 0$, then this matrix cannot be computed.

The coefficients in a nonlinear model are not equal to the slopes (or the elasticities) with respect to the variables. For the Box–Cox model $\ln Y = \alpha + \beta X^{(\lambda)} + \epsilon$,

$$\frac{\partial E[\ln y \,|\, \mathbf{x}]}{\partial \ln x} = x \frac{\partial E[\ln y \,|\, \mathbf{x}]}{\partial x} = \beta x^{\lambda} = \eta.$$

A standard error for this estimator can be obtained using the **delta method**. The derivatives are $\partial \eta / \partial \beta = x^{\lambda} = \eta / \beta$ and $\partial \eta / \partial \lambda = \eta \ln x$. Collecting terms, we obtain

$$Asy.Var[\hat{\eta}] = (\eta/\beta)^2 \left\{ Asy.Var[\hat{\beta}] + (\beta \ln x)^2 \, Asy.Var[\hat{\lambda}] + (2\beta \ln x) \, Asy.Cov[\hat{\beta}, \hat{\lambda}] \right\}.$$

7.2.7 LOGLINEAR MODELS

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 α^0 is a constant term and θ is an additional parameter such that

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

(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,

[9] See Fomby, Hill, and Johnson (1984, pp. 426–431).

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

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

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

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

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

The parameter θ 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. However, since the model is nonlinear, the coefficients do not give the magnitudes of the interesting effects in the equation. In particular, for this model,

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

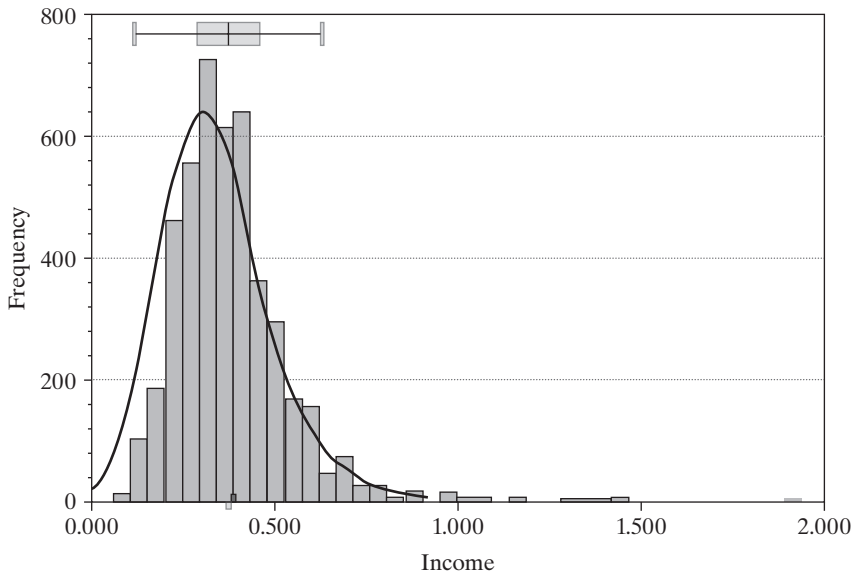
The implication is that the analyst must be careful in interpreting the estimation results, as interest usually focuses on partial effects, not coefficients.

Example 7.6 Interaction Effects in a Loglinear Model for Income

In *Incentive Effects in the Demand for Health Care: A Bivariate Panel Count Data Estimation*, Riphahn, Wambach, and Million (2003) 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.[10] Among the variables reported in the panel are household income, with numerous

[10]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.

FIGURE 7.1 Histogram and Kernel Density Estimate for Income.



other sociodemographic variables such as age, gender, and education. For this example, we will model the distribution of income using the 1988 wave of the data set, 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. Figure 7.1 displays a histogram and a kernel density estimator for the household income variable for these observations. Table 7.2 provides descriptive statistics for the exogenous variables used in this application.

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

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

As we have constructed the model, the derivative result, $\partial E[y|\mathbf{x}]/\partial x_k = \beta_k \exp(\alpha + \mathbf{x}'\boldsymbol{\beta})$, must be modified 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

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

TABLE 7.2 Descriptive Statistics for Variables Used in Nonlinear Regression

Variable	*Mean*	*Std.Dev.*	*Minimum*	*Maximum*
Income	0.344896	0.164054	0.0050	2
Age	43.4452	11.2879	25	64
Educ	11.4167	2.36615	7	18
Female	0.484267	0.499808	0	1

Another consideration is how to compute the partial effects, as sample averages or at the means of the variables. For example, $\partial E[y|\mathbf{x}]/\partial Age = E[y|\mathbf{x}] \times (\beta_2 + 2\beta_3 Age + \beta_7 Educ)$. We will estimate the average partial effects by averaging these values over the sample observations. 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.

The average value of *Age* in the sample is 43.4452 and the average value of *Education* is 11.4167. The partial effect of a year of education is estimated to be 0.015736 if it is computed by computing the partial effect for each individual and averaging the results. 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 4.5% (i.e., 0.015736/0.35).

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 7.2 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% between ages 25 and 48. 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 7.2 shows precisely the opposite. The difference is accounted for by the interaction term, *Female* × *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.

