9 NONLINEAR REGRESSION MODELS

9.1 INTRODUCTION

Although the linear model is flexible enough to allow great variety in the shape of the regression, it still rules out many useful functional forms. In this chapter, we examine regression models that are intrinsically nonlinear in their parameters. This allows a much wider range of functional forms than the linear model can accommodate.[1]

9.2 NONLINEAR REGRESSION MODELS

The general form of the nonlinear regression model is

$$y_i = h(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i. \tag{9-1}$$

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

$$y = e^{\beta_0} x_1^{\beta_1} x_2^{\beta_2} e^{\varepsilon}$$

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

Example 9.1 CES Production Function

In Example 7.5, we examined a constant elasticity of substitution production function model:

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

No transformation renders this equation linear in the parameters. We did find, however, that a linear Taylor series approximation to this function around the point $\rho = 0$ produced an intrinsically linear equation that could be fit by least squares. Nonetheless, the true model is nonlinear in the sense that interests us in this chapter.

Example 9.2 Translog Demand System

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

$$-\ln V = \beta_0 + \sum_{k=1}^{K} \beta_k \ln(p_k/M) + \sum_{k=1}^{K}\sum_{l=1}^{K} \gamma_{kl} \ln(p_k/M)\ln(p_l/M)$$

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

where V is indirect utility, p_k is the price for the kth commodity and M is income. Roy's identity applied to this logarithmic function produces a budget share equation for the kth commodity that is of the form

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

where $\beta_M = \sum_k \beta_k$ and $\gamma_{Mj} = \sum_k \gamma_{kj}$. No transformation of the budget share equation produces a linear model. This is an intrinsically nonlinear regression model. (It is also one among a system of equations, an aspect we will ignore for the present.)

9.2.1 ASSUMPTIONS OF THE NONLINEAR REGRESSION MODEL

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

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

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

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

2. **Identifiability of the model parameters:** The parameter vector in the model is identified (estimable) if there is no nonzero parameter $\boldsymbol{\beta}^0 \neq \boldsymbol{\beta}$ such that $h(\mathbf{x}_i, \boldsymbol{\beta}^0) = h(\mathbf{x}_i, \boldsymbol{\beta})$ for all $\mathbf{x}_i$. In the linear model, this was the full rank assumption, but the simple absence of "multicollinearity" among the variables in $\mathbf{x}$ is not sufficient to produce this condition in the nonlinear regression model. Note that the model given in Example 9.2 is not identified. If the parameters in the model are all multiplied by the same nonzero constant, the same conditional mean function results. This condition persists even if all the variables in the model are linearly independent. The indeterminacy was removed in the study cited by imposing the **normalization** $\beta_M = 1$.

3. **Zero mean of the disturbance:** It follows from Assumption **1** that we may write

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

where $E[\varepsilon_i \mid h(\mathbf{x}_i, \boldsymbol{\beta})] = 0$. 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 return to this point below.

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

$$E\left[\varepsilon_i^2 \,\middle|\, h(\mathbf{x}_j, \boldsymbol{\beta}),\ j = 1, \ldots, n\right] = \sigma^2, \quad \text{a finite constant,} \qquad \textbf{(9-2)}$$

and nonautocorrelation

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

5. **Data generating process:** The data generating process for $\mathbf{x}_i$ is assumed to be a well behaved population such that first and second moments of the data can be assumed to converge to fixed, finite population counterparts. The crucial assumption is that the process generating $\mathbf{x}_i$ is strictly exogenous to that generating ε_i. The data on $\mathbf{x}_i$ are assumed to be "well behaved."
6. **Underlying probability model:** There is a well defined probability distribution generating 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 0 and variance σ^2 conditioned on $h(\mathbf{x}_i, \boldsymbol{\beta})$. Thus, at this point, our statement of the model is **semiparametric.** (See Section 16.3.) 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. In Chapter 17, we will fully parameterize the model by assuming that the disturbances are normally distributed. This will allow us to be more specific about certain test statistics and, in addition, allow some generalizations of the regression model. The assumption is not necessary here.

9.2.2 THE ORTHOGONALITY CONDITION AND THE SUM OF SQUARES

Assumptions 1 and 3 imply that $E[\varepsilon_i \mid h(\mathbf{x}_i, \boldsymbol{\beta})] = 0$. In the linear model, it follows, *because of the linearity of the conditional mean,* that ε_i and $\mathbf{x}_i$, itself, 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 below 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, $E[\boldsymbol{\varepsilon} \mid \mathbf{X}] = \mathbf{0}$ 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 perhaps a treatment analogous to that for the linear model would apply. But the convergence results needed to obtain the asymptotic properties of the estimator still have to be strengthened. The dynamic nonlinear regression model is beyond the reach of our treatment here. Strict independence of 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).

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

$$\ln f(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) = -(1/2)\left[\ln 2\pi + \ln \sigma^2 + \varepsilon_i^2/\sigma^2\right]. \tag{9-3}$$

For this special case, we have from item D.2 in Theorem 17.2 (on maximum likelihood estimation), that the derivatives of the log density with respect to the parameters have mean zero. That is,

$$E\left[\frac{\partial \ln f(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta}}\right] = E\left[\frac{1}{\sigma^2}\left(\frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)\varepsilon_i\right] = \mathbf{0}, \tag{9-4}$$

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

In the context of the linear model, the **orthogonality condition** $E[\mathbf{x}_i\varepsilon_i] = 0$ produces least squares as a **GMM estimator** for the model. (See Chapter 18.) 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 (9-4) produces a moment condition which will define the nonlinear least squares estimator as a GMM estimator.

Example 9.3 First-Order Conditions for a Nonlinear Model

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

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

by nonlinear least squares [see (9-10)] are

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

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

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

These equations do not have an explicit solution.

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

DEFINITION 9.1 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.

9.2.3 THE LINEARIZED REGRESSION

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

$$h(\mathbf{x}, \boldsymbol{\beta}) \approx h(\mathbf{x}, \boldsymbol{\beta}^0) + \sum_{k=1}^{K} \frac{\partial h(\mathbf{x}, \boldsymbol{\beta}^0)}{\partial \beta_k^0}\left(\beta_k - \beta_k^0\right). \tag{9-5}$$

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

$$h(\mathbf{x}, \boldsymbol{\beta}) \approx \left[h(\mathbf{x}, \boldsymbol{\beta}^0) - \sum_{k=1}^{K} \beta_k^0 \left(\frac{\partial h(\mathbf{x}, \boldsymbol{\beta}^0)}{\partial \beta_k^0}\right)\right] + \sum_{k=1}^{K} \beta_k \left(\frac{\partial h(\mathbf{x}, \boldsymbol{\beta}^0)}{\partial \beta_k^0}\right). \tag{9-6}$$

Let x_k^0 equal the kth partial derivative,[2] $\partial h(\mathbf{x}, \boldsymbol{\beta}^0)/\partial \beta_k^0$. 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

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

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:

$$y^0 = y - h^0 + \mathbf{x}^{0\prime}\boldsymbol{\beta}^0 = \mathbf{x}^{0\prime}\boldsymbol{\beta} + \varepsilon^0. \tag{9-7}$$

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 (9-6). That is,

$$\varepsilon^0 = \varepsilon + \left[h(\mathbf{x}, \boldsymbol{\beta}) - \left\{h^0 - \sum_{k=1}^{K} x_k^0 \beta_k^0 + \sum_{k=1}^{K} x_k^0 \beta_k\right\}\right]. \tag{9-8}$$

Since all the errors are accounted for, (9-7) 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 (9-7) by linear least squares. (Whether this estimator is consistent or not remains to be seen.)

Example 9.4 Linearized Regression

For the model in Example 9.3, 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 $\boldsymbol{\beta}^0$,

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

could be regressed on the three variables previously defined to estimate β_1, β_2, and β_3.

