Chapter 8

Nonlinear Least Squares Theory

What we have analyzed so far are the OLS estimator for linear specifications. Yet, it is hard to believe that linear specifications are "universal" in characterizing all economic relationships. As an alternative, it is natural to consider specifications that are nonlinear in parameters. For example, the function $\alpha+\beta x^\gamma$ offers more flexibility than the simple linear function $\alpha + \beta x$. Such extension also poses some difficulties in practice. First, deciding an appropriate nonlinear function is typically difficult. Second, it is usually cumbersome to estimate nonlinear specifications and analyze the properties of the resulting estimators. Last, but not the least, estimation results of nonlinear specification may not be easily interpreted.

Despite these difficulties, there are more and more empirical evidence suggesting that many economic relationships are in fact nonlinear. Examples include nonlinear production functions, regime switching in output series, and time series models for asymmetric dynamic patterns. In this chapter, we concentrate on estimation of and hypothesis testing for nonlinear regressions and the nonlinear least squares (NLS) estimator. For more information about nonlinear regressions we refer to Gallant (1987), Gallant and White (1988), Davidson and MacKinnon (1993), and Bierens (1994).

8.1 Nonlinear Specifications

We consider the nonlinear specification

$$y = f(\boldsymbol{x};\boldsymbol{\beta}) + e(\boldsymbol{\beta}), \tag{8.1}$$

where f is a given function with $\boldsymbol{x}$ an $\ell \times 1$ vector of explanatory variables and $\boldsymbol{\beta}$ a $k \times 1$ vector of parameters, and $e(\boldsymbol{\beta})$ is the error of this specification. Although $f(\boldsymbol{x};\boldsymbol{\beta})$ may be a linear function, we are primarily interested in the specifications that are nonlinear in

parameters. Clearly, the number of explanatory variables ℓ need not be the same as the number of parameters k.

There are numerous nonlinear specifications considered in empirical applications. A flexible nonlinear specification can be obtained by transforming a regressor x by the so-called *Box-Cox transform*:

$$\frac{x^\gamma - 1}{\gamma}.$$

The Box-Cox transform yields $x - 1$ when $\gamma = 1$, $1 - 1/x$ when $\gamma = -1$, and a value close to $\ln x$ when γ approaches zero. Instead of postulating a functional form, we can allow the data to determine an appropriate function by specifying a model with the Box-Cox transform of regressors. It is typical to apply this transform to positively valued variables.

In the study of firm behavior, the celebrated CES (constant elasticity of substitution) production function suggests to characterize the output y by the following nonlinear function:

$$y = \alpha\big[\delta L^{-\gamma} + (1-\delta)K^{-\gamma}\big]^{-\lambda/\gamma},$$

where L denotes labor, K denotes capital, α, γ, δ and λ are parameters such that $\alpha > 0$, $0 < \delta < 1$ and $\gamma \geq -1$. It can be see that the elasticity of substitution for the CES production function is

$$\frac{\mathrm{d}\ln(K/L)}{\mathrm{d}\ln(\mathrm{MP}_L/\mathrm{MP}_K)} = \frac{1}{(1+\gamma)} \geq 0,$$

where MP denotes marginal product. The CES function includes the linear, Cobb-Douglas, Leontief production functions as special cases. To estimate the CES production function, the following nonlinear specification is usually considered in practice:

$$\ln y = \ln\alpha - \frac{\lambda}{\gamma}\ln\big[\delta L^{-\gamma} + (1-\delta)K^{-\gamma}\big] + e;$$

for a different estimation strategy, see Exercise 8.3. On the other hand, the translog (transcendental logarithmic) production function is nonlinear in variables but linear in parameters:

$$\ln y = \beta_1 + \beta_2 \ln L + \beta_3 \ln K + \beta_4 (\ln L)(\ln K) + \beta_5 (\ln L)^2 + \beta_6 (\ln K)^2,$$

and hence can be estimated by the OLS method.

In the time series context, a nonlinear AR(p) specification is

$$y_t = f(y_{t-1}, \ldots, y_{t-p}) + e_t.$$

To be more specific, the *exponential autoregressive* (EXPAR) specification takes the following form:

$$y_t = \sum_{j=1}^{p} \left[\alpha_j + \beta_j \exp\left(-\gamma y_{t-1}^2\right)\right] y_{t-j} + e_t,$$

which was designed to describe physical vibration whose amplitude depends on the magnitude of y_{t-1}. In some cases one may replace y_{t-1}^2 in the exponential function with y_{t-j}^2 for $j = 1, \ldots, p$.