TABLE 7.3 Estimated Regression Equations

	Nonlinear Least Squares			*Linear Least Squares*	
Variable	***Estimate***	***Std. Error***	***t Ratio***	***Estimate***	***Projection***
Constant	−2.58070	0.17455	14.78	−0.13050	0.10746
Age	0.06020	0.00615	9.79	0.01791	0.00066
Age^2	−0.00084	0.00006082	−13.83	−0.00027	
Education	−0.00616	0.01095	−0.56	−0.00281	0.01860
Female	0.17497	0.05986	2.92	0.07955	0.00075
Female × *Educ*	−0.01476	0.00493	−2.99	−0.00685	
Age × *Educ*	0.00134	0.00024	5.59	0.00055	
e'e		106.09825		106.24323	
s		0.15387		0.15410	
R^2		0.12005		0.11880	

FIGURE 7.2 Expected Incomes vs. Age for Men and Women with EDUC = 16.

Finally, in addition to the quadratic term in age, the model contains an interaction term, *Age* × *Education*. The coefficient is positive and highly significant. But, it is not obvious how this should be interpreted. In a linear model,

$$Income = \beta_1 + \beta_2 Age + \beta_3 Age^2 + \beta_4 Education + \beta_5 Female + \beta_6 Female \times Education + \beta_7 Age \times Education + \varepsilon,$$

we would find that $\beta_7 = \partial^2 E[Income|x]/\partial\, Age\, \partial Education$. 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 β_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, β_7 is indeed zero. For convenience, let $\mu(x)$ equal the conditional mean function. Then, the partial effect with respect to *Age* is

$$\partial\mu(x)/\partial Age = \mu(x) \times (\beta_2 + 2\beta_3 Age),$$

and

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

which is nonzero even if there is no **interaction term** in the model. The interaction effect in the model that includes the product term, $\beta_7 Age \times Education$, is

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

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:

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

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 β_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 (2010b).

Section 4.4.5 considered the linear projection as a feature of the joint distribution of y and $\mathbf{x}$. It was noted that, assuming the conditional mean function in the joint distribution is $E[y|\mathbf{x}] = \mu(\mathbf{x})$, then the slopes of linear projection, $\gamma = [E\{\mathbf{x}\mathbf{x}'\}]^{-1}E[\mathbf{x}y]$, might resemble the slopes of $\mu(\mathbf{x})$, $\boldsymbol{\delta} = \partial\mu(\mathbf{x})/\partial\mathbf{x}$ at least for some $\mathbf{x}$. In a loglinear, single-index function model such as the one analyzed here, this would relate to the linear least squares regression of y on $\mathbf{x}$. Table 7.4 reports two sets of least squares regression coefficients. The ones on the right show the regression of *Income* on all of the first- and second-order terms that appear in the conditional mean. This would not be the projection of y on $\mathbf{x}$. At best it might be seen as an approximation to $\mu(\mathbf{x})$. The rightmost coefficients report the projection. Both results 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. Table 7.4 reports the results of the computations. The average partial effects for the nonlinear regression are obtained by computing the derivatives for each observation and averaging the results. For the linear approximation, the derivatives are linear functions of the variables, so the average partial effects are simply computed at the means of the variables. Finally, the coefficients of the linear projection are immediate estimates of the partial effects. We find, for example, the partial effect of education in the nonlinear model is 0.01574. Although the linear least squares coefficients are very different, if the partial effect for education is computed for the linear approximation the result of 0.01789 is reasonably close, and results from the fact that in the center of the data, the exponential function is passably linear. The linear projection is less effective at reproducing the partial effects. The comparison for the other variables is mixed. The conclusion from Example 4.4 is unchanged. The substantive comparison here would be between the slopes of the nonlinear regression and the slopes of the linear projection. They resemble each other, but not as closely as one might hope.

TABLE 7.4 Estimated Partial Effects

Variable	*Nonlinear Regression*	*Linear Approximation*	*Linear Projection*
Age	0.00095	0.00091	0.00066
Educ	0.01574	0.01789	0.01860
Female	0.00084	0.00135	0.00075

Example 7.7 Generalized Linear Models for the Distribution of Healthcare Costs

Jones, Lomas, and Rice (2014, 2015) examined the distribution of healthcare costs in the UK. Two aspects of the analysis were different from our examinations to this point. First, while nearly all of the development we have considered so far involves regression, that is, the conditional mean (or median) of the distribution of the dependent variable, their interest was in other parts of the distribution, specifically conditional and unconditional tail probabilities for relatively outlying parts of the distribution. Second, the variable under study is nonnegative, highly asymmetric (skewness 13.03), and leptokurtic (kurtosis 363.13−the distribution has a thick right tail). Some values from the estimated survival function (Jones et al., 2015, Table 1) are $S(£500) = 0.8296$, $S(£1{,}000) = 0.5589$, $S(£5{,}000) = 0.1383$, and $S(£10{,}000) = 0.0409$. The skewness and kurtosis values would compare to 0.0 and 3.0, respectively, for the normal distribution. The survival function values for the normal distribution with this mean and standard deviation would be 0.6608, 0.6242, 0.3193, and 0.0732, respectively. The model is constructed with these features of the data in mind. Several methods of fitting the distribution were examined, including a set of nine parametric models. Several of these were special cases of the *generalized beta of the second kind*. The functional forms are *generalized linear models* constructed from a *family* of distributions, such as the normal or exponential, and a link function, $g(\mathbf{x}'\boldsymbol{\beta})$ such that $\text{link}(g(\mathbf{x}'\boldsymbol{\beta})) = \mathbf{x}'\boldsymbol{\beta}$. Thus, if the link function is "ln" (log link), then $g(\mathbf{x}'\boldsymbol{\beta}) = \exp(\mathbf{x}'\boldsymbol{\beta})$. Among the nine special cases examined are