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

9.2.4 LARGE SAMPLE PROPERTIES OF THE NONLINEAR LEAST SQUARES ESTIMATOR

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

It is necessary to make some assumptions about the regressors. The precise requirements are discussed in some detail in Judge et al. (1985), Amemiya (1985), and Davidson and MacKinnon (1993). In the linear regression model, to obtain our asymptotic results, we assume that the sample moment matrix $(1/n)\mathbf{X}'\mathbf{X}$ converges to a positive definite matrix $\mathbf{Q}$. By analogy, we impose the same condition on the derivatives of the regression function, which are called the **pseudoregressors** in the linearized model *when they are computed at the true parameter values*. Therefore, for the nonlinear regression model, the analog to (5-1) is

$$\text{plim}\,\frac{1}{n}\mathbf{X}^{0\prime}\mathbf{X}^0 = \text{plim}\,\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial h(\mathbf{x}_i,\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}_0}\right)\left(\frac{\partial h(\mathbf{x}_i,\boldsymbol{\beta}_0)}{\partial\boldsymbol{\beta}_0'}\right) = \mathbf{Q}^0, \qquad \textbf{(9-9)}$$

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

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

This is the orthogonality condition noted earlier in (5-4). In particular, note that orthogonality of the disturbances and the data is not the same condition. Finally, asymptotic normality can be established under general conditions if

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

With these in hand, the asymptotic properties of the nonlinear least squares estimator have been derived. They are, in fact, essentially those we have already seen for the 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. [Amemiya (1985).]

The nonlinear least squares criterion function is

$$S(\mathbf{b}) = \frac{1}{2}\sum_{i=1}^{n}[y_i - h(\mathbf{x}_i,\mathbf{b})]^2 = \frac{1}{2}\sum_{i=1}^{n}e_i^2, \qquad \textbf{(9-10)}$$

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

$$\mathbf{g}(\mathbf{b}) = -\sum_{i=1}^{n}[y_i - h(\mathbf{x}_i, \mathbf{b})]\frac{\partial h(\mathbf{x}_i, \mathbf{b})}{\partial \mathbf{b}} = \mathbf{0}. \tag{9-11}$$

In the linear model of Chapter 2, this produces a set of linear equations, the normal equations (3-4). But in this more general case, (9-11) is a set of nonlinear equations that do not have an explicit solution. Note that σ^2 is not relevant to the solution [nor was it in (3-4)]. 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 9.1 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. *$q(\boldsymbol{\beta}^0)$ has a unique minimum at the true parameter vector, $\boldsymbol{\beta}$,*

then, the nonlinear least squares estimator defined by (9-10) and (9-11) 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. In order 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. An argument to this effect appears in Mittelhammer et al. (2000, p. 190–191).

THEOREM 9.2 Asymptotic Normality of the Nonlinear Least Squares Estimator

If the pseudoregressors defined in (9-3) 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 estimate of the asymptotic covariance matrix is

$$\text{Est.Asy. Var}[\mathbf{b}] = \hat{\sigma}^2(\mathbf{X}^{0\prime}\mathbf{X}^0)^{-1}. \qquad \textbf{(9-12)}$$

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.

The requirement that the matrix in (9-9) 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 9.5 gives an application.

9.2.5 COMPUTING THE NONLINEAR LEAST SQUARES ESTIMATOR

Minimizing the sum of squares is a standard problem in nonlinear optimization that can be solved by a number of methods. (See Section E.6.) The method of Gauss–Newton is often used. In the linearized regression model, if a value of $\boldsymbol{\beta}^0$ is available, then the linear regression model shown in (9-7) 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

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

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.6.5 is $\delta = \mathbf{e}^0\mathbf{X}^0(\mathbf{X}^{0\prime}\mathbf{X}^0)^{-1}\mathbf{X}^{0\prime}\mathbf{e}^0$. We note, finally, that iteration of the linearized regression, although a very effective algorithm for many problems, does not always work. As does Newton's method, this algorithm sometimes "jumps off" to a wildly errant second iterate, after which it may be impossible to compute the residuals for the next iteration. The choice of starting values for the iterations can be crucial. There is art as well as science in the computation of nonlinear least squares estimates. [See McCullough and Vinod (1999).] In the absence of information about starting values, a workable strategy is to try the Gauss–Newton iteration first. If it 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.

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

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}[y_i - h(\mathbf{x}_i, \mathbf{b})]^2. \tag{9-13}$$

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 (1993) argue that on average, (9-13) 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 which is the case for the program they are using. With this in hand, the estimator of the asymptotic covariance matrix for the nonlinear least squares estimator is given in (9-12).

Once the nonlinear least squares estimates are in hand, inference and hypothesis tests can proceed in the same fashion as prescribed in Chapter 7. A minor problem can arise in evaluating the fit of the regression in that the familiar measure,

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{9-14}$$

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

9.3 APPLICATIONS

We will examine two applications. The first is a nonlinear extension of the consumption function examined in Example 2.1. The Box–Cox transformation presented in Section 9.3.2 is a device used to search for functional form in regression.

9.3.1 A Nonlinear Consumption Function

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

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

in which γ equals 1. With this restriction, the model is linear. If γ is free to vary, however, then this version becomes a nonlinear regression. The linearized model is

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

The nonlinear least squares procedure reduces to iterated regression of

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

Quarterly data on consumption, real disposable income, and several other variables for 1950 to 2000 are listed in Appendix Table F5.1. We will use these to fit the nonlinear consumption function. This turns out to be a particularly straightforward estimation problem. **Iterations** are begun at the linear least squares estimates for α and β and 1 for γ. As shown below, the solution is reached in 8 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 noted above.)

Begin NLSQ iterations. Linearized regression.

Iteration = 1; Sum of squares = 1536321.88; Gradient = 996103.930
Iteration = 2; Sum of squares = $.1847 \times 10^{12}$; Gradient = $.1847 \times 10^{12}$
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 = .752189847
Iteration = 6; Sum of squares = 504403.216; Gradient = .526642396E-04
Iteration = 7; Sum of squares = 504403.216; Gradient = .511324981E-07
Iteration = 8; Sum of squares = 504403.216; Gradient = .606793426E-10

The linear and nonlinear least squares regression results are shown in Table 9.1.

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

TABLE 9.1 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	.01091
γ	1.0000	—	1.24483	.01205
e′e	1,536,321.881		504,403.1725	
σ	87.20983		50.0946	
R^2	.996448		.998834	
Var[b]	—		0.000119037	
Var[c]	—		0.00014532	
Cov[b, c]	—		−0.000131491	

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 procedures outlined earlier are used at the last iteration 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.) The estimates for the linear model are shown in Table 9.1 as well. Eight iterations are required for convergence. The value of δ is shown at the right. 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.

For hypothesis testing and confidence intervals, the usual 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 percent significance level, and we thus reject the linear model in favor of the nonlinear regression. We are also interested in the marginal propensity to consume. In this expanded model, $H_0: \gamma = 1$ is a test that the marginal propensity to consume is constant, not that it is 1. (That would be a joint test of both $\gamma = 1$ and $\beta = 1$.) In this model, the marginal propensity to consume is

$$\text{MPC} = \frac{dc}{dY} = \beta\gamma Y^{\gamma-1},$$

which varies with Y. To test the hypothesis that this value is 1, we require a particular value of Y. Since it is the most recent value, we choose $\text{DPI}_{2000.4} = 6634.9$. At this value, the MPC is estimated as 1.08264. We estimate its standard error using the delta method, with the square root of

$$[\partial \text{MPC}/\partial b \quad \partial \text{MPC}/\partial c]\begin{bmatrix} \text{Var}[b] & \text{Cov}[b,c] \\ \text{Cov}[b,c] & \text{Var}[c] \end{bmatrix}\begin{bmatrix} \partial \text{MPC}/\partial b \\ \partial \text{MPC}/\partial c \end{bmatrix}$$

$$= [cY^{c-1} \quad bY^{c-1}(1 + c\ln Y)]\begin{bmatrix} 0.00011904 & -0.000131491 \\ -0.000131491 & 0.00014532 \end{bmatrix}\begin{bmatrix} cY^{c-1} \\ bY^{c-1}(1 + c\ln Y) \end{bmatrix}$$

$$= 0.00007469,$$

which gives a standard error of 0.0086425. For testing the hypothesis that the MPC is equal to 1.0 in 2000.4, we would refer

$$z = \frac{1.08264 - 1}{0.0086425} = -9.562$$

to a standard normal table. This difference is certainly statistically significant, so we would reject the hypothesis.

Example 9.5 Multicollinearity in Nonlinear Regression

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 0.999312 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 $\mathbf{D}^{-1}\mathbf{X}^{0\prime}\mathbf{X}^0\mathbf{D}^{-1}$ 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 value for a problematic data set. By the standards discussed in Section 4.9.1, the collinearity problem in this "data set" is severe.

