7 NONLINEAR, SEMIPARAMETRIC AND NONPARAMETRIC REGRESSION MODELS[1]

7.1 INTRODUCTION

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

$$y = x_1\beta_1 + x_2\beta_2 + \ldots + \varepsilon. \qquad (7\text{-}1)$$

Chapters 2-5 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,

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

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),

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

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

Linear and nonlinear least squares are used to estimate the parameters of the **conditional mean function**, $E[y|\mathbf{x}]$. As we saw in Example 4.5, other relationships between y and $\mathbf{x}$, such as the **conditional median**, might be of interest. Section 7.3 revisits this idea with an examination of the conditional median function and the least absolute deviations estimator. This section will also relax the restriction that the model coefficients are always the same in the different parts of the distribution of y (given $\mathbf{x}$). The LAD estimator estimates the parameters of the conditional median, that is, 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 consider thus far are semiparametric in nature, and less parametric as we move from Section 7.2 to 7.3. The **partially linear regression** examined in Section 7.4 extends (7-1) such that $y = f(x) + \mathbf{z}'\boldsymbol{\beta} + \varepsilon$. The endpoint of this progression is a model in which the relationship between y and x is not forced to conform to a particular parameterized function. Using largely graphical and kernel density methods, we consider in Section 7.5 how to analyze a **nonparametric regression** relationship that essentially imposes little more than $E[y|\mathbf{x}] = h(\mathbf{x})$.

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

[2] 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).

Au: KT "linear regression model" is not in chap. list. See msp 7-28. Where term in chap. list is used

Au: Has LAD been spelled out? If not, spell it out here

Au: The KTs "conditional mean function," "conditional median" and "nonparametric regression" are not in chap. list

7.2 NONLINEAR REGRESSION MODELS

The general form of the nonlinear regression model is

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

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

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

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

Example 7.1 CES Production Function

In Example 6.5, we examined a constant elasticity of substitution production function model:

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

No transformation reduces this equation to one that is linear in the parameters. In Example 6.5, a linear Taylor series approximation to this function around the point $\rho = 0$ is used to produce an intrinsically linear equation that can be fit by least squares. Nonetheless, the underlying model in (7-5) is nonlinear in the sense that interests us in this chapter.

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

Example 11.2 Translog Demand System

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

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

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.)

7.2.1

~~11.2.1~~ 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:

1. **Functional form:** The conditional mean function for y_i given $\mathbf{x}_i$ is
$$E[y_i \mid \mathbf{x}_i] = h(\mathbf{x}_i, \beta), \quad i = 1, \ldots, n,$$
where $h(\mathbf{x}_i, \beta)$ is a continuously differentiable function of β.
2. **Identifiability of the model parameters:** The parameter vector in the model is identified (estimable) if there is no nonzero parameter $\beta^0 \neq \beta$ such that $h(\mathbf{x}_i, \beta^0) = h(\mathbf{x}_i, \beta)$ 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 ~~11.2~~ 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** $\beta_M = 1$.

Example 7.2 illustrates the problem.

3. **Zero mean of the disturbance:** It follows from Assumption 1 that we may write
$$y_i = h(\mathbf{x}_i, \beta) + \varepsilon_i.$$
where $E[\varepsilon_i \mid h(\mathbf{x}_i, \beta)] = 0$. This states that the disturbance at observation i is uncorrelated with the conditional mean function for all observations in the sample. This is not quite the same as assuming that the disturbances and the exogenous variables are uncorrelated, which is the familiar assumption, however.
4. **Homoscedasticity and nonautocorrelation:** As in the linear model, we assume conditional homoscedasticity,
$$E[\varepsilon_i^2 \mid h(\mathbf{x}_j, \beta), \ j = 1, \ldots, n] = \sigma^2, \quad \text{a finite constant,} \qquad (11\text{-}2)$$

7-6

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 zero 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 14.3.) 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 16, 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.

(See Section 12.3.)

14

11.2.2 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 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, 2004).

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

$$\ln f(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2) = -(1/2)\left[\ln 2\pi + \ln \sigma^2 + \varepsilon_i^2/\sigma^2\right]. \tag{11-3}$$

For this special case, we have from item D.2 in Theorem 16.2 (on maximum likelihood estimation), that the derivatives of the log density with respect to the parameters have

Example 11.2 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 useable 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

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

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 unuseable as an emprical 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.

Au: KT "identification problem" is not in chap. list.