- Gamma family, log link:

$$f(cost\,|\,\mathbf{x}) = \frac{[g(\mathbf{x}'\boldsymbol{\beta})]^{-P}}{\Gamma(P)}\exp[-cost/g(\mathbf{x}'\boldsymbol{\beta})]cost^{P-1},$$

$$g(\mathbf{x}'\boldsymbol{\beta}) = \exp(\mathbf{x}'\boldsymbol{\beta});\ E[cost\,|\,\mathbf{x}] = Pg(\mathbf{x}'\boldsymbol{\beta}).$$

- Lognormal family, identity link:

$$f(cost\,|\,\mathbf{x}) = \frac{1}{\sigma cost\sqrt{2\pi}}\exp\left[-\frac{(\ln cost - g(\mathbf{x}'\boldsymbol{\beta}))^2}{2\sigma^2}\right],$$

$$g(\mathbf{x}'\boldsymbol{\beta}) = \mathbf{x}'\boldsymbol{\beta};\ E[cost\,|\,\mathbf{x}] = \exp\left[g(\mathbf{x}'\boldsymbol{\beta}) + \tfrac{1}{2}\sigma^2\right].$$

- Finite mixture of two gammas, inverse square root link:

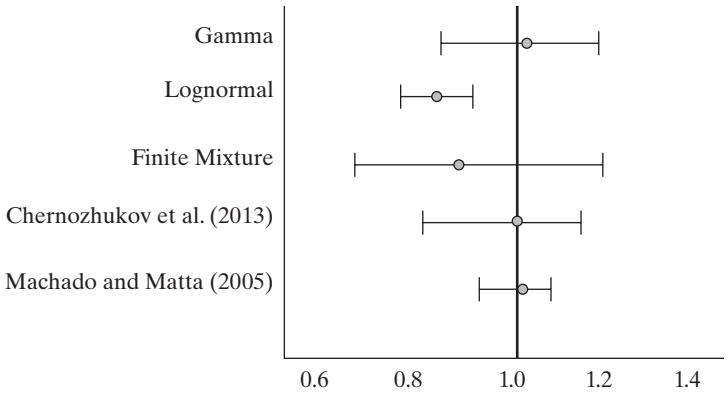
$$f(cost\,|\,\mathbf{x}) = \sum_{j=1}^{2}\alpha_j\frac{[g(\mathbf{x}'\boldsymbol{\beta}_j)]^{-P_j}}{\Gamma(P_j)}\exp[-cost/g(\mathbf{x}'\boldsymbol{\beta}_j)]cost^{P_j-1},\ 0 \le \alpha_j \le 1,\ \sum_{j=1}^{2}\alpha_j = 1,$$

$$g(\mathbf{x}'\boldsymbol{\beta}) = 1/(\mathbf{x}'\boldsymbol{\beta})^2;\ E[cost\,|\,\mathbf{x}] = \alpha_1P_1g(\mathbf{x}'\boldsymbol{\beta}_1) + \alpha_2P_2g(\mathbf{x}'\boldsymbol{\beta}_2).$$

(The models have been reparameterized here to simplify them and show their similarities.) In each case, there is a conditional mean function. However, the quantity of interest in the study is not the regression function; it is the survival function, $S(cost\,|\,\mathbf{x},k) = \text{Prob}(cost \ge k\,|\,\mathbf{x})$. The measure of a model's performance is its ability to estimate the sample survival rate for values of k; the one of particular interest is the largest, $k = 10{,}000$. The main interest is the marginal rate, $E_{\mathbf{x}}[S(cost\,|\,\mathbf{x},k)] = \int_{\mathbf{x}} S(cost\,|\,\mathbf{x},k)f(\mathbf{x})d\mathbf{x}$. This is estimated by estimating $\boldsymbol{\beta}$ and the ancillary parameters of the specific model, then estimating $S(cost\,|\,k)$ with $(1/n)\Sigma_{i=1}^{n}S(cost_i\,|\,\mathbf{x}_i,k:\hat{\boldsymbol{\beta}})$ The covariates include a set of morbidity characteristics and an interacted cubic function of age and sex. Several semiparametric and nonparametric methods are examined along with the parametric regression–based models. Figure 7.3 shows the bias and variability of the three parametric estimators and two of the proposed semiparametric methods.[11] Overall, none of the 14 methods examined emerges as best overall by a set of fitting criteria that includes bias and variability.

[11]Derived from the results in Figure 4 in Jones et al. (2015).

FIGURE 7.3 Performance of Several Estimators of $S(cost|k)$.



7.2.8 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, $\boldsymbol{\beta}^0$,

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

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

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

Let x_k^0 equal the kth partial derivative,[12] $\partial h(\mathbf{x}, \boldsymbol{\beta}^0)/\partial \beta_k^0$. 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

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

[12] You should verify that for the linear regression model, these derivatives are the independent variables.

which implies that

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

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

$$y^0 = y - h^0 + \mathbf{x}^{0\prime}\boldsymbol{\beta}^0 = \mathbf{x}^{0\prime}\boldsymbol{\beta} + \varepsilon^0. \tag{7-30}$$

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

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

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 could 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.