9.3.2 THE BOX–COX TRANSFORMATION

The Box–Cox transformation is a device for generalizing the linear model. The transformation is[3]

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

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

$$y = \alpha + \sum_{k=1}^{K} \beta_k x_k^{(\lambda)} + \varepsilon \qquad \textbf{(9-15)}$$

is a linear regression that can be estimated by least squares.[4] 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.[5] At the same time, it is also possible to transform y, say, by

[3] Box and Cox (1964). To be defined for all values of λ, x must be strictly positive. See also Zarembka (1974).

[4] In most applications, some of the regressors—for example, dummy variable—will not be transformed. For such a variable, say ν_k, $\nu_k^{(\lambda)} = \nu_k$, and the relevant derivatives in (9-16) will be zero.

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

$y^{(\theta)}$. Transformation of the dependent variable, however, amounts to a specification of the whole model, not just the functional form. We will examine this case more closely in Section 17.6.2.

Example 9.6 Flexible Cost Function

Caves, Christensen, and Trethaway (1980) analyzed the costs of production for railroads providing freight and passenger service. Continuing a long line of literature on the costs of production in regulated industries, a translog cost function (see Section 14.3.2) would be a natural choice for modeling this multiple-output technology. Several of the firms in the study, however, produced no passenger service, which would preclude the use of the translog model. (This model would require the log of zero.) An alternative is the Box–Cox transformation, which is computable for zero output levels. A constraint must still be placed on λ in their model, as $0^{(\lambda)}$ is defined only if λ is strictly positive. A positive value of λ is not assured. A question does arise in this context (and other similar ones) as to whether zero outputs should be treated the same as nonzero outputs or whether an output of zero represents a discrete corporate decision distinct from other variations in the output levels. In addition, as can be seen in (9-16), this solution is only partial. The zero values of the regressors preclude computation of appropriate standard errors.

If λ in (9-15) is taken to be an unknown parameter, then the regression becomes nonlinear in the parameters. Although no transformation will reduce it to linearity, 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. When λ equals zero, the transformation is, by L'Hôpital's rule,

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

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 (and, with normality of the disturbance, maximum likelihood) estimates of the parameters.

After determining the optimal value of λ, it is sometimes treated as if it were a *known* value in the least squares results. But $\hat{\lambda}$ is an estimate of an unknown parameter. It is not hard to show that the least squares standard errors will always underestimate the correct asymptotic standard errors.[6] To get the appropriate values, we need the derivatives of the right-hand side of (9-15) with respect to α, $\boldsymbol{\beta}$, and λ. In the notation of Section 9.2.3, these are

$$\begin{aligned}\frac{\partial h(.)}{\partial \alpha} &= 1,\\ \frac{\partial h(.)}{\partial \beta_k} &= x_k^{(\lambda)}, \qquad\qquad (9\text{-}16)\\ \frac{\partial h(.)}{\partial \lambda} &= \sum_{k=1}^{K}\beta_k\frac{\partial x_k^{(\lambda)}}{\partial \lambda} = \sum_{k=1}^{K}\beta_k\left[\frac{1}{\lambda}\left(x_k^\lambda \ln x_k - x_k^{(\lambda)}\right)\right].\end{aligned}$$

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

We can now use (9-12) and (9-13) 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. This was the point noted at the end of Example 9.6.

It is important to remember that the coefficients in a nonlinear model are not equal to the slopes (i.e., here the demand elasticities) with respect to the variables. For the Box–Cox model,[7]

$$\begin{aligned}\ln Y &= \alpha + \beta\left[\frac{X^\lambda - 1}{\lambda}\right] + \varepsilon\\ \frac{dE[\ln Y\,|\,X]}{d\ln X} &= \beta X^\lambda = \eta.\end{aligned}\tag{9-17}$$

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

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

9.4 HYPOTHESIS TESTING AND PARAMETRIC RESTRICTIONS

In most cases, the sorts of hypotheses one would test in this context will involve fairly simple linear restrictions. The tests can be carried out using the usual formulas discussed in Chapter 7 and the asymptotic covariance matrix presented earlier. For more involved hypotheses and for nonlinear restrictions, the procedures are a bit less clear-cut. Three principal testing procedures were discussed in Section 6.4 and Appendix C: the Wald, likelihood ratio, and Lagrange multiplier tests. For the linear model, all three statistics are transformations of the standard F statistic (see Section 17.6.1), so the tests are essentially identical. In the nonlinear case, they are equivalent only asymptotically. We will work through the Wald and Lagrange multiplier tests for the general case and then apply them to the example of the previous section. Since we have not assumed normality of the disturbances (yet), we will postpone treatment of the likelihood ratio statistic until we revisit this model in Chapter 17.

9.4.1 SIGNIFICANCE TESTS FOR RESTRICTIONS: *F* AND WALD STATISTICS

The hypothesis to be tested is

$$H_0\colon \mathbf{r}(\boldsymbol{\beta}) = \mathbf{q}.\tag{9-18}$$

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

[7] We have used the result $d\ln Y/d\ln X = X d\ln Y/dX$.

on the parameters. If there is more than one restriction, then they must be functionally independent. These two conditions imply that the $J \times K$ matrix

$$\mathbf{R}(\boldsymbol{\beta}) = \frac{\partial \mathbf{r}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} \tag{9-19}$$

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.)

Let **b** be the unrestricted, nonlinear least squares estimator, and let $\mathbf{b}_*$ be the estimator obtained when the constraints of the hypothesis are imposed.[8] Which test statistic one uses depends on how difficult the computations are. Unlike the linear model, the various testing procedures vary in complexity. For instance, in our example, the Lagrange multiplier is by far the simplest to compute. Of the four methods we will consider, only this test does not require us to compute a nonlinear regression.

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

$$F[J, n-K] = \frac{[S(\mathbf{b}_*) - S(\mathbf{b})]/J}{S(\mathbf{b})/(n-K)}. \tag{9-20}$$

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

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

$$\begin{aligned} W &= [\mathbf{r}(\mathbf{b}) - \mathbf{q}]'\{\text{Est.Asy. Var}[\mathbf{r}(\mathbf{b}) - \mathbf{q}]\}^{-1}[\mathbf{r}(\mathbf{b}) - \mathbf{q}] \\ &= [\mathbf{r}(\mathbf{b}) - \mathbf{q}]'\{\mathbf{R}(\mathbf{b})\hat{\mathbf{V}}\mathbf{R}'(\mathbf{b})\}^{-1}[\mathbf{r}(\mathbf{b}) - \mathbf{q}], \end{aligned} \tag{9-21}$$

where

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

and $\mathbf{R}(\mathbf{b})$ is evaluated at **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

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

Wald statistic is not invariant to how the hypothesis is framed. In cases in which there are more than one equivalent ways to specify $\mathbf{r}(\boldsymbol{\beta}) = \mathbf{q}$, W can give different answers depending on which is chosen.

9.4.2 TESTS BASED ON THE LM STATISTIC

The **Lagrange multiplier test** is based on the decrease in the sum of squared residuals that would result if the restrictions in the restricted model were released. The formalities of the test are given in Sections 17.5.3 and 17.6.1. For the nonlinear regression model, the test has a particularly appealing form.[9] 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 (9-6). Let $\mathbf{X}^0_*$ be this matrix *computed at the restricted estimates.* Then the Lagrange multiplier statistic for the nonlinear regression model is

$$\text{LM} = \frac{\mathbf{e}_*'\mathbf{X}^0_*[\mathbf{X}^{0\prime}_*\mathbf{X}^0_*]^{-1}\mathbf{X}^{0\prime}_*\mathbf{e}_*}{\mathbf{e}_*'\mathbf{e}_*/n}. \tag{9-22}$$

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.

Example 9.7 Hypotheses Tests in a Nonlinear Regression Model

We test the hypothesis $H_0: \gamma = 1$ in the consumption function of Section 9.3.1.

- **F statistic.** The F statistic is

$$F[1{,}204-3] = \frac{(1{,}536{,}321.881 - 504{,}403.57)/1}{504{,}403.57/(204-3)} = 411.29.$$

  The critical value from the tables is 4.18, so the hypothesis is rejected.
- **Wald statistic.** For our example, the Wald statistic is based on the distance of $\hat{\gamma}$ from 1 and is simply the square of the asymptotic t ratio we computed at the end of the example:

$$W = \frac{(1.244827-1)^2}{0.01205^2} = 412.805.$$

  The critical value from the chi-squared table is 3.84.
- **Lagrange multiplier.** For our example, the elements in $\mathbf{x}_i^*$ are

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

  To compute this at the restricted estimates, we use the ordinary least squares estimates for α and β and 1 for γ so that

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

[9] This test is derived in Judge et al. (1985). A lengthy discussion appears in Mittelhammer et al. (2000).