As another example, consider the *self-exciting threshold autoregressive* (SETAR) specification:

$$y_t = \begin{cases} a_0 + a_1 y_{t-1} + \cdots + a_p y_{t-p} + e_t, & \text{if } y_{t-d} \in (-\infty, c], \\ b_0 + b_1 y_{t-1} + \cdots + b_p y_{t-p} + e_t, & \text{if } y_{t-d} \in (c, \infty), \end{cases} \tag{8.2}$$

where d is known as the "delay parameter" which is an integer between 1 and p, and c is the "threshold parameter." Note that the SETAR model is different from the structural change model in that the parameters switch from one regime to another depending on whether a past realization y_{t-d} exceeds the threshold value c. This specification can be easily extended to allow for r threshold parameters, so that the specification switches among $r+1$ different dynamic structures.

The SETAR specification (8.2) can be written as

$$y_t = a_0 + \sum_{j=1}^{p} a_j y_{t-j} + \left(\delta_0 + \sum_{j=1}^{p} \delta_j y_{t-j}\right) \mathbf{1}_{\{y_{t-d} > c\}} + e_t, \tag{8.3}$$

where $a_j + \delta_j = b_j$, and $\mathbf{1}$ denotes the indicator function. By replacing the indicator function above with a "smooth" function h we are able to avoid abrupt changes of parameters and admit smoother transitions of structures. It is typical to choose the function h as a distribution function, e.g.,

$$h(y_{t-d}; c, s) = \frac{1}{1 + \exp[-(y_{t-d} - c)/s]},$$

where c is still the threshold value and s is a scale parameter. This leads to the following *smooth threshold autoregressive* (STAR) specification:

$$y_t = a_0 + \sum_{j=1}^{p} a_j y_{t-j} + \left(\delta_0 + \sum_{j=1}^{p} \delta_j y_{t-j}\right) h(y_{t-d}; c, s) + e_t,$$

cf. (8.3). Clearly, this specification behaves similarly to a SETAR specification when $|(y_{t-d} - c)/s|$ is very large. See Tong (1990) for other nonlinear time series models and their motivations.

Another well known nonlinear specification is the so-called *artificial neural network* which is designed to mimic the behavior of biological neural systems. This model has been widely used in cognitive science, pattern recognition, engineering, biology and linguistics. A 3-layer neural network can be expressed as

$$f(x_1 \ldots, x_p; \boldsymbol{\beta}) = g\left(\alpha_0 + \sum_{i=1}^{q} \alpha_i \, h\Big(\gamma_{i0} + \sum_{j=1}^{p} \gamma_{ij} x_j\Big)\right), \tag{8.4}$$

where $\boldsymbol{\beta}$ is the parameter vector containing all α and γ, g and h are some pre-specified functions. In the jargon of the neural network literature, this specification contains p "inputs units" in the input layer (each corresponding to an explanatory variable x_j), q "hidden units" in the hidden (middle) layer with the i^{th} hidden-unit activation:

$$h_i = h(\gamma_{i0} + \sum_{j=1}^{p} \gamma_{ij} x_j), \quad i = 1, \ldots, q$$

and one "output unit" in the output layer with the activation $o = g(\beta_0 + \sum_{i=1}^{q} \beta_i h_i)$. The functions h and g are known as "activation functions," and the parameters in these functions are the "connection weights" between units. That is, the input values simultaneously activate q hidden units, and the resulting hidden-unit activations in turn determine an output value. The output values resulted from (8.4) are supposed to capture the behavior of the "target" (dependent) variable y.

In the context of nonlinear regression, we can write

$$y = g\left(\alpha_0 + \sum_{i=1}^{q} \alpha_i \, h\Big(\gamma_{i0} + \sum_{j=1}^{p} \gamma_{ij} x_j\Big)\right) + e.$$

For a multivariate target $\boldsymbol{y}$, networks with multiple outputs can be constructed similarly with g being a vector-valued function. In practice, it is typical to choose h as a "sigmoid" (S-shaped) function bounded within a certain range. Two leading choices of h are the logistic function $h(x) = 1/(1 + e^{-x})$ which is bounded between 0 and 1 and the hyperbolic tangent function

$$h(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}},$$

which is bounded between -1 and 1. The function g may be the identity function or the same as h. Different choices of g and h yield different nonlinear regression models.