Example 7.8 Linearized Regression

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

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

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

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

can be linearly regressed on the three pseudoregressors to estimate β_1, β_2, and β_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 $\boldsymbol{\beta}^0$, 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

$$\begin{aligned}
\mathbf{b}_{t+1} &= \left[\sum_{i=1}^{n} \mathbf{x}_i^0 \mathbf{x}_i^{0\prime} \right]^{-1} \left[\sum_{i=1}^{n} \mathbf{x}_i^0 \left(y_i - h_i^0 + \mathbf{x}_i^{0\prime} \mathbf{b}_t \right) \right] \\
&= \mathbf{b}_t + \left[\sum_{i=1}^{n} \mathbf{x}_i^0 \mathbf{x}_i^{0\prime} \right]^{-1} \left[\sum_{i=1}^{n} \mathbf{x}_i^0 \left(y_i - h_i^0 \right) \right] \\
&= \mathbf{b}_t + (\mathbf{X}^{0\prime}\mathbf{X}^0)^{-1} \mathbf{X}^{0\prime} \mathbf{e}^0 \\
&= \mathbf{b}_t + \boldsymbol{\Delta}_t,
\end{aligned} \tag{7-32}$$

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 $\mathbf{X}^{0\prime}\mathbf{e}^0$ 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 $\mathbf{X}^{0\prime}\mathbf{e}^0$ exactly equal to zero. A useful, scale-free counterpart to the convergence criterion discussed in Section E.3.6 is $\delta = \mathbf{e}^{0\prime}\mathbf{X}^0(\mathbf{X}^{0\prime}\mathbf{X}^0)^{-1}\mathbf{X}^{0\prime}\mathbf{e}^0$. [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.[13] In the absence of information about starting values, a workable strategy is to try the Gauss–Newton iteration first. If it 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 7.9 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

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

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 $\boldsymbol{\beta}$ 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 α and β at the linear least squares values that would result in the special case of $\gamma = 1$ and use 1 for the starting value for γ. The **iterations** are begun at the least squares estimates for α and β and 1 for γ.

[13] See McCullough and Vinod (1999).

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, δ, 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 $\text{Med}[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$. Under these assumptions, least squares remains a natural choice for estimation of $\boldsymbol{\beta}$. But, as we explored in Example 4.3, **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

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

This result suggests a motivation for LAD in its own right, rather than as a robust (to outliers) alternative to least squares.[14] 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

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

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

[14] In Example 4.3, we considered the possibility that in small samples with possibly thick-tailed disturbance distributions, the LAD estimator might have a smaller variance than least squares.

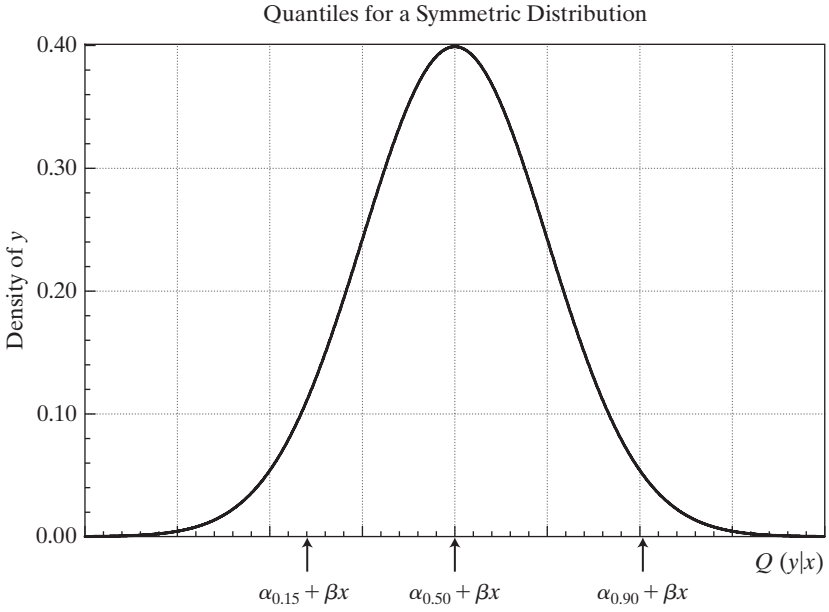
is semiparametric—it requires a much less detailed specification of the distribution of $y|\mathbf{x}$. In the simplest linear model with fixed coefficient vector, $\boldsymbol{\beta}$, the quantiles of $y|\mathbf{x}$ would be defined by variation of the constant term. The implication of the model is shown in Figure 7.4. For a fixed β 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 α_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 α_q in Figure 7.4.

7.3.1 LEAST ABSOLUTE DEVIATIONS ESTIMATION

Least squares can be 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.[15] 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

FIGURE 7.4 Quantile Regression Model.



[15]For some applications, see Taylor (1974), Amemiya (1985, pp. 70–80), Andrews (1974), Koenker and Bassett (1978), Li and Racine (2007), Henderson and Parmeter (2015), and a survey written at a very accessible level by Birkes and Dodge (1993). A somewhat more rigorous treatment is given by Hardle (1990).