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

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

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

9.4.3 A SPECIFICATION TEST FOR NONLINEAR REGRESSIONS: THE P_E TEST

MacKinnon, White, and Davidson (1983) have extended the J test discussed in Section 8.3.3 to nonlinear regressions. One result of this analysis is a simple test for linearity versus loglinearity.

The specific hypothesis to be tested is

$$H_0 : y = h^0(\mathbf{x}, \boldsymbol{\beta}) + \varepsilon_0$$

versus

$$H_1 : g(y) = h^1(\mathbf{z}, \boldsymbol{\gamma}) + \varepsilon_1,$$

where $\mathbf{x}$ and $\mathbf{z}$ are regressor vectors and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the parameters. As the authors note, using y instead of, say, $j(y)$ in the first function is nothing more than an implicit definition of the units of measurement of the dependent variable.

An intermediate case is useful. If we assume that $g(y)$ is equal to y but we allow $h^0(.)$ and $h^1(.)$ to be nonlinear, then the necessary modification of the J test is straightforward, albeit perhaps a bit more difficult to carry out. For this case, we form the compound model

$$\begin{aligned} y &= (1-\alpha)h^0(\mathbf{x}, \boldsymbol{\beta}) + \alpha h^1(\mathbf{z}, \boldsymbol{\gamma}) + \varepsilon \\ &= h^0(\mathbf{x}, \boldsymbol{\beta}) + \alpha[h^1(\mathbf{z}, \boldsymbol{\gamma}) - h^0(\mathbf{x}, \boldsymbol{\beta})] + \varepsilon. \end{aligned} \qquad \textbf{(9-23)}$$

Presumably, both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ could be estimated in isolation by nonlinear least squares. Suppose that a nonlinear least squares estimate of $\boldsymbol{\gamma}$ has been obtained. One approach is to insert this estimate in (9-23) and then estimate $\boldsymbol{\beta}$ and α by nonlinear least squares. The J test amounts to testing the hypothesis that α equals zero. Of course, the model is symmetric in $h^0(.)$ and $h^1(.)$, so their roles could be reversed. The same conclusions drawn earlier would apply here.

Davidson and MacKinnon (1981) propose what may be a simpler alternative. Given an estimate of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}$, approximate the first $h^0(\mathbf{x}, \boldsymbol{\beta})$ in (9-23) with a linear Taylor series at this point. The result is

$$h^0(\mathbf{x}, \boldsymbol{\beta}) \approx h^0(\mathbf{x}, \hat{\boldsymbol{\beta}}) + \left[\frac{\partial h^0(.)}{\partial \hat{\boldsymbol{\beta}}'}\right](\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) = \hat{h}^0 + \hat{\mathbf{H}}^0\boldsymbol{\beta} - \hat{\mathbf{H}}^0\hat{\boldsymbol{\beta}}. \qquad \textbf{(9-24)}$$

(Note $\mathbf{H}^0$ is a row vector of derivatives.) Using this device, they replace (9-23) with

$$y - \hat{h}^0 + \hat{\mathbf{H}}^0\hat{\boldsymbol{\beta}} = \hat{\mathbf{H}}^0\boldsymbol{\beta} + \alpha[h^1(\mathbf{z}, \hat{\boldsymbol{\gamma}}) - h^0(\mathbf{x}, \hat{\boldsymbol{\beta}})] + \boldsymbol{\varepsilon},$$

in which $\boldsymbol{\beta}$ and α can be estimated by linear least squares. As before, the J test amounts to testing the significance of $\hat{\alpha}$. If it is found that $\hat{\alpha}$ is significantly different from zero, then H_0 is rejected. For the authors' asymptotic results to hold, any initial consistent estimator of $\boldsymbol{\beta}$ will suffice for $\hat{\boldsymbol{\beta}}$; the nonlinear least squares estimator that they suggest seems a natural choice.[10]

Now we can generalize the test to allow a nonlinear function, $g(y)$, in H_1. Davidson and MacKinnon require $g(y)$ to be monotonic, continuous, and continuously differentiable and not to introduce any new parameters. (This requirement excludes the Box–Cox model, which is considered in Section 9.3.2.) The compound model that forms the basis of the test is

$$(1-\alpha)[y - h^0(\mathbf{x}, \boldsymbol{\beta})] + \alpha[g(y) - h^1(\mathbf{z}, \boldsymbol{\gamma})] = \varepsilon. \qquad \textbf{(9-25)}$$

Again, there are two approaches. As before, if $\hat{\boldsymbol{\gamma}}$ is an estimate of $\boldsymbol{\gamma}$, then $\boldsymbol{\beta}$ and α can be estimated by maximum likelihood conditional on this estimate.[11] This method promises to be extremely messy, and an alternative is proposed. Rewrite (9-25) as

$$y - h^0(\mathbf{x}, \boldsymbol{\beta}) = \alpha[h^1(\mathbf{z}, \boldsymbol{\gamma}) - g(y)] + \alpha[y - h^0(\mathbf{x}, \boldsymbol{\beta})] + \varepsilon.$$

Now use the same linear Taylor series expansion for $h^0(\mathbf{x}, \boldsymbol{\beta})$ on the left-hand side and replace both y and $h^0(\mathbf{x}, \boldsymbol{\beta})$ with $\hat{h}^0$ on the right. The resulting model is

$$y - \hat{h}^0 + \hat{\mathbf{H}}^0\boldsymbol{\beta} = \hat{\mathbf{H}}^0\boldsymbol{\beta} + \alpha[\hat{h}^1 - g(\hat{h}^0)] + e. \qquad \textbf{(9-26)}$$

As before, with an initial estimate of $\boldsymbol{\beta}$, this model can be estimated by least squares.

This modified form of the J test is labeled the P_E *test.* As the authors discuss, it is probably not as powerful as any of the Wald or Lagrange multiplier tests that we have considered. In their experience, however, it has sufficient power for applied research and is clearly simple to carry out.

The P_E test can be used to test a linear specification against a loglinear model. For this test, both $h^0(.)$ and $h^1(.)$ are linear, whereas $g(y) = \ln y$. Let the two competing models be denoted

$$H_0 : y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$$

and

$$H_1 : \ln y = \ln(\mathbf{x})'\boldsymbol{\gamma} + \varepsilon.$$

[We stretch the usual notational conventions by using $\ln(\mathbf{x})$ for $(\ln x_1, \ldots, \ln x_k)$.] Now let **b** and **c** be the two linear least squares estimates of the parameter vectors. The P_E test for H_1 as an alternative to H_0 is carried out by testing the significance of the coefficient $\hat{\alpha}$ in the model

$$y = \mathbf{x}'\boldsymbol{\beta} + \alpha[\widehat{\ln y} - \ln(\mathbf{x}'\mathbf{b})] + \phi. \qquad \textbf{(9-27)}$$

The second term is the difference between predictions of $\ln y$ obtained directly from the loglinear model and obtained as the log of the prediction from the linear model. We can also reverse the roles of the two formulas and test H_0 as the alternative. The

[10] This procedure assumes that H_0 is correct, of course.

[11] Least squares will be inappropriate because of the transformation of y, which will translate to a Jacobian term in the log-likelihood. See the later discussion of the Box–Cox model.

TABLE 9.2 Estimated Money Demand Equations

	a	b_r	c_Y	R^2	s
Linear	−228.714 (13.891)	−23.849 (2.044)	0.1770 (0.00278)	0.95548	76.277
P_E test for the linear model, $\hat{\alpha} = -121.496$ (46.353), $t = -2.621$					
Loglinear	−8.9473 (0.2181)	−0.2590 (0.0236)	1.8205 (0.0289)	0.96647	0.14825
P_E test for the loglinear model, $\hat{\alpha} = -0.0003786$ (0.0001969), $t = 1.925$					

compound regression is

$$\ln y = \ln(\mathbf{x})'\boldsymbol{\gamma} + \alpha\left(\hat{y} - e^{\ln(\mathbf{x})'\mathbf{c}}\right) + \varepsilon. \tag{9-28}$$

The test of linearity vs. loglinearity has been the subject of a number of studies. Godfrey and Wickens (1982) discuss several approaches.