The class of neural networks is highly nonlinear in parameters but possesses two appealing properties. First, a neural network is a *universal approximator*, in the sense that it is capable of approximating any Borel-measurable function to any degree of accuracy,

provided that the number of hidden units q is sufficiently large. Given this property, neural networks may be understood as series expansions, where hidden-unit activation functions play the role of basis functions. Second, to achieve a given degree of approximation accuracy, neural networks are more parsimonious than other expansions, such as the polynomial and trigonometric expansions, because the number of hidden units can grow at a much slower rate. For more details of artificial neural networks and their relationships to econometrics we refer to Kuan and White (1994), Kuan (2008), and the references cited therein.

8.2 The NLS Method

We consider the nonlinear specification (8.1):

$$y = f(\boldsymbol{x};\boldsymbol{\beta}) + e(\boldsymbol{\beta}),$$

where $f\colon \mathbb{R}^\ell \times \Theta_1 \mapsto \mathbb{R}$, Θ_1 is the parameter space, a compact subspace of $\mathbb{R}^k$, and $e(\boldsymbol{\beta})$ is the specification error. Given T observations of y and $\boldsymbol{x}$, let

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}, \qquad \boldsymbol{f}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T;\boldsymbol{\beta}) = \begin{bmatrix} f(\boldsymbol{x}_1;\boldsymbol{\beta}) \\ f(\boldsymbol{x}_2;\boldsymbol{\beta}) \\ \vdots \\ f(\boldsymbol{x}_T;\boldsymbol{\beta}) \end{bmatrix}.$$

The nonlinear specification (8.1) now can be expressed as

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T;\boldsymbol{\beta}) + \boldsymbol{e}(\boldsymbol{\beta}),$$

where $\boldsymbol{e}(\boldsymbol{\beta})$ is the vector of errors.

8.2.1 NLS Estimator

Our objective is to find a k-dimensional surface that "best" fits the data $(y_t, \boldsymbol{x}_t)$, $t = 1,\ldots,T$. Analogous to the OLS method, the NLS method suggests to minimize the following NLS criterion function with respect to $\boldsymbol{\beta}$:

$$\begin{aligned} Q_T(\boldsymbol{\beta}) &= \frac{1}{T}[\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T;\boldsymbol{\beta})]'[\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T;\boldsymbol{\beta})] \\ &= \frac{1}{T}\sum_{t=1}^{T}[y_t - f(\boldsymbol{x}_t;\boldsymbol{\beta})]^2. \end{aligned} \tag{8.5}$$

Note that Q_T is also a function of the data y_t and x_t; we omit the arguments y_t and $\boldsymbol{x}_t$ for convenience.

The first order condition of the NLS minimization problem is a system of k nonlinear equations with k unknowns:

$$\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}) = -\frac{2}{T}\nabla_{\boldsymbol{\beta}} \boldsymbol{f}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T;\boldsymbol{\beta})\,[\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T;\boldsymbol{\beta})] \stackrel{\text{set}}{=} \mathbf{0},$$

where

$$\nabla_{\boldsymbol{\beta}} \boldsymbol{f}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T;\boldsymbol{\beta}) = \left[\begin{array}{cccc} \nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_1;\boldsymbol{\beta}) & \nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_2;\boldsymbol{\beta}) & \ldots & \nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_T;\boldsymbol{\beta}) \end{array}\right],$$

is a $k \times T$ matrix. A solution to the first order condition is the NLS estimator, denoted as $\hat{\boldsymbol{\beta}}_T$. The NLS estimator does not have a closed form because the first order condition is a system of nonlinear functions; see also Exercise 8.1.

Clearly, $\hat{\boldsymbol{\beta}}_T$ is a desired minimizer of $Q_T(\boldsymbol{\beta})$ provided that the second order conditions hold: $\nabla^2_{\boldsymbol{\beta}} Q_T(\hat{\boldsymbol{\beta}}_T)$ is positive definite. We thus impose the following identification requirement for nonlinear regressions.