7.2.2 THE NONLINEAR LEAST SQUARES ESTIMATOR

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

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

The first order conditions for the minimization are

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

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 "pseudo-regressors," $\mathbf{x}_i^0(\beta) = \mathbf{x}_i^0$. We find that the nonlinear least squares estimator is found as the solutions to

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

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

Assumptions 1 and 3 imply that $E[\varepsilon_i \mid h(\mathbf{x}_i,\beta)] = 0$. 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 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, $E[\varepsilon|\mathbf{X}] = \mathbf{0}$ 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 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, 2004).

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

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

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,

$$E\left[\frac{\partial \ln f(y_i \mid \mathbf{x}_i, \boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta}}\right] = E\left[\frac{1}{\sigma^2}\left(\frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)\varepsilon_i\right] = \mathbf{0}, \tag{11-4}$$

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 15.) 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 (11-4) produces a moment condition which will define the nonlinear least squares estimator as a GMM estimator.

Example 11.3 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 (11-10)] 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 11.1 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.

11.2.3 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

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

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

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

7.2.3 LARGE SAMPLE PROPERTIES OF THE NONLINEAR LEAST SQUARES ESTIMATOR

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

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

$$\operatorname{plim} \frac{1}{n}\mathbf{X}^{0\prime}\mathbf{X}^0 = \operatorname{plim} \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_0}\right)\left(\frac{\partial h(\mathbf{x}_i, \boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}_0'}\right) = \mathbf{Q}^0, \tag{11-9}$$

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

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

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

The nonlinear least squares criterion function is

$$S(\mathbf{b}) = \frac{1}{2}\sum_{i=1}^{n}[y_i - h(\mathbf{x}_i, \mathbf{b})]^2 = \frac{1}{2}\sum_{i=1}^{n}e_i^2, \tag{11-10}$$

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

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

In the linear model of Chapter 3, this produces a set of linear equations, the normal equations (3-4). But in this more general case, (11-11) 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 11.1 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 (11-10) and (11-11) 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. 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).

Au: KT "nonlinear least squares" already KT in chap. How also?

7-14

Au: Provide correct x-ref equation number- 3x

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 ten observations from one based on ten thousand. 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.

7-11

AU: Provide x-ref & equation number

THEOREM 11.2 Asymptotic Normality of the Nonlinear Least Squares Estimator

If the pseudoregressors defined in (11-9) are "well behaved," then

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

where

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

The sample estimator of the asymptotic covariance matrix is 7-15

$$\text{Est. Asy. Var}[\mathbf{b}] = \hat{\sigma}^2(\mathbf{X}^{0\prime}\mathbf{X}^0)^{-1}. \quad (11\text{-}12)$$

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 in the class of GMM estimators. (7-12)

The requirement that the matrix in (11-9) 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 7.4 gives an application.

KT

11.2.5 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.3.) The method of Gauss–Newton is often used. In the linearized regression model, if a value of $\boldsymbol{\beta}^0$ is available, then the linearized regression model shown in (11-7) 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

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

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.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 (11-22).] We note, finally, that iteration of the linearized regression, although a very effective algorithm for many problems, does not always work. As does Newton's method, this algorithm sometimes "jumps off" to a wildly errant second iterate, after which it may be impossible to compute the residuals for the next iteration. The choice of starting values for the iterations can be crucial. There is art as well as science in the computation of nonlinear least squares estimates. [See McCullough and Vinod (1999).] In the absence of information about starting values, a workable strategy is to try the Gauss–Newton iteration first. If it 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:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}[y_i - h(\mathbf{x}_i, \mathbf{b})]^2. \qquad (7\text{-}16) \quad \text{(11-13)}$$

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 (2004) argue that on average, (11-13) 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 (11-12): (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,

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \qquad \text{(11-14)}$$

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

7.2.4 HYPOTHESIS TESTING AND PARAMETRIC RESTRICTIONS

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

7.2.4 HYPOTHESIS TESTING AND PARAMETRIC RESTRICTIONS

In most cases, the sorts of hypotheses one would test in this context will involve fairly simple linear restrictions. The tests can be carried out using the usual 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 and Lagrange multiplier tests. For the linear model, these statistics are transformations of the standard F statistic (see Section 16.9.1), 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 Section 16.9.5.