Example 9.8 Money Demand

A large number of studies have estimated money demand equations, some linear and some log-linear.[12] Quarterly data from 1950 to 2000 for estimation of a money demand equation are given in Appendix Table F5.1. The interest rate is the quarterly average of the monthly average 90 day T-bill rate. The money stock is M1. Real GDP is seasonally adjusted and stated in 1996 constant dollars. Results of the P_E test of the linear versus the loglinear model are shown in Table 9.2.

Regressions of M1 on a constant, r and Y, and ln M1 on a constant, ln r and ln Y, produce the results given in Table 9.2 (standard errors are given in parentheses). Both models appear to fit quite well,[13] and the pattern of significance of the coefficients is the same in both equations. After computing fitted values from the two equations, the estimates of α from the two models are as shown in Table 9.2. Referring these to a standard normal table, we reject the linear model in favor of the loglinear model.

9.5 ALTERNATIVE ESTIMATORS FOR NONLINEAR REGRESSION MODELS

Section 9.2 discusses the "standard" case in which the only complication to the classical regression model of Chapter 2 is that the conditional mean function in $y_i = h(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i$ is a nonlinear function of $\boldsymbol{\beta}$. This fact mandates an alternative estimator, nonlinear least squares, and some new interpretation of the "regressors" in the model. In this section, we will consider two extensions of these results. First, as in the linear case, there can be situations in which the assumption that Cov$[\mathbf{x}_i, \varepsilon_i] = \mathbf{0}$ is not reasonable. These situations will, as before, require an instrumental variables treatment, which we consider in Section 9.5.1. Second, there will be models in which it is convenient to estimate the parameters in two steps, estimating one subset at the first step and then using these estimates in a second step at which the remaining parameters are estimated.

[12] A comprehensive survey appears in Goldfeld (1973).

[13] The interest elasticity is in line with the received results. The income elasticity is quite a bit larger.

We will have to modify our asymptotic results somewhat to accommodate this estimation strategy. The two-step estimator is discussed in Section 9.5.2.

9.5.1 NONLINEAR INSTRUMENTAL VARIABLES ESTIMATION

In Section 5.4, we extended the linear regression model to allow for the possibility that the regressors might be correlated with the disturbances. The same problem can arise in nonlinear models. The consumption function estimated in Section 9.3.1 is almost surely a case in point, and we reestimated it using the instrumental variables technique for linear models in Example 5.3. In this section, we will extend the method of instrumental variables to nonlinear regression models.

In the nonlinear model,

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

the covariates $\mathbf{x}_i$ may be correlated with the disturbances. We would expect this effect to be transmitted to the pseudoregressors, $\mathbf{x}_i^0 = \partial h(\mathbf{x}_i, \boldsymbol{\beta})/\partial \boldsymbol{\beta}$. If so, then the results that we derived for the linearized regression would no longer hold. Suppose that there is a set of variables $[\mathbf{z}_1, \ldots, \mathbf{z}_L]$ such that

$$\operatorname{plim}(1/n)\mathbf{Z}'\boldsymbol{\varepsilon} = \mathbf{0} \tag{9-29}$$

and

$$\operatorname{plim}(1/n)\mathbf{Z}'\mathbf{X}^0 = \mathbf{Q}^0_{\mathbf{zx}} \neq \mathbf{0},$$

where $\mathbf{X}^0$ is the matrix of pseudoregressors in the linearized regression, evaluated at the true parameter values. If the analysis that we did for the linear model in Section 5.4 can be applied to this set of variables, then we will be able to construct a consistent estimator for $\boldsymbol{\beta}$ using the instrumental variables. As a first step, we will attempt to replicate the approach that we used for the linear model. The linearized regression model is given in (9-7),

$$\mathbf{y} = \mathbf{h}(\mathbf{X}, \boldsymbol{\beta}) + \boldsymbol{\varepsilon} \approx \mathbf{h}^0 + \mathbf{X}^0(\boldsymbol{\beta} - \boldsymbol{\beta}^0) + \boldsymbol{\varepsilon}$$

or

$$\mathbf{y}^0 \approx \mathbf{X}^0\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\mathbf{y}^0 = \mathbf{y} - \mathbf{h}^0 + \mathbf{X}^0\boldsymbol{\beta}^0.$$

For the moment, we neglect the approximation error in linearizing the model. In (9-29), we have assumed that

$$\operatorname{plim}(1/n)\mathbf{Z}'\mathbf{y}^0 = \operatorname{plim}\,(1/n)\mathbf{Z}'\mathbf{X}^0\boldsymbol{\beta}. \tag{9-30}$$

Suppose, as we did before, that there are the same number of instrumental variables as there are parameters, that is, columns in $\mathbf{X}^0$. (Note: This number need not be the number of variables. See our preceding example.) Then the "estimator" used before is suggested:

$$\mathbf{b}_{\mathrm{IV}} = (\mathbf{Z}'\mathbf{X}^0)^{-1}\mathbf{Z}'\mathbf{y}^0. \tag{9-31}$$

The logic is sound, but there is a problem with this estimator. The unknown parameter vector $\boldsymbol{\beta}$ appears on both sides of (9-30). We might consider the approach we used for our first solution to the nonlinear regression model. That is, with some initial estimator in hand, iterate back and forth between the instrumental variables regression and recomputing the pseudoregressors until the process converges to the fixed point that we seek. Once again, the logic is sound, and in principle, this method does produce the estimator we seek.

If we add to our preceding assumptions

$$\frac{1}{\sqrt{n}}\mathbf{Z}'\boldsymbol{\varepsilon} \xrightarrow{d} N[\mathbf{0}, \sigma^2\mathbf{Q}_{\mathbf{zz}}],$$

then we will be able to use the same form of the asymptotic distribution for this estimator that we did for the linear case. Before doing so, we must fill in some gaps in the preceding. First, despite its intuitive appeal, the suggested procedure for finding the estimator is very unlikely to be a good algorithm for locating the estimates. Second, we do not wish to limit ourselves to the case in which we have the same number of instrumental variables as parameters. So, we will consider the problem in general terms. The estimation criterion for nonlinear instrumental variables is a quadratic form,

$$\begin{aligned}\text{Min}_{\boldsymbol{\beta}}\, S(\boldsymbol{\beta}) &= \tfrac{1}{2}\big\{[\mathbf{y}-\mathbf{h}(\mathbf{X},\boldsymbol{\beta})]'\mathbf{Z}\big\}(\mathbf{Z}'\mathbf{Z})^{-1}\big\{\mathbf{Z}'[\mathbf{y}-\mathbf{h}(\mathbf{X},\boldsymbol{\beta})]\big\}\\ &= \tfrac{1}{2}\boldsymbol{\varepsilon}(\boldsymbol{\beta})'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}(\boldsymbol{\beta}).\end{aligned}$$

The first-order conditions for minimization of this weighted sum of squares are

$$\frac{\partial S(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = -\mathbf{X}^{0\prime}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{\varepsilon}(\boldsymbol{\beta}) = \mathbf{0}.$$

This result is the same one we had for the linear model with $\mathbf{X}^0$ in the role of $\mathbf{X}$. You should check that when $\boldsymbol{\varepsilon}(\boldsymbol{\beta}) = \mathbf{y}-\mathbf{X}\boldsymbol{\beta}$, our results for the linear model in Section 9.5.1 are replicated exactly. This problem, however, is highly nonlinear in most cases, and the repeated least squares approach is unlikely to be effective. But it is a straightforward minimization problem in the frameworks of Appendix E, and instead, we can just treat estimation here as a problem in nonlinear optimization.

We have approached the formulation of this instrumental variables estimator more or less strategically. However, there is a more structured approach. The orthogonality condition

$$\text{plim}(1/n)\mathbf{Z}'\boldsymbol{\varepsilon} = \mathbf{0}$$

defines a GMM estimator. With the homoscedasticity and nonautocorrelation assumption, the resultant minimum distance estimator produces precisely the criterion function suggested above. We will revisit this estimator in this context, in Chapter 18.

With well-behaved *pseudoregressors* and instrumental variables, we have the general result for the nonlinear instrumental variables estimator; this result is discussed at length in Davidson and MacKinnon (1993).