[ID-2] $f(\boldsymbol{x};\cdot)$ is twice continuously differentiable in the second argument on Θ_1, such that for given data $(y_t, \boldsymbol{x}_t)$, $t = 1,\ldots,T$, $\nabla^2_{\boldsymbol{\beta}} Q_T(\hat{\boldsymbol{\beta}}_T)$ is positive definite.

Although $\hat{\boldsymbol{\beta}}_T$ is a minimum of $Q_T(\boldsymbol{\beta})$ under [ID-2], it is not necessarily a unique solution. Indeed, for a given data set, there may exist multiple, local minima of $Q_T(\boldsymbol{\beta})$.

Writing $\boldsymbol{f}(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_T;\boldsymbol{\beta})$ as $\boldsymbol{f}(\boldsymbol{\beta})$, we have

$$\nabla^2_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}) = -\frac{2}{T}\nabla^2_{\boldsymbol{\beta}} \boldsymbol{f}(\boldsymbol{\beta})\,[\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\beta})] + \frac{2}{T}[\nabla_{\boldsymbol{\beta}} \boldsymbol{f}(\boldsymbol{\beta})][\nabla_{\boldsymbol{\beta}} \boldsymbol{f}(\boldsymbol{\beta})]'.$$

For linear regressions, $\boldsymbol{f}(\boldsymbol{\beta}) = \boldsymbol{X}\boldsymbol{\beta}$ so that $\nabla_{\boldsymbol{\beta}} \boldsymbol{f}(\boldsymbol{\beta}) = \boldsymbol{X}'$ and $\nabla^2_{\boldsymbol{\beta}} \boldsymbol{f}(\boldsymbol{\beta}) = \mathbf{0}$. It follows that $\nabla^2_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}) = 2(\boldsymbol{X}'\boldsymbol{X})/T$, which is positive definite if, and only if, $\boldsymbol{X}$ has full column rank. This shows that [ID-2] is analogous to [ID-1] for the OLS method. Note, however, that [ID-1] is a global identification condition because $2(\boldsymbol{X}'\boldsymbol{X})/T$ does not depend on $\boldsymbol{\beta}$.

Let $\hat{\boldsymbol{y}}$ denote the vector of NLS fitted values with the t^{th} element $\hat{y}_t = f(\boldsymbol{x}_t, \hat{\boldsymbol{\beta}}_T)$, and $\hat{\boldsymbol{e}}$ denote the vector of NLS residuals $\boldsymbol{y} - \hat{\boldsymbol{y}}$ with the t^{th} element $\hat{e}_t = y_t - \hat{y}_t$. Denote the transpose of $\nabla_{\boldsymbol{\beta}} \boldsymbol{f}(\boldsymbol{\beta})$ as $\boldsymbol{\Xi}(\boldsymbol{\beta})$. Then by the first order condition,

$$\boldsymbol{\Xi}(\hat{\boldsymbol{\beta}}_T)'\hat{\boldsymbol{e}} = [\nabla_{\boldsymbol{\beta}} \boldsymbol{f}(\hat{\boldsymbol{\beta}}_T)]\hat{\boldsymbol{e}} = \mathbf{0}.$$

That is, the residual vector is orthogonal to every column vector of $\boldsymbol{\Xi}(\hat{\boldsymbol{\beta}}_T)$. Geometrically, $\boldsymbol{f}(\boldsymbol{\beta})$ defines a surface on Θ_1, and for any $\boldsymbol{\beta}$ in Θ_1, $\boldsymbol{\Xi}(\boldsymbol{\beta})$ is a k-dimensional linear subspace tangent at the point $\boldsymbol{f}(\boldsymbol{\beta})$. Thus, $\boldsymbol{y}$ orthogonally projects onto this surface at $\boldsymbol{f}(\hat{\boldsymbol{\beta}}_T)$ such that the residual vector is orthogonal to the tangent space at that point. In contrast with linear regressions, there may be more than one orthogonal projection and hence multiple

solutions to the NLS minimization problem. There is also no guarantee that the sum of NLS residuals is zero; see Exercise 8.2.

Remark: Due to the nonlinearity of f, the marginal response to the change of the i^{th} regressor, $\partial f(\boldsymbol{x}_t;\boldsymbol{\beta})/\partial x_{ti}$, may be a function of *all* parameters.