11.4.1 SIGNIFICANCE TESTS FOR RESTRICTIONS: F AND WALD STATISTICS

The hypothesis to be tested is

$$H_0 : \mathbf{r}(\boldsymbol{\beta}) = \mathbf{q}, \tag{11-18}$$

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

$$\mathbf{R}(\boldsymbol{\beta}) = \frac{\partial \mathbf{r}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}, \tag{11-19}$$

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 5.8, 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.[3] 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

$$F[J, n-K] = \frac{[S(\mathbf{b}_*) - S(\mathbf{b})]/J}{S(\mathbf{b})/(n-K)}. \tag{11-20}$$

[3] 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.

in (5-29)

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

$$W = [\mathbf{r}(\mathbf{b}) - \mathbf{q}]'\{\text{Est. Asy. Var}[\mathbf{r}(\mathbf{b}) - \mathbf{q}]\}^{-1}[\mathbf{r}(\mathbf{b}) - \mathbf{q}]$$
$$= [\mathbf{r}(\mathbf{b}) - \mathbf{q}]'\{\mathbf{R}(\mathbf{b})\hat{\mathbf{V}}\mathbf{R}'(\mathbf{b})\}^{-1}[\mathbf{r}(\mathbf{b}) - \mathbf{q}], \quad (11\text{-}21)$$

where

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

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

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

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

AU: KT "Lagrange Multiplier test" already KT in chap. Here also?

14.6.3

~~11.4.2 TESTS BASED ON THE LM STATISTIC~~

new paragraph

FN 4

The **Lagrange multiplier test** is based on the decrease in the sum of squared residuals that would result if the restrictions in the restricted model were released. The formalities of the test are given in Section 16.6.3. For the nonlinear regression model, the test has a particularly appealing form.[4] 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 (11-6). Let $\mathbf{X}^0_*$ be this matrix *computed at the restricted estimates*. Then the Lagrange multiplier statistic for the nonlinear regression model is

Ans: Provide correct x-ref equation no.

$$\text{LM} = \frac{\mathbf{e}_*'\mathbf{X}^0_*[\mathbf{X}^{0\prime}_*\mathbf{X}^0_*]^{-1}\mathbf{X}^{0\prime}_*\mathbf{e}_*}{\mathbf{e}_*'\mathbf{e}_*/n}. \quad (11\text{-}22)$$

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.

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

7.2.5 APPLICATIONS

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

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

In this case, the interaction effect is $\partial^2 E[y|x,z]/\partial x \partial z = \beta_4$. There is no interaction effect if 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 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. 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 F5.1. We will use these to fit the nonlinear consumption function. (Details of the computation of the estimates are given in Section 7.2.6.) The restricted linear and unrestricted nonlinear least squares regression results are shown in Table 7.1.

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	
Var[b]	—		0.000119037	
Var[c]	—		0.00014532	
Cov[b, c]	—		−0.000131491	

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.) The estimates for the linear model are shown in Table 7.1 as well.

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 Section 4.8.1, 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 values of 1.96 for the 5 percent significance level, and we thus reject the linear model in favor of the nonlinear regression. The three procedures for hypotheses produce the same conclusion.

• The *F* statistic is

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

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

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

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

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

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

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

To compute this at the restricted estimates, we use the ordinary least squares estimates for 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. 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.

We are also interested in the marginal propensity to consume. In this expanded model, H0 : γ=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 γ =1 and β =1.) In this model, the marginal propensity to consume is

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

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

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

$$= [cY^{c-1} \quad bY^{c-1}(1 + c\ln Y)] \begin{bmatrix} 0.000119037 & -0.000131491 \\ -0.000131491 & 0.00014532 \end{bmatrix} \begin{bmatrix} cY^{c-1} \\ bY^{c-1}(1 + c\ln Y) \end{bmatrix}$$

$$= 0.00007469,$$

which gives a standard error of 0.0086423. For testing the hypothesis that the MPC is equal to 1.0 in 2000.4 we would refer z = (0.86971 − 1)/0.0086423 = −15.076 to the standard normal table. This difference is certainly statistically significant, so we would reject the hypothesis.

Example 7.5 The Box-Cox Transformation

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

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

Special cases of interest are $\lambda = 1$, which produces a linear transformation, $x^{(1)} = x - 1$, and $\lambda = 0$. When 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,

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

7-23 (11-15)

is a linear regression that can be estimated by least squares. However, if λ in (11-15) 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.[5] To be defined for all values of λ, x must be strictly positive. In most applications, some of the regressors—for example, a dummy variable—will not be transformed. For such a variable, say v_k, $v_k^{(\lambda)} = v_k$, and the relevant derivatives in (11-16) 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.