THEOREM 9.3 Asymptotic Distribution of the Nonlinear Instrumental Variables Estimator

With well-behaved instrumental variables and pseudoregressors,

$$\mathbf{b}_{\mathrm{IV}} \overset{a}{\sim} N\left[\boldsymbol{\beta}, \sigma^2\left(\mathbf{Q}^0_{\mathbf{xz}}(\mathbf{Q}_{\mathbf{zz}})^{-1}\mathbf{Q}^0_{\mathbf{zx}}\right)^{-1}\right].$$

We estimate the asymptotic covariance matrix with

$$\text{Est.Asy. Var}[\mathbf{b}_{\mathrm{IV}}] = \hat{\sigma}^2[\hat{\mathbf{X}}^{0\prime}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\hat{\mathbf{X}}^0]^{-1},$$

where $\hat{\mathbf{X}}^0$ *is* $\mathbf{X}^0$ *computed using* $\mathbf{b}_{\mathrm{IV}}$.

As a final observation, note that the "two-stage least squares" interpretation of the instrumental variables estimator for the linear model still applies here, with respect to the IV estimator. That is, at the final estimates, the first-order conditions (normal equations) imply that

$$\mathbf{X}^{0\prime}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} = \mathbf{X}^{0\prime}\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}^0\boldsymbol{\beta},$$

which says that the estimates satisfy the normal equations for a linear regression of $\mathbf{y}$ (not $\mathbf{y}^0$) on the predictions obtained by regressing the columns of $\mathbf{X}^0$ on $\mathbf{Z}$. The interpretation is not quite the same here, because to compute the predictions of $\mathbf{X}^0$, we must have the estimate of $\boldsymbol{\beta}$ in hand. Thus, this two-stage least squares approach does not show *how to compute* $\mathbf{b}_{\mathrm{IV}}$; it shows a characteristic of $\mathbf{b}_{\mathrm{IV}}$.

Example 9.9 Instrumental Variables Estimates of the Consumption Function

The consumption function in Section 9.3.1 was estimated by nonlinear least squares without accounting for the nature of the data that would certainly induce correlation between $\mathbf{X}^0$ and ε. As we did earlier, we will reestimate this model using the technique of instrumental variables. For this application, we will use the one-period lagged value of consumption and one- and two-period lagged values of income as instrumental variables estimates. Table 9.3 reports the nonlinear least squares and instrumental variables estimates. Since we are using two periods of lagged values, two observations are lost. Thus, the least squares estimates are not the same as those reported earlier.

The instrumental variable estimates differ considerably from the least squares estimates. The differences can be deceiving, however. Recall that the MPC in the model is $\beta Y^{\gamma-1}$. The 2000.4 value for DPI that we examined earlier was 6634.9. At this value, the instrumental variables and least squares estimates of the MPC are 0.8567 with an estimated standard error of 0.01234 and 1.08479 with an estimated standard error of 0.008694, respectively. These values do differ a bit but less than the quite large differences in the parameters might have led one to expect. We do note that both of these are considerably greater than the estimate in the linear model, 0.9222 (and greater than one, which seems a bit implausible).

9.5.2 TWO-STEP NONLINEAR LEAST SQUARES ESTIMATION

In this section, we consider a special case of this general class of models in which the nonlinear regression model depends on a second set of parameters that is estimated separately.

TABLE 9.3 Nonlinear Least Squares and Instrumental Variable Estimates

	Instrumental Variables		*Least Squares*	
Parameter	*Estimate*	*Standard Error*	*Estimate*	*Standard Error*
α	627.031	26.6063	468.215	22.788
β	0.040291	0.006050	0.0971598	0.01064
γ	1.34738	0.016816	1.24892	0.1220
σ	57.1681	—	49.87998	—
$\mathbf{e}'\mathbf{e}$	650,369.805	—	495,114.490	—

The model is

$$y = h(\mathbf{x}, \boldsymbol{\beta}, \mathbf{w}, \boldsymbol{\gamma}) + \varepsilon.$$

We consider cases in which the auxiliary parameter $\boldsymbol{\gamma}$ is estimated separately in a model that depends on an additional set of variables $\mathbf{w}$. This first step might be a least squares regression, a nonlinear regression, or a maximum likelihood estimation. The parameters $\boldsymbol{\gamma}$ will usually enter $h(.)$ through some function of $\boldsymbol{\gamma}$ and $\mathbf{w}$, such as an expectation. The second step then consists of a nonlinear regression of y on $h(\mathbf{x}, \boldsymbol{\beta}, \mathbf{w}, \mathbf{c})$ in which $\mathbf{c}$ is the first-round estimate of $\boldsymbol{\gamma}$. To put this in context, we will develop an example.

The estimation procedure is as follows.

1. Estimate $\boldsymbol{\gamma}$ by least squares, nonlinear least squares, or maximum likelihood. We assume that this estimator, however obtained, denoted $\mathbf{c}$, is consistent and asymptotically normally distributed with asymptotic covariance matrix $\mathbf{V}_c$. Let $\hat{\mathbf{V}}_c$ be any appropriate estimator of $\mathbf{V}_c$.
2. Estimate $\boldsymbol{\beta}$ by nonlinear least squares regression of y on $h(\mathbf{x}, \boldsymbol{\beta}, \mathbf{w}, \mathbf{c})$. Let $\sigma^2\mathbf{V}_b$ be the asymptotic covariance matrix of this estimator of $\boldsymbol{\beta}$, assuming $\boldsymbol{\gamma}$ is known and let $s^2\hat{\mathbf{V}}_b$ be any appropriate estimator of $\sigma^2\mathbf{V}_b = \sigma^2(\mathbf{X}^{0\prime}\mathbf{X}^0)^{-1}$, where $\mathbf{X}^0$ is the matrix of pseudoregressors evaluated at the true parameter values $\mathbf{x}_i^0 = \partial h(\mathbf{x}_i, \boldsymbol{\beta}, \mathbf{w}_i, \boldsymbol{\gamma})/\partial\boldsymbol{\beta}$.

The argument for consistency of $\mathbf{b}$ is based on the Slutsky Theorem, D.12 as we treat $\mathbf{b}$ as a function of $\mathbf{c}$ and the data. We require, as usual, well-behaved pseudoregressors. As long as $\mathbf{c}$ is consistent for $\boldsymbol{\gamma}$, the large-sample behavior of the estimator of $\boldsymbol{\beta}$ conditioned on $\mathbf{c}$ is the same as that conditioned on $\boldsymbol{\gamma}$, that is, as if $\boldsymbol{\gamma}$ were known. Asymptotic normality is obtained along similar lines (albeit with greater difficulty). The asymptotic covariance matrix for the two-step estimator is provided by the following theorem.

THEOREM 9.4 Asymptotic Distribution of the Two-Step Nonlinear Least Squares Estimator [Murphy and Topel (1985)]

Under the standard conditions assumed for the nonlinear least squares estimator, the second-step estimator of $\boldsymbol{\beta}$ is consistent and asymptotically normally distributed with asymptotic covariance matrix

$$\mathbf{V}_b^* = \sigma^2\mathbf{V}_b + \mathbf{V}_b[\mathbf{C}\mathbf{V}_c\mathbf{C}' - \mathbf{C}\mathbf{V}_c\mathbf{R}' - \mathbf{R}\mathbf{V}_c\mathbf{C}']\mathbf{V}_b,$$

THEOREM 9.4 (Continued)
where

$$\mathbf{C} = n\ \text{plim}\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^0\hat{\varepsilon}_i^2\left(\frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta}, \mathbf{w}_i, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}'}\right)$$

and

$$\mathbf{R} = n\ \text{plim}\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i^0\hat{\boldsymbol{\varepsilon}}_i\left(\frac{\partial g(\mathbf{w}_i, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}'}\right).$$

The function $\partial g(.)/\partial\boldsymbol{\gamma}$ in the definition of **R** *is the gradient of the ith term in the log-likelihood function if $\boldsymbol{\gamma}$ is estimated by maximum likelihood. (The precise form is shown below.) If $\boldsymbol{\gamma}$ appears as the parameter vector in a regression model,*

$$z_i = f(\mathbf{w}_i, \boldsymbol{\gamma}) + u_i, \qquad \textbf{(9-32)}$$

then $\partial g(.)/\partial\boldsymbol{\gamma}$ will be a derivative of the sum of squared deviations function,

$$\frac{\partial g(.)}{\partial \boldsymbol{\gamma}} = u_i\frac{\partial f(\mathbf{w}_i, \boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}.$$

If this is a linear regression, then the derivative vector is just $\mathbf{w}_i$.