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 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), Koenker and Hallock (2001), Huber (1967), and Rogers (1993) have analyzed this regression.[16] Their results suggest an estimator for the asymptotic covariance matrix of the quantile regression estimator,

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

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.[17] [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 σ^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.[18] 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 4.3, the results begin overwhelmingly to favor least squares. For better or

[16]Powell (1984) has extended the LAD estimator to produce a robust estimator for the case in which data on the dependent variable are censored, that is, when negative values of y_i are recorded as zero. See 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).

[17]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 with the author.)

[18]See Section 12.4 and, for example, Johnston and DiNardo (1997, pp. 370–375).

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), $K[.]$ is a weighting, or kernel function, and e_i, $i = 1, \ldots, n$ is the set of residuals. There are no hard and fast rules for choosing h; one popular choice is that used by Stata (2014), $h = .9s/n^{1/5}$. 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:

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

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

Example 7.10 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 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 7.5 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, $0.25/\hat{f}^2(0) = 0.25/1.5467^2 = 0.104502$. 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 a semiparametric 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

FIGURE 7.5 Standardized Residuals for a Production Function.

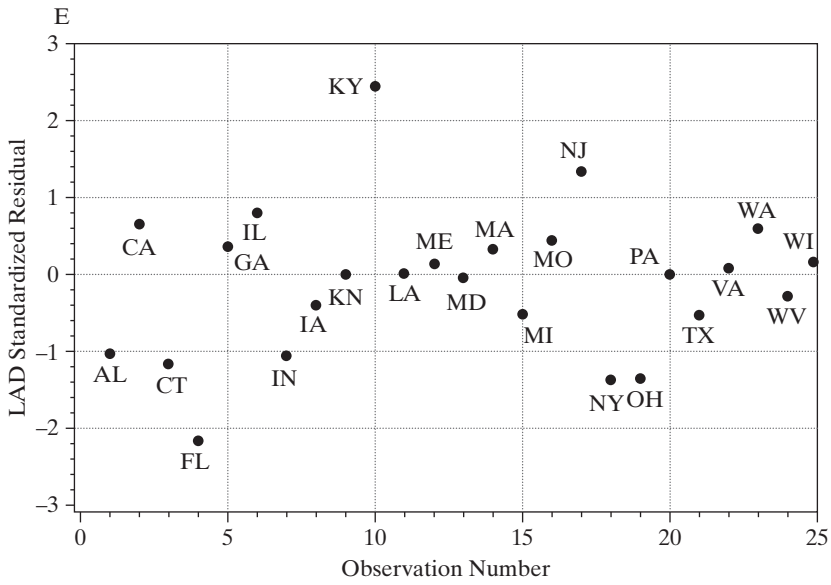


TABLE 7.5 LS and LAD Estimates of a Production Function

	Least Squares			*LAD*				
					Bootstrap		*Kernel Density*	
Coefficient	*Estimate*	*Standard Error*	*t Ratio*	*Estimate*	*Standard Error*	*t Ratio*	*Standard Error*	*t Ratio*
Constant	2.293	0.107	21.396	2.275	0.202	11.246	0.183	12.374
β_k	0.279	0.081	3.458	0.261	0.124	2.099	0.138	1.881
β_l	0.927	0.098	9.431	0.927	0.121	7.637	0.169	5.498
Σe^2	0.7814			0.7984				
$\Sigma \lvert e \rvert$	3.3652			3.2541				

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 distribution.

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

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

where

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

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.[19]

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.[20] 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

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

where

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

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 ε_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 7.11 Quantile Regression for Smoking Behavior

Laporte, Karimova, and Ferguson (2010) employed Becker and Murphy's (1988) model of rational addiction to study the behavior of a sample of Canadian smokers. The rational addiction model is a model of inter-temporal optimization, meaning that, rather than making independent decisions about how much to smoke in each period, the individual plots out an optimal lifetime smoking trajectory, conditional on future values of exogenous variables such as price. The optimal control problem which yields that trajectory incorporates the individual's attitudes to the harm smoking can do to her health and the rate at which she will trade the present against the future. This means that factors like the individual's degree of myopia are built into the trajectory of cigarette consumption which she will follow, and that consumption trajectory is what yields the forward-looking second-order difference equation which characterizes rational addiction behavior.[21]

The proposed empirical model is a dynamic regression,

$$C_t = \alpha + \mathbf{x}_t'\boldsymbol{\beta} + \gamma_1 C_{t+1} + \gamma_0 C_{t-1} + \varepsilon_t.$$

[19]See Koenker and D'Oray (1987) and Koenker (2005).

[20]See Buchinsky (1998).

[21]Laporte et al., p. 1064.

If it is assumed that $\mathbf{x}_t$ is fixed at $\mathbf{x}_*$ and ε_t is fixed at its expected value of zero, then a long run equilibrium consumption occurs where $C_t = C_{t-1} = C^*$ so that

$$C^* = \frac{\alpha + \mathbf{x}_*'\boldsymbol{\beta}}{1 - \gamma_1 - \gamma_0}.$$

(Some restrictions on the coefficients must hold for a finite positive equilibrium to exist. We can see, for example, $\gamma_0 + \gamma_1$ must be less than one.) The long run partial effects are then $\partial C^*/\partial x_{*k} = \beta_k/(1 - \gamma_0 - \gamma_1)$. Various covariates enter the model including, gender, whether smoking is restricted in the workplace, self-assessment of poor diet, price, and whether the individual jumped to zero consumption.