In some applications, the motivation for the transformation is to program around zero values in a loglinear model. Caves, Christensen, and Trethaway (1980) analyzed the costs of production for railroads providing freight and passenger service. Continuing a long line of literature on the costs of production in regulated industries, a translog cost function (see Section 10.4.2) would be a natural choice for modeling this multiple-output technology. Several of the firms in the study, however, produced no passenger service, which would preclude the use of the translog model. (This model would require the log of zero.) An alternative is the Box–Cox transformation, which is computable for zero output levels. A question does arise in this context (and other similar ones) as to whether zero outputs should be treated the same as nonzero outputs or whether an output of zero represents a discrete corporate decision distinct from other variations in the output levels. In addition, as can be seen in (11-16) this solution is only partial. The zero values of the regressors preclude computation of appropriate standard errors.

Nonlinear least squares is straightforward. In most instances, we can expect to find the least squares value of λ between −2 and 2. Typically, then, λ is estimated by scanning this range for the value that minimizes the sum of squares. Note what happens of there are zeros for x in the sample. Then, 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. 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.

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

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 (~~11-15~~ 7-23) with respect to α, $\boldsymbol{\beta}$, and λ. ~~In the notation of Section 11.2.3, these are~~ The pseudoregressors are

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

We can now use (~~11-12~~ 7-15) and (~~11-13~~ 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. This was the point noted earlier.

It is important to remember that the coefficients in a nonlinear model are not equal to the slopes (or the elasticities) with respect to the variables. For the particular Box–Cox model in (~~11-15~~ 7-23),

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

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

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

AU: Is "Asy. Var." OK? Italics?

The application in Example 7.4 is a Box-Cox model of the sort discussed here. We can rewrite (~~7-15~~ 7-23) as (See Section 7.2.6)

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

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

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

AU: OK to spell out "i.e." in text? that is/

Example 7.6 Interaction Effects in a Loglinear Model for Income

A recent study in health economics is "Incentive Effects in the Demand for Health Care: A Bivariate Panel Count Data Estimation" by Riphahn, Wambach, and Million (2003). The authors were interested in counts of physician visits and hospital visits and in the impact that the presence of private insurance had on the utilization counts of interest, i.e., 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. Among the variables reported in the panel are household income, with numerous other sociodemographic variables such as age, gender and education. For this example, we will model the distribution of income using the last wave of the data set (1988), a cross section with 4483 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 4481 observations. Figures 7.1 and 7.2 display a histogram and a kernel density estimator for the household income variable for these observations.

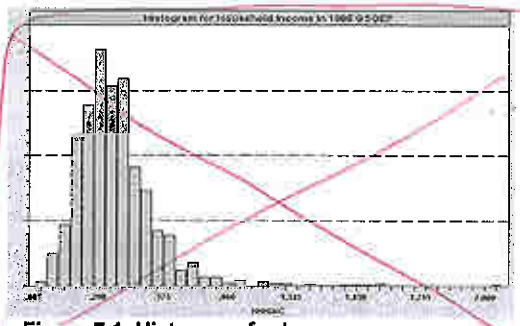

Figure 7.1 Histogram for Income

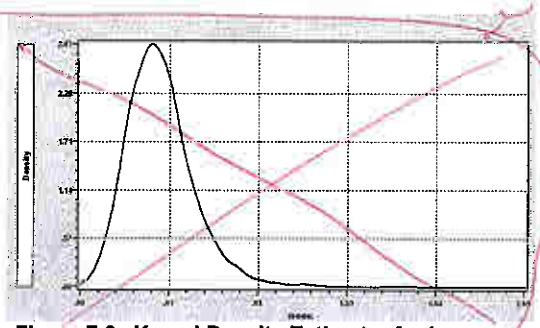

Figure 7.2 Kernel Density Estimator for Income

new figures on next page nsp 7-22

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

$$Income = \exp(\beta_1+\beta_2 Age+\beta_3 Age^2+\beta_4 Education+ \beta_5 Female+\beta_6 Female\times Education+\beta_7 Age\times Education) + \varepsilon.$$

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

Table 7.2 Descriptive Statistics for Variables Used in Nonlinear Regression

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

The data are published on the *Journal of Applied Econometrics* data archive website, at http://qed.econ.queensu.ca/jae/2003-v18.4/riphahn-wambach-million/. The variables in the data file are listed in Appendix Table F11.1. The number of observations in each year varies from one to seven with a total number of observations of 27,326. We will use these data in several examples here and later in the book.