Implementation of the theorem requires that the asymptotic covariance matrix computed as usual for the second-step estimator based on **c** instead of the true $\boldsymbol{\gamma}$ must be corrected for the presence of the estimator **c** in **b**.

Before developing the application, we note how some important special cases are handled. If $\boldsymbol{\gamma}$ enters $h(.)$ as the coefficient vector in a prediction of another variable in a regression model, then we have the following useful results.

Case 1 Linear regression models. If $h(.) = \mathbf{x}_i'\boldsymbol{\beta} + \delta E[z_i \mid \mathbf{w}_i] + \varepsilon_i$, where $E[z_i \mid \mathbf{w}_i] = \mathbf{w}_i'\boldsymbol{\gamma}$, then the two models are just fit by linear least squares as usual. The regression for y includes an additional variable, $\mathbf{w}_i'\mathbf{c}$. Let d be the coefficient on this new variable. Then

$$\hat{\mathbf{C}} = d\sum_{i=1}^{n} e_i^2\mathbf{x}_i\mathbf{w}_i'$$

and

$$\hat{\mathbf{R}} = \sum_{i=1}^{n}(e_iu_i)\mathbf{x}_i\mathbf{w}_i'.$$

Case 2 Uncorrelated linear regression models. In Case 1, if the two regression disturbances are uncorrelated, then $\mathbf{R} = \mathbf{0}$.

Case 2 is general. The terms in **R** vanish asymptotically if the regressions have uncorrelated disturbances, whether either or both of them are linear. This situation will be quite common.

Case 3 Prediction from a nonlinear model. In Cases 1 and 2, if $E[z_i \mid \mathbf{w}_i]$ is a nonlinear function rather than a linear function, then it is only necessary to change $\mathbf{w}_i$ to $\mathbf{w}_i^0 = \partial E[z_i \mid \mathbf{w}_i]/\partial \gamma$—a vector of pseudoregressors—in the definitions of $\mathbf{C}$ and $\mathbf{R}$.

Case 4 Subset of regressors. In case 2 (but not in case 1), if $\mathbf{w}$ contains all the variables that are in $\mathbf{x}$, then the appropriate estimator is simply

$$\mathbf{V}_b^* = s_e^2 \left(1 + \frac{c^2 s_u^2}{s_e^2}\right) (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1},$$

where $\mathbf{X}^*$ includes all the variables in $\mathbf{x}$ as well as the prediction for z.

All these cases carry over to the case of a nonlinear regression function for y. It is only necessary to replace $\mathbf{x}_i$, the actual regressors in the linear model, with $\mathbf{x}_i^0$, the pseudoregressors.

9.5.3 TWO-STEP ESTIMATION OF A CREDIT SCORING MODEL

Greene (1995c) estimates a model of consumer behavior in which the dependent variable of interest is the number of major derogatory reports recorded in the credit history of a sample of applicants for a type of credit card. In fact, this particular variable is one of the most significant determinants of whether an application for a loan or a credit card will be accepted. This dependent variable y is a discrete variable that at any time, for most consumers, will equal zero, but for a significant fraction who have missed several revolving credit payments, it will take a positive value. The typical values are zero, one, or two, but values up to, say, 10 are not unusual. This count variable is modeled using a Poisson regression model. This model appears in Sections B.4.8, 22.2.1, 22.3.7, and 21.9. The probability density function for this discrete random variable is

$$\text{Prob}[y_i = j] = \frac{e^{-\lambda_i}\lambda_i^j}{j!}.$$

The expected value of y_i is λ_i, so depending on how λ_i is specified and despite the unusual nature of the dependent variable, this model is a linear or nonlinear regression model. We will consider both cases, the linear model $E[y_i \mid \mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta}$ and the more common loglinear model $E[y_i \mid \mathbf{x}_i] = e^{\mathbf{x}_i'\boldsymbol{\beta}}$, where $\mathbf{x}_i$ might include such covariates as age, income, and typical monthly credit account expenditure. This model is usually estimated by maximum likelihood. But since it is a bona fide regression model, least squares, either linear or nonlinear, is a consistent, if inefficient, estimator.

In Greene's study, a secondary model is fit for the outcome of the credit card application. Let z_i denote this outcome, coded 1 if the application is accepted, 0 if not. For purposes of this example, we will model this outcome using a **logit** model (see the extensive development in Chapter 21, esp. Section 21.3). Thus

$$\text{Prob}[z_i = 1] = P(\mathbf{w}_i, \boldsymbol{\gamma}) = \frac{e^{\mathbf{w}_i'\boldsymbol{\gamma}}}{1 + e^{\mathbf{w}_i'\boldsymbol{\gamma}}},$$

where $\mathbf{w}_i$ might include age, income, whether the applicants own their own homes, and whether they are self-employed; these are the sorts of variables that "credit scoring" agencies examine.

Finally, we suppose that the probability of acceptance enters the regression model as an additional explanatory variable. (We concede that the power of the underlying theory wanes a bit here.) Thus, our nonlinear regression model is

$$E[y_i \mid \mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta} + \delta P(\mathbf{w}_i, \boldsymbol{\gamma}) \quad \text{(linear)}$$

or

$$E[y_i \mid \mathbf{x}_i] = e^{\mathbf{x}_i'\boldsymbol{\beta} + \delta P(\mathbf{w}_i, \boldsymbol{\gamma})} \quad \text{(loglinear, nonlinear).}$$

The two-step estimation procedure consists of estimation of $\boldsymbol{\gamma}$ by maximum likelihood, then computing $\hat{P}_i = P(\mathbf{w}_i, \mathbf{c})$, and finally estimating by either linear or nonlinear least squares $[\boldsymbol{\beta}, \delta]$ using $\hat{P}_i$ as a constructed regressor. We will develop the theoretical background for the estimator and then continue with implementation of the estimator.

For the Poisson regression model, when the conditional mean function is linear, $\mathbf{x}_i^0 = \mathbf{x}_i$. If it is loglinear, then

$$\mathbf{x}_i^0 = \partial\lambda_i/\partial\boldsymbol{\beta} = \partial\exp(\mathbf{x}_i'\boldsymbol{\beta})/\partial\boldsymbol{\beta} = \lambda_i\mathbf{x}_i,$$

which is simple to compute. When $P(\mathbf{w}_i, \boldsymbol{\gamma})$ is included in the model, the pseudoregressor vector $\mathbf{x}_i^0$ includes this variable and the coefficient vector is $[\boldsymbol{\beta}, \delta]$. Then

$$\hat{\mathbf{V}}_b = \frac{1}{n}\sum_{i=1}^{n}[y_i - h(\mathbf{x}_i, \mathbf{w}_i, \mathbf{b}, \mathbf{c})]^2 \times (\mathbf{X}^{0\prime}\mathbf{X}^0)^{-1},$$

where $\mathbf{X}^0$ is computed at $[\mathbf{b}, d, \mathbf{c}]$, the final estimates.

For the logit model, the gradient of the log-likelihood and the estimator of $\mathbf{V}_c$ are given in Section 21.3.1. They are

$$\partial \ln f(z_i \mid \mathbf{w}_i, \boldsymbol{\gamma})/\partial\boldsymbol{\gamma} = [z_i - P(\mathbf{w}_i, \boldsymbol{\gamma})]\mathbf{w}_i$$

and

$$\hat{\mathbf{V}}_c = \left[\sum_{i=1}^{n}[z_i - P(\mathbf{w}_i, \hat{\boldsymbol{\gamma}})]^2\mathbf{w}_i\mathbf{w}_i'\right]^{-1}.$$

Note that for this model, we are actually inserting a prediction from a regression model of sorts, since $E[z_i \mid \mathbf{w}_i] = P(\mathbf{w}_i, \boldsymbol{\gamma})$. To compute $\mathbf{C}$, we will require

$$\partial h(.)/\partial\boldsymbol{\gamma} = \lambda_i\delta\,\partial P_i/\partial\boldsymbol{\gamma} = \lambda_i\delta P_i(1 - P_i)\mathbf{w}_i.$$

The remaining parts of the corrected covariance matrix are computed using

$$\hat{\mathbf{C}} = \sum_{i=1}^{n}(\hat{\lambda}_i\hat{\mathbf{x}}_i^0\hat{\varepsilon}_i^2)[\hat{\lambda}_i d\hat{P}_i(1 - \hat{P}_i)]\mathbf{w}_i'$$

and

$$\hat{\mathbf{R}} = \sum_{i=1}^{n}(\hat{\lambda}_i\hat{\mathbf{x}}_i^0\hat{\varepsilon}_i)(z_i - \hat{P}_i)\mathbf{w}_i'.$$

(If the regression model is linear, then the three occurrences of λ_i are omitted.)