8.2.2 Nonlinear Optimization Algorithms

When a solution to the first order condition of the NLS minimization problem cannot be obtained analytically, NLS estimates must be computed using some numerical methods. In particular, it is common to employ an *iterative algorithm* that starts from some initial value of the parameter and then repeatedly calculates next available value according to a particular rule until an optimum is reached approximately. It should be noted that an iterative algorithm need not locate the global optimum unless the objective function is globally concave (convex). When an objective function has multiple optima, an iterative algorithm is likely to get stuck at one local optimum. Some methods, such as the *simulated annealing algorithm*, have been proposed to search for the global solution. These methods have not yet been standard in practice because they are typically difficult to implement and computationally very demanding. We will therefore confine ourselves to those commonly used "local" methods.

To minimize $Q_T(\boldsymbol{\beta})$, a generic, iterative algorithm can be expressed as:

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} + s^{(i)}\boldsymbol{d}^{(i)}; \tag{8.6}$$

that is, the $(i+1)^{\text{th}}$ iterated value $\boldsymbol{\beta}^{(i+1)}$ is obtained from $\boldsymbol{\beta}^{(i)}$, the value from the previous iteration, by adjusting the amount $s^{(i)}\boldsymbol{d}^{(i)}$, where $\boldsymbol{d}^{(i)}$ characterizes the direction of change in the parameter space and $s^{(i)}$ controls the amount of change. Different algorithms are based on different choices of s and $\boldsymbol{d}$. As maximizing Q_T is equivalent to minimizing $-Q_T$, the methods discussed here are readily modified to the algorithms for maximization problems.

The *gradient method* determines the direction based on the gradient vector. Consider the first-order Taylor expansion of $Q(\boldsymbol{\beta})$ about $\boldsymbol{\beta}^\dagger$:

$$Q_T(\boldsymbol{\beta}) \approx Q_T(\boldsymbol{\beta}^\dagger) + [\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}^\dagger)]'(\boldsymbol{\beta} - \boldsymbol{\beta}^\dagger).$$

Replacing $\boldsymbol{\beta}$ with $\boldsymbol{\beta}^{(i+1)}$ and $\boldsymbol{\beta}^\dagger$ with $\boldsymbol{\beta}^{(i)}$ we have

$$Q_T\big(\boldsymbol{\beta}^{(i+1)}\big) \approx Q_T\big(\boldsymbol{\beta}^{(i)}\big) + \big[\nabla_{\boldsymbol{\beta}} Q_T\big(\boldsymbol{\beta}^{(i)}\big)\big]' s^{(i)}\boldsymbol{d}^{(i)}.$$

This approximation is valid provided that $\boldsymbol{\beta}^{(i+1)}$ is in the neighborhood of $\boldsymbol{\beta}^{(i)}$. Let $\boldsymbol{g}(\boldsymbol{\beta})$ denote the gradient vector of Q_T: $\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta})$, and $\boldsymbol{g}^{(i)}$ denote $\boldsymbol{g}(\boldsymbol{\beta})$ evaluated at $\boldsymbol{\beta}^{(i)}$. Setting

$\boldsymbol{d}^{(i)} = -\boldsymbol{g}^{(i)}$, we have

$$Q_T\big(\boldsymbol{\beta}^{(i+1)}\big) \approx Q_T\big(\boldsymbol{\beta}^{(i)}\big) - s^{(i)}\big[\boldsymbol{g}^{(i)\prime}\boldsymbol{g}^{(i)}\big],$$

where $\boldsymbol{g}^{(i)\prime)}\boldsymbol{g}^{(i)}$ is non-negative. Thus, Q_T would be decreasing when s is a positive (and small) number. Clearly, when $\boldsymbol{\beta}^{(i)}$ is already a minimum of Q_T, $\boldsymbol{g}^{(i)}$ is zero so that no further adjustment is possible. The algorithm (8.6) then becomes

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} - s^{(i)}\boldsymbol{g}^{(i)}. \tag{8.7}$$

Note that with $\boldsymbol{d}^{(i)} = \boldsymbol{g}^{(i)}$,

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} + s^{(i)}\boldsymbol{g}^{(i)},$$

which is an algorithm searching for a maximum of Q_T.