The analysis in the study is done primarily through graphical descriptions of the quantile regressions. Figure 7.6 (Figure 4 from the article) shows the estimates of the coefficient on a gender dummy variable in the model. The center line is the quantile-based coefficient on the dummy variable. The bands show 95% confidence intervals. (The authors do not mention how the standard errors are computed.) The dotted horizontal line shows the least squares estimate of the same coefficient. Note that it coincides with the 50th quantile estimate of this parameter.

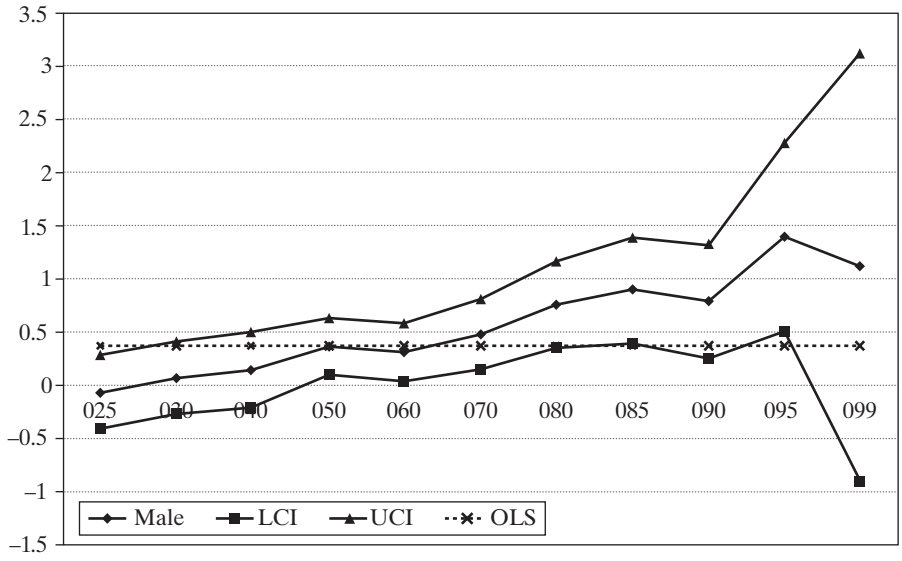
Example 7.12 Income Elasticity of Credit Card Expenditures

Greene (1992, 2007c) analyzed the default behavior and monthly expenditure behavior of a 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 quantile regression approach might be based on

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

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

FIGURE 7.6 Male Coefficient in Quantile Regressions.



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.[22] 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 7.6 presents least squares estimates of the coefficients of the conditional mean function as well as full results for several quantiles.[23] Standard errors are shown for the least squares and median ($q = 0.5$) results. 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 $(1 - 1.08344)/0.03212 = 2.60$. 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 7.7 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 7.8 shows the implied quantile regressions for $q = 0.1$, 0.3, 0.5, 0.7, and 0.9.

TABLE 7.6 Estimated Quantile Regression Models

	Estimated Parameters			
Quantile	***Constant***	***ln Income***	***Age***	***Dependents***
0.1	−6.73560	1.40306	−0.03081	−0.04297
0.2	−4.31504	1.16919	−0.02460	−0.04630
0.3	−3.62455	1.12240	−0.02133	−0.04788
0.4	−2.98830	1.07109	−0.01859	−0.04731
(Median) 0.5	−2.80376	1.07493	−0.01699	−0.04995
Std.Error	(0.24564)	(0.03223)	(0.00157)	(0.01080)
t	−11.41	33.35	−10.79	−4.63
Least Squares	−3.05581	1.08344	−0.01736	−0.04461
Std.Error	(0.23970)	(0.03212)	(0.00135)	(0.01092)
t	−12.75	33.73	−12.88	−4.08
0.6	−2.05467	1.00302	−0.01478	−0.04609
0.7	−1.63875	0.97101	−0.01190	−0.03803
0.8	−0.94031	0.91377	−0.01126	−0.02245
0.9	−0.05218	0.83936	−0.00891	−0.02009

[22]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 true zeros and not missing data. A more detailed approach—the one used in the 1992 study—would model separately the presence or absence of the observation on spending and then model spending conditionally on acceptance of the application. We will revisit this issue in Chapter 19 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.

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

FIGURE 7.7 Estimates of Income Elasticity of Expenditure.