TABLE 9.4 Two-Step Estimates of a Credit Scoring Model

	Step 1. $P(w_i, \gamma)$		*Step 2.* $\mathrm{E}[y_i \mid x_i] = x_i'\beta + \delta P_i$			*Step 2.* $\mathrm{E}[y_i \mid x_i] = \mathrm{e}^{\mathbf{x}_i'\beta + \delta \mathbf{P}_i}$		
Variable	*Est.*	*St.Er.*	*Est.*	*St.Er.**	*St.Er.**	*Est.*	*St.Er.*	*St.Er.**
Constant	2.7236	1.0970	−1.0628	1.1907	1.2681	−7.1969	6.2708	49.3854
Age	−0.7328	0.02961	0.021661	0.018756	0.020089	0.079984	0.08135	0.61183
Income	0.21919	0.14296	0.03473	0.07266	0.082079	−0.1328007	0.21380	1.8687
Self-empl	−1.9439	1.01270						
Own Rent	0.18937	0.49817						
Expend			−0.000787	0.000368	0.000413	−0.28008	0.96429	0.96969
$\boldsymbol{P}(\boldsymbol{w}_i, \boldsymbol{\gamma})$			1.0408	1.0653	1.177299	6.99098	5.7978	49.34414
ln L	−53.925							
$\mathbf{e}'\mathbf{e}$			95.5506			80.31265		
s			0.977496			0.89617		
R^2			0.05433			0.20514		
Mean	0.73		0.36			0.36		

Data used in the application are listed in Appendix Table F9.1. We use the following model:

$$\text{Prob}[z_i = 1] = P(\text{age, income, own rent, self-employed}),$$

$$E[y_i] = h(\text{age, income, expend}).$$

We have used 100 of the 1,319 observations used in the original study. Table 9.4 reports the results of the various regressions and computations. The column denoted St.Er.* contains the corrected standard error. The column marked St.Er. contains the standard errors that would be computed ignoring the two-step nature of the computations. For the linear model, we used $\mathbf{e}'\mathbf{e}/n$ to estimate σ^2.

As expected, accounting for the variability in **c** increases the standard errors of the second-step estimator. The linear model appears to give quite different results from the nonlinear model. But this can be deceiving. In the linear model, $\partial E[y_i \mid \mathbf{x}_i, P_i]/\partial \mathbf{x}_i = \boldsymbol{\beta}$ whereas in the nonlinear model, the counterpart is not $\boldsymbol{\beta}$ but $\lambda_i \boldsymbol{\beta}$. The value of λ_i at the mean values of all the variables in the second-step model is roughly 0.36 (the mean of the dependent variable), so the marginal effects in the nonlinear model are [0.0224, −0.0372, −0.07847, 1.9587], respectively, including P_i but not the constant, which are reasonably similar to those for the linear model. To compute an asymptotic covariance matrix for the estimated marginal effects, we would use the delta method from Sections D.2.7 and D.3.1. For convenience, let $\mathbf{b}_p = [\mathbf{b}', d]'$, and let $\mathbf{v}_i = [\mathbf{x}_i', \hat{P}_i]'$, which just adds P_i to the regressor vector so we need not treat it separately. Then the vector of marginal effects is

$$\mathbf{m} = \exp(\mathbf{v}_i'\mathbf{b}_p) \times \mathbf{b}_p = \lambda_i \mathbf{b}_p.$$

The matrix of derivatives is

$$\mathbf{G} = \partial \mathbf{m}/\partial \mathbf{b}_p = \lambda_i(\mathbf{I} + \mathbf{b}_p \mathbf{v}_i'),$$

so the estimator of the asymptotic covariance matrix for **m** is

$$\text{Est.Asy. Var}[\mathbf{m}] = \mathbf{G}\mathbf{V}_b^*\mathbf{G}'.$$

TABLE 9.5 Maximum Likelihood Estimates of Second-Step Regression Model

	Constant	*Age*	*Income*	*Expend*	*P*
Estimate	−6.3200	0.073106	0.045236	−0.00689	4.6324
Std.Error	3.9308	0.054246	0.17411	0.00202	3.6618
Corr.Std.Error	9.0321	0.102867	0.402368	0.003985	9.918233

One might be tempted to treat λ_i as a constant, in which case only the first term in the quadratic form would appear and the computation would amount simply to multiplying the asymptotic standard errors for $\mathbf{b}_p$ by λ_i. This approximation would leave the asymptotic t ratios unchanged, whereas making the full correction will change the entire covariance matrix. The approximation will generally lead to an understatement of the correct standard errors.

Finally, although this treatment is not discussed in detail until Chapter 18, we note at this point that nonlinear least squares is an inefficient estimator in the Poisson regression model; maximum likelihood is the preferred, efficient estimator. Table 9.5 presents the maximum likelihood estimates with both corrected and uncorrected estimates of the asymptotic standard errors of the parameter estimates. (The full discussion of the model is given in Section 21.9.) The corrected standard errors are computed using the methods shown in Section 17.7. A comparison of these estimates with those in the third set of Table 9.4 suggests the clear superiority of the maximum likelihood estimator.

9.6 SUMMARY AND CONCLUSIONS

In this chapter, we extended the regression model to a form which 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 (since the derivatives of the regression are often nonconstant, in contrast to those in the linear model.) Finally, we added two additional levels of generality to the model. A nonlinear instrumental variables estimator is suggested to accommodate the possibility that the disturbances in the model are correlated with the included variables. In the second application, two-step nonlinear least squares is suggested as a method of allowing a model to be fit while including functions of previously estimated parameters.

Key Terms and Concepts

- Box–Cox transformation
- Consistency
- Delta method
- GMM estimator
- Identification
- Instrumental variables estimator
- Iteration
- Linearized regression model
- LM test
- Logit
- Multicollinearity
- Nonlinear model
- Normalization
- Orthogonality condition
- Overidentifying restrictions
- P_E test
- Pseudoregressors
- Semiparametric
- Starting values
- Translog
- Two-step estimation
- Wald test

Exercises

1. Describe how to obtain nonlinear least squares estimates of the parameters of the model $y = \alpha x^{\beta} + \varepsilon$.
2. Use MacKinnon, White, and Davidson's P_E test to determine whether a linear or loglinear production model is more appropriate for the data in Appendix Table F6.1. (The test is described in Section 9.4.3 and Example 9.8.)
3. 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 F9.2, estimate α, β_k, β_l, and λ by using the scanning method suggested in Section 9.3.2. (Do not forget to scale Y, K, and L by the number of establishments.) Use (9-16), (9-12), and (9-13) 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: $\partial \ln Y/\partial \ln K = K \partial \ln Y/\partial K$.]
4. For the model in Exercise 3, test the hypothesis that $\lambda = 0$ using a Wald test, a likelihood ratio test, and a Lagrange multiplier test. Note that the restricted model is the Cobb–Douglas log-linear model.
5. To extend Zellner and Revankar's model in a fashion similar to theirs, we can use the Box–Cox transformation for the dependent variable as well. Use the method of Example 17.6 (with $\theta = \lambda$) to repeat the study of the preceding two exercises. How do your results change?
6. Verify the following differential equation, which applies to the Box–Cox transformation:

$$\frac{d^i x^{(\lambda)}}{d\lambda^i} = \left(\frac{1}{\lambda}\right)\left[x^{\lambda}(\ln x)^i - \frac{i d^{i-1} x^{(\lambda)}}{d\lambda^{i-1}}\right]. \qquad \textbf{(9-33)}$$

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

$$\lim_{\lambda \to 0} \frac{d^i x^{(\lambda)}}{d\lambda^i} = \frac{(\ln x)^{i+1}}{i+1}. \qquad \textbf{(9-34)}$$

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