Given the search direction, one may choose $s^{(i)}$ to maximize the step length. To this end, consider the first order condition below:

$$\frac{\partial Q_T\big(\boldsymbol{\beta}^{(i+1)}\big)}{\partial s^{(i)}} = \nabla_{\boldsymbol{\beta}} Q_T\big(\boldsymbol{\beta}^{(i+1)}\big)\,\frac{\partial \boldsymbol{\beta}^{(i+1)}}{\partial s^{(i)}} = -\boldsymbol{g}^{(i+1)\prime}\boldsymbol{g}^{(i)} = 0.$$

Let $\boldsymbol{H}^{(i)}$ denote the Hessian matrix of Q_T evaluated at $\boldsymbol{\beta}^{(i)}$:

$$\boldsymbol{H}^{(i)} = \nabla^2_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(i)}} = \nabla_{\boldsymbol{\beta}}\, g(\boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(i)}}.$$

By Taylor's expansion of g, we have

$$\boldsymbol{g}^{(i+1)} \approx \boldsymbol{g}^{(i)} + \boldsymbol{H}^{(i)}\big(\boldsymbol{\beta}^{(i+1)} - \boldsymbol{\beta}^{(i)}\big) = \boldsymbol{g}^{(i)} - \boldsymbol{H}^{(i)} s^{(i)}\boldsymbol{g}^{(i)}.$$

It follows that

$$0 = \boldsymbol{g}^{(i+1)\prime}\boldsymbol{g}^{(i)} \approx \boldsymbol{g}^{(i)\prime}\boldsymbol{g}^{(i)} - s^{(i)}\boldsymbol{g}^{(i)\prime}\boldsymbol{H}^{(i)}\boldsymbol{g}^{(i)},$$

or equivalently,

$$s^{(i)} = \frac{\boldsymbol{g}^{(i)\prime}\boldsymbol{g}^{(i)}}{\boldsymbol{g}^{(i)\prime}\boldsymbol{H}^{(i)}\boldsymbol{g}^{(i)}},$$

which is non-negative whenever $\boldsymbol{H}^{(i)}$ is positive definite. The algorithm (8.7) now reads:

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} - \left[\frac{\boldsymbol{g}^{(i)\prime}\boldsymbol{g}^{(i)}}{\boldsymbol{g}^{(i)\prime}\boldsymbol{H}^{(i)}\boldsymbol{g}^{(i)}}\right]\boldsymbol{g}^{(i)}, \tag{8.8}$$

which is known as the *steepest descent algorithm*. If $H^{(i)}$ is not positive definite, $s^{(i)}$ is not guaranteed to be non-negative so that this algorithm may point to a wrong direction.

As the steepest descent algorithm adjusts parameters along the opposite of the gradient direction, it may run into difficulty when, e.g., the nonlinear function being optimized is flat around the optimum. The algorithm may iterate back and forth without much progress in approaching an optimum. To alleviate this problem, the *Newton method* also takes into account the second order derivatives. Consider the second-order Taylor expansion of $Q(\boldsymbol{\beta})$ around some $\boldsymbol{\beta}^\dagger$:

$$Q_T(\boldsymbol{\beta}) \approx Q_T(\boldsymbol{\beta}^\dagger) + \boldsymbol{g}^{\dagger\prime}(\boldsymbol{\beta}-\boldsymbol{\beta}^\dagger) + \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}^\dagger)'\boldsymbol{H}^\dagger(\boldsymbol{\beta}-\boldsymbol{\beta}^\dagger),$$

where $\boldsymbol{g}^\dagger$ and $\boldsymbol{H}^\dagger$ are $\boldsymbol{g}$ and $\boldsymbol{H}$ evaluated at $\boldsymbol{\beta}^\dagger$, respectively. Basing on this expansion, the first order condition of $Q_T(\boldsymbol{\beta})$ can be expressed as $\boldsymbol{g}^\dagger + \boldsymbol{H}^\dagger(\boldsymbol{\beta}-\boldsymbol{\beta}^\dagger) \approx \mathbf{0}$, so that

$$\boldsymbol{\beta} \approx \boldsymbol{\beta}^\dagger - (\boldsymbol{H}^\dagger)^{-1}\boldsymbol{g}^\dagger. \tag{8.9}$$

This suggests the following algorithm:

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} - \big(\boldsymbol{H}^{(i)}\big)^{-1}\boldsymbol{g}^{(i)}, \tag{8.10}$$

with the step length 1 and the direction vector $-\big(\boldsymbol{H}^{(i)}\big)^{-1}\boldsymbol{g}^{(i)}$. The algorithm (8.10) is known as the *Newton-Raphson algorithm*.