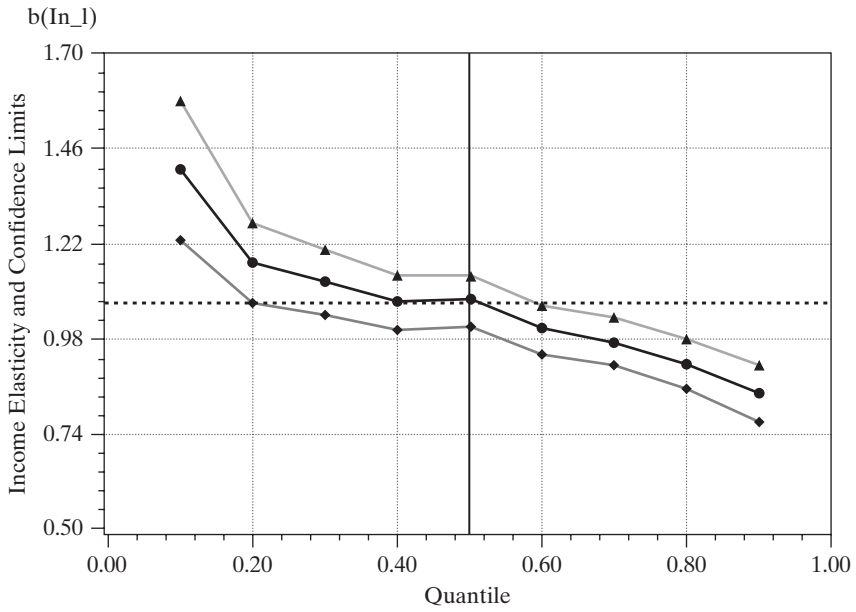
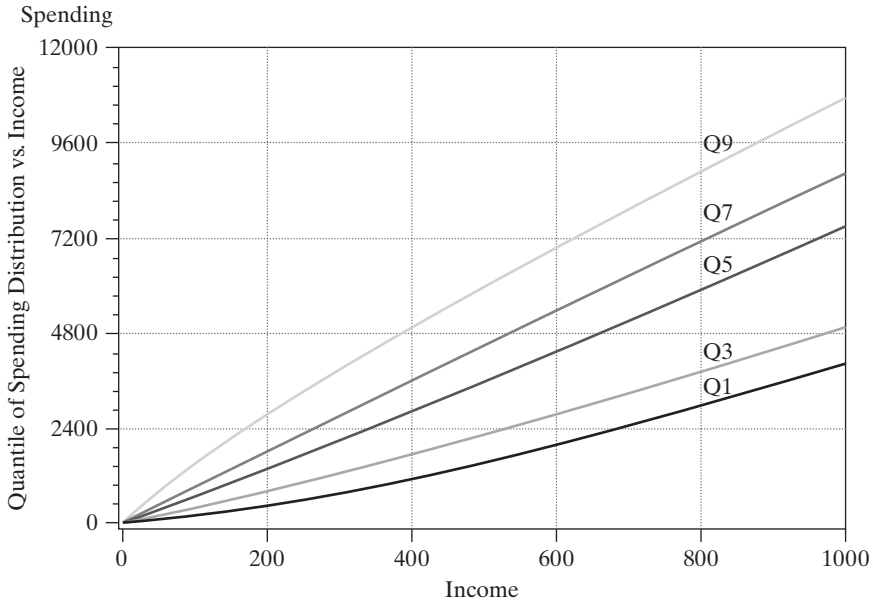


FIGURE 7.8 Quantile Regressions for Spending vs. Income.



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 a 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 is another approach.[24] 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}), $j = 1, \ldots, n/2$, in which $x_{j1} = x_{j2}$ within every pair. If so, then estimation of $\boldsymbol{\beta}$ could be based on the simple transformed model,

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

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 hand, a noisy estimate of $f(x_i)$ could be estimated with $y_i - \mathbf{z}_i'\hat{\boldsymbol{\beta}}_d$ (the estimate contains the estimation error as well as ε_i).[25]

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 $x_1 < x_2 < \cdots < x_n$. Suppose, as well, that this variable is well behaved in the sense that, as the sample size increases, this sorted data vector more completely 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.[26] 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 $M = 1, \ldots, 10$. 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

[24] Analyzed in detail by Yatchew (1998, 2000) and Härdle, Liang, and Gao (2000).

[25] See Estes and Honoré (1995) who suggest this approach (with simple differencing of the data).

[26] Yatchew (1997, 1998) goes more deeply into the underlying theory.

consistent, asymptotically normally distributed, and have asymptotic covariance matrix,[27]

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

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

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

Yatchew suggests that the partial residuals, $y_{d,i} - \mathbf{z}'_{d,i}\hat{\boldsymbol{\beta}}_d$, 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 7.13 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 10.3.2) 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 10.3.1 for description of the data.) There are 158 observations in the data set, but the last 35 are holding companies that 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 specified is a restricted version of the Christensen and Greene model,

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

where $c = \ln[C/(Q \times PF)]$, $k = \ln(PK/PF)$, $\mathrm{l} = \ln(PL/PF)$, 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.

$$\begin{array}{llcccccccl}
\text{(parametric)} & c & = & -7.32 & + & 0.069k & + & 0.241 - 0.569q + 0.057q^2/2 + \varepsilon, & & \\
 & & & (0.333) & & (0.065) & & (0.069)\ (0.042)\ (0.006) & & s = 0.13949 \\
\text{(partially linear)} & c_d & = & & & 0.108k_d & + & 0.163l_d + f(q) + v & & \\
 & & & & & (0.076) & & (0.081) & & s = 0.16529
\end{array}$$

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

[27] Yatchew (2000, p. 191) denotes this covariance matrix $E[\mathrm{Cov}[\mathbf{z}|\mathbf{x}]]$.

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

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

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.[28] 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

[28]See Cleveland (1979) and Schimek (2000).