F7.1

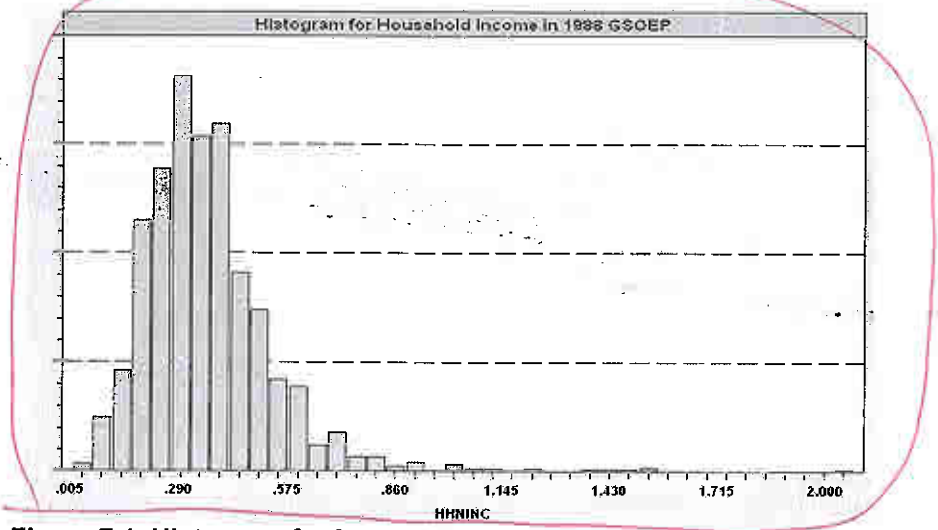

Figure 7.1 Histogram for Income

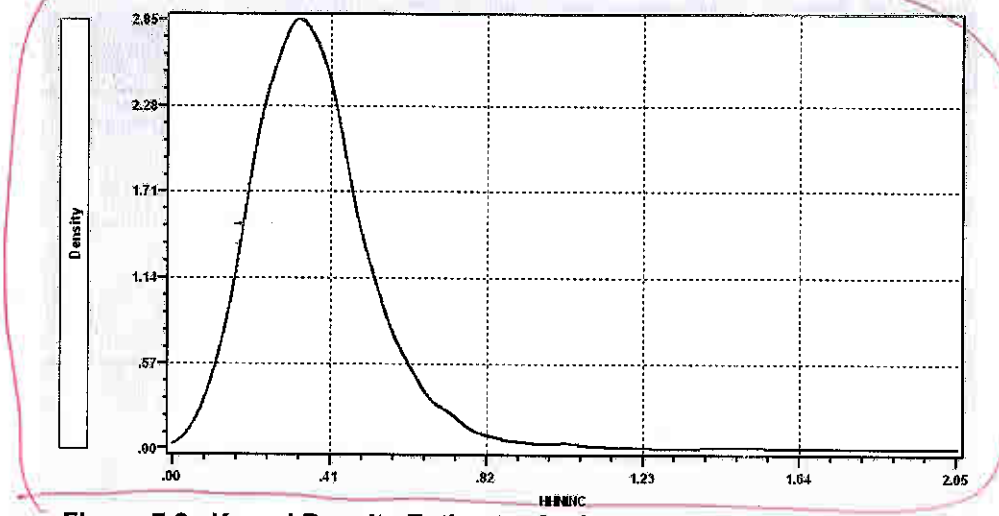

Figure 7.2 Kernel Density Estimator for Income

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

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

(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 $\exp[\ln g(\theta)+\alpha^0+\mathbf{x}'\beta]$, then absorb $\ln g(\theta)$ in the constant term in $\ln E[y|\mathbf{x}] = \alpha+\mathbf{x}'\beta$. The lognormal distribution (see Section B.4.4) is often used to model incomes. For the lognormal random variable,

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

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

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

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,

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

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

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

has a virtue in that the nonlinear least squares estimator will be consistent even if the distributional assumption is incorrect[2] it is *robust* to this type of misspecification since it does not make explicit use of a distributional assumption.

[2] A nonlinear regression treatment of the lognormal model is developed in Amemiya (1973).