From Taylor's expansion it is easy to see that

$$Q_T\big(\boldsymbol{\beta}^{(i+1)}\big) - Q_T\big(\boldsymbol{\beta}^{(i)}\big) \approx -\frac{1}{2}\,\boldsymbol{g}^{(i)\prime}\big(\boldsymbol{H}^{(i)}\big)^{-1}\boldsymbol{g}^{(i)},$$

where the right-hand side is negative provided that $\boldsymbol{H}^{(i)}$ is positive definite. When this approximation is good, the Newton-Raphson algorithm usually (but not always) results in a decrease in the value of Q_T. This algorithm may point to a wrong direction if $\boldsymbol{H}^{(i)}$ is not positive definite; for example, when Q is concave at $\boldsymbol{\beta}^i$. When Q_T is (locally) quadratic, the second-order expansion is exact. Then by (8.9), $\boldsymbol{\beta} = \boldsymbol{\beta}^\dagger - (\boldsymbol{H}^\dagger)^{-1}\boldsymbol{g}^\dagger$ must be a minimum of $Q_T(\boldsymbol{\beta})$. This immediately suggests that the Newton-Raphson algorithm (8.10) can reach the minimum in a single step. If the quadratic approximation is good, this algorithm can locate the minimum more quickly than the steepest descent algorithm.

There are some drawbacks of the Newton-Raphson algorithm. First, the Hessian matrix need not be positive definite. Second, the Hessian matrix must be inverted at each iteration step. An algorithm that avoids computing the second order derivatives is the *Gauss-Newton algorithm*. Observe that

$$\boldsymbol{H}(\boldsymbol{\beta}) = -\frac{2}{T}\nabla^2_{\boldsymbol{\beta}}\boldsymbol{f}(\boldsymbol{\beta})[\boldsymbol{y}-\boldsymbol{f}(\boldsymbol{\beta})] + \frac{2}{T}\boldsymbol{\Xi}(\boldsymbol{\beta})'\boldsymbol{\Xi}(\boldsymbol{\beta}),$$

where $\boldsymbol{\Xi}(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}}\boldsymbol{f}(\boldsymbol{\beta})$. When $\boldsymbol{\beta}$ is close to the minimum and $\boldsymbol{f}$ is correctly specified for the conditional mean function, the first term on the right-hand side has mean zero. Ignoring

the first term, an approximation to $\boldsymbol{H}(\boldsymbol{\beta})$ is thus $2\boldsymbol{\Xi}(\boldsymbol{\beta})'\boldsymbol{\Xi}(\boldsymbol{\beta})/T$. This approximation is convenient because it requires only the first order derivatives and is guaranteed to be positive semi-definite. With this approximation, the Newton-Raphson algorithm (8.10) simplifies to

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} + \left[\boldsymbol{\Xi}\left(\boldsymbol{\beta}^{(i)}\right)'\boldsymbol{\Xi}\left(\boldsymbol{\beta}^{(i)}\right)\right]^{-1}\boldsymbol{\Xi}\left(\boldsymbol{\beta}^{(i)}\right)\left[\boldsymbol{y} - \boldsymbol{f}\left(\boldsymbol{\beta}^{(i)}\right)\right]. \tag{8.11}$$

It is easily seen that the adjustment term in (8.11) can be obtained as the OLS estimate of regressing $\boldsymbol{y} - \boldsymbol{f}\left(\boldsymbol{\beta}^{(i)}\right)$ on $\boldsymbol{\Xi}\left(\boldsymbol{\beta}^{(i)}\right)$; this is known as the *Gauss-Newton regression*. The iterated $\boldsymbol{\beta}$ values can be computed by performing the Gauss-Newton regression repeatedly. The performance of the Gauss-Newton algorithm may be quite different from that of the Newton-Raphson algorithm because (8.11) relies on an approximation to the Hessian matrix.