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**,

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

where s is the sample standard deviation and IQR is the interquartile range (0.75 quantile minus 0.25 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**,

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

and the **Epanechnikov kernel**,

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

This produces the kernel weighted regression estimator,

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

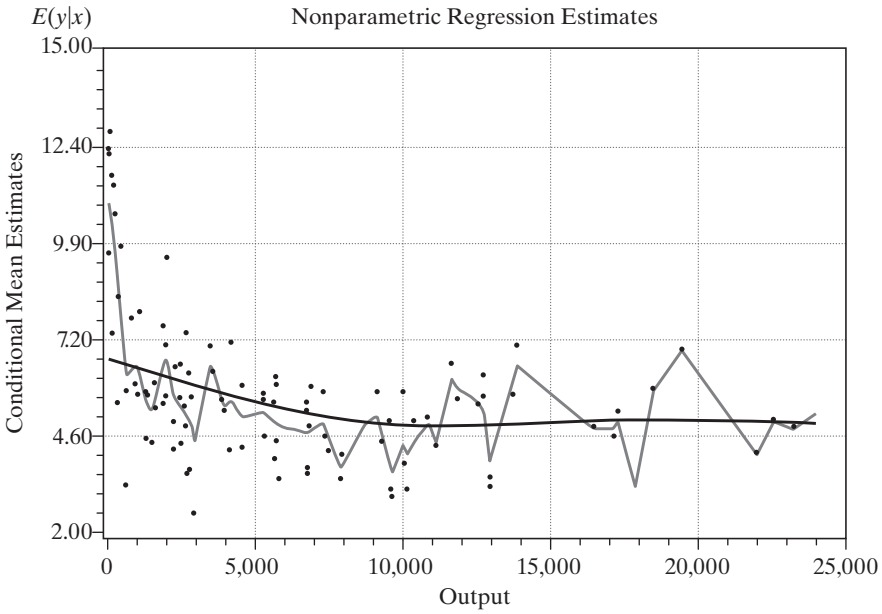
which has become a standard tool in nonparametric analysis.

Example 7.14 A Nonparametric Average Cost Function

In Example 7.13, we fit a partially linear regression for the relationship between average cost and output for electricity supply. Figure 7.9 shows 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 functions in Figure 7.9 use bandwidths of 2,000 and 100. Because 2,000 is a fairly large proportion of the range of variation of output, this function is quite smooth. The other function in Figure 7.9 uses a bandwidth of only 100. 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

FIGURE 7.9 Nonparametric Cost Functions.



of the data set. For the econometrician, however, there are a few drawbacks. There is no danger of misspecifying the conditional mean function; however, the great generality of the approach limits the ability to test one's specification or the underlying theory.[29] Most relationships are more complicated than a simple conditional mean of one variable. In Example 7.14, 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.[30] A promising approach is the partially linear model considered earlier. Henderson and Parmeter (2015) describe extensions of the kernel regression that accommodate multiple regression.

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).

[29]See, for example, Blundell, Browning, and Crawford's (2003) extensive study of British expenditure patterns.

[30]See Härdle (1990), Li and Racine (2007), and Henderson and Parmeter (2015).

Key Terms and Concepts

- Bandwidth
- Bootstrap
- Box–Cox transformation
- Conditional mean function
- Conditional median
- Delta method
- Epanechnikov kernel
- GMM estimator
- Identification condition
- Identification problem
- Indirect utility function
- Interaction term
- Iteration
- Jacobian
- Kernel density estimator
- Kernel functions
- Lagrange multiplier test
- Least absolute deviations (LAD)
- Linear regression model
- Linearized regression model
- Logistic kernel
- Median regression
- Nearest neighbor
- Neighborhood
- Nonlinear least squares
- Nonlinear regression model
- Nonparametric regression
- Orthogonality condition
- Overidentifying restrictions
- Partially linear model
- Pseudoregressors
- Quantile regression model
- Roy's identity
- Semiparametric
- Silverman's rule of thumb
- Smoothing function
- Starting values

Exercises

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

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

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

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

   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

1. 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 F7.2, estimate α, β_k, β_l, and λ by using the scanning method suggested in Example 7.5. (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:* $\partial \ln Y/\partial \ln K = K\partial \ln Y/\partial K$.)
2. For the model in Application 1, 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 λ is $\partial E[y|x]/\partial\lambda\,|\,\lambda = 0$. Result (7-35) will be useful.

3. 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 to 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/misra1c.shtml). The nonlinear regression model is

$$y_i = h(x, \boldsymbol{\beta}) + \varepsilon$$
$$= \beta_1\left(1 - \frac{1}{\sqrt{1 + 2\beta_2 x_i}}\right) + \varepsilon_i.$$

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:

	Estimate	*Estimated Standard Error*
β_1	6.3642725809E+02	4.6638326572E+00
β_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" to the solution. For this problem, the starting values provided are $\boldsymbol{\beta}^1 = (500, 0.0001)$ and

$\boldsymbol{\beta}^2 = (600, 0.0002)$. The exercise here is to reproduce the NIST results with your software. [For a detailed analysis of the NIST nonlinear least squares benchmarks with several well-known computer programs, see McCullough (1999).]

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

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

Example 6.19 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,

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

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 *NLOGIT*. The exercises are suggested as departure points for explorations using a few of the many estimation techniques listed in this chapter.

5. 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

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

6. To continue the analysis in Application 5, 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.