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

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

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

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

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

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

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

Table 7.3 Estimated Regression Equations

	Nonlinear Least Squares			Linear Least Squares		
Variable	**Estimate**	**Std.Error**	**t**	**Estimate**	**Std.Error**	**t**
Constant	-2.58070	.17455	14.78	-.13050	.06261	-2.08
Age	.06020	.00615	9.79	.01791	.00214	8.37
Age²	-.00084	.00006082	-13.83	-.00027	.00001985	-13.51
Education	-.00616	.01095	-.56	-.00281	.00418	-.67
Female	.17497	.05986	2.92	.07955	.02339	3.40
Female×Educ	-.01476	.00493	-2.99	-.00685	.00202	-3.39
Age×Educ	.00134	.00024	5.59	.00055	.00009394	5.88
e'e	106.09825			106.24323		
s	.15387			.15410		
R^2	.12005			.11880		

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

Figure 7.3 Expected Incomes

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

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 far from 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 we estimated, that includes the product term, 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) then reassemble the components:

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

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 (2010).

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

7.2.6 COMPUTING THE NONLINEAR LEAST SQUARES ESTIMATOR

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

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

$h(\mathbf{x}, \boldsymbol{\beta})$ at a particular value for the parameter vector, $\boldsymbol{\beta}^0$:

$$h(\mathbf{x}, \boldsymbol{\beta}) \approx h(\mathbf{x}, \boldsymbol{\beta}^0) + \sum_{k=1}^{K} \frac{\partial h(\mathbf{x}, \boldsymbol{\beta}^0)}{\partial \beta_k^0}(\beta_k - \beta_k^0). \tag{11-5}$$

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

$$h(\mathbf{x}, \boldsymbol{\beta}) \approx \left[h(\mathbf{x}, \boldsymbol{\beta}^0) - \sum_{k=1}^{K} \beta_k^0 \left(\frac{\partial h(\mathbf{x}, \boldsymbol{\beta}^0)}{\partial \beta_k^0}\right)\right] + \sum_{k=1}^{K} \beta_k \left(\frac{\partial h(\mathbf{x}, \boldsymbol{\beta}^0)}{\partial \beta_k^0}\right). \tag{11-6}$$

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

$$y^0 = y - h^0 + \mathbf{x}^{0\prime}\boldsymbol{\beta}^0 = \mathbf{x}^{0\prime}\boldsymbol{\beta} + \varepsilon^0. \tag{11-7}$$

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 (11-6). That is,

$$\varepsilon^0 = \varepsilon + \left[h(\mathbf{x}, \boldsymbol{\beta}) - \left\{h^0 - \sum_{k=1}^{K} x_k^0 \beta_k^0 + \sum_{k=1}^{K} x_k^0 \beta_k\right\}\right]. \tag{11-8}$$

Because all the errors are accounted for, (11-7) 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 (11-7) by linear least squares. (Whether this estimator is consistent or not remains to be seen.)

Example 11.4 Linearized Regression

For the model in Example 11.3, the regressors in the linearized equation would be

$$x_1^0 = \frac{\partial h(.)}{\partial \beta_1^0} = 1,$$

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

$$x_3^0 = \frac{\partial h(.)}{\partial \beta_3^0} = \beta_2^0 x e^{\beta_3^0 x}.$$

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

7-29

With a set of value of the parameters β^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 ~~could~~ be linearly regressed on the three variables previously defined to estimate β_1, β_2 and β_3.

7-30

The linearized regression model shown in (~~11-7~~) can be estimated by linear least squares. Once a parameter vector is obtained, it can play the role of a new β^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

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

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 (11-22).] We note, finally, that iteration of the linearized regression, although a very effective algorithm for many problems, does not always work. As does Newton's method, this algorithm sometimes "jumps off" to a wildly errant second iterate, after which it may be impossible to compute the residuals for the next iteration. The choice of starting values for the iterations can be crucial. There is art as well as science in the computation of nonlinear least squares estimates. [See McCullough and Vinod (1999).] In the absence of information about starting values, a workable strategy is to try the Gauss–Newton iteration first. If it 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:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}[y_i - h(\mathbf{x}_i, \mathbf{b})]^2. \tag{11-13}$$

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 (2004) argue that on average, (11-13) 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 (11-12).

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,

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{11-14}$$

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

Example 7.8 Nonlinear Least Squares

Example 7.4 considered analysis of a nonlinear consumption function

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

The linearized regression model is

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

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

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

on

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

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.

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