To maintain a correct search direction of the steepest descent and Newton-Raphson algorithms, it is important to ensure that $\boldsymbol{H}^{(i)}$ is positive definite at each iteration. A simple approach is to correct $\boldsymbol{H}^{(i)}$, if necessary, by adding an appropriate matrix to it. A popular correction is

$$\boldsymbol{H}_c^{(i)} = \boldsymbol{H}^{(i)} + c^{(i)}\boldsymbol{I},$$

where $c^{(i)}$ is a positive number chosen to "force" $\boldsymbol{H}_c^{(i)}$ to be a positive definite matrix. Let $\tilde{\boldsymbol{H}} = \boldsymbol{H}^{-1}$. One may also compute

$$\tilde{\boldsymbol{H}}_c^{(i)} = \tilde{\boldsymbol{H}}^{(i)} + c\boldsymbol{I},$$

because it is the inverse of $\boldsymbol{H}^{(i)}$ that matters in the algorithm. Such a correction is used in, for example, the *Marquardt-Levenberg algorithm*.

The *quasi-Newton method*, on the other hand, corrects $\tilde{\boldsymbol{H}}^{(i)}$ iteratively by adding a symmetric, correction matrix $\boldsymbol{C}^{(i)}$:

$$\tilde{\boldsymbol{H}}^{(i+1)} = \tilde{\boldsymbol{H}}^{(i)} + \boldsymbol{C}^{(i)},$$

with the initial value $\tilde{\boldsymbol{H}}^{(0)} = \boldsymbol{I}$. This method includes the Davidon-Fletcher-Powell (DFP) algorithm and the Broydon-Fletcher-Goldfarb-Shanno (BFGS) algorithm, where the latter is the algorithm used in the GAUSS program. In the DFP algorithm,

$$\boldsymbol{C}^{(i)} = \frac{\boldsymbol{\delta}^{(i)}\boldsymbol{\delta}^{(i)\prime}}{\boldsymbol{\delta}^{(i)\prime}\boldsymbol{\gamma}^{(i)}} + \frac{\tilde{\boldsymbol{H}}^{(i)}\boldsymbol{\gamma}^{(i)}\boldsymbol{\gamma}^{(i)\prime}\tilde{\boldsymbol{H}}^{(i)}}{\boldsymbol{\gamma}^{(i)\prime}\tilde{\boldsymbol{H}}^{(i)}\boldsymbol{\gamma}^{(i)}},$$

where $\boldsymbol{\delta}^{(i)} = \boldsymbol{\beta}^{(i+1)} - \boldsymbol{\beta}^{(i)}$ and $\boldsymbol{\gamma}^{(i)} = \boldsymbol{g}^{(i+1)} - \boldsymbol{g}^{(i)}$. The BFGS algorithm contains an additional term in the correction matrix.

To implement an iterative algorithm, one must choose a vector of initial values to start the algorithm and a stopping rule to terminate the iteration procedure. Initial values are usually specified by the researcher or obtained using a random number generator; prior information, if available, should also be taken into account. For example, if the parameter is a probability, the algorithm may be initialized by, say, 0.5 or by a number randomly generated from the uniform distribution on $[0, 1]$. Without prior information, it is also typical to generate initial values from some (e.g., normal) distribution. In practice, one would generate many sets of initial values and then choose the one that leads to a better result (for example, a better fit of data). Of course, this search process is computationally demanding.

When an algorithm results in no further improvement, a stopping rule must be invoked to terminate the iterations. Typically, an algorithm stops when one of the following convergence criteria is met: for a pre-determined, small positive number c,

1. $\left\|\boldsymbol{\beta}^{(i+1)} - \boldsymbol{\beta}^{(i)}\right\| < c$, where $\|\cdot\|$ denotes the Euclidean norm,
2. $\left\|\boldsymbol{g}\left(\boldsymbol{\beta}^{(i)}\right)\right\| < c$, or
3. $\left|Q_T\left(\boldsymbol{\beta}^{(i+1)}\right) - Q_T\left(\boldsymbol{\beta}^{(i)}\right)\right| < c$.

For the Gauss-Newton algorithm, one may stop the algorithm when TR^2 is "close" to zero, where R^2 is the coefficient of determination of the Gauss-Newton regression. As the residual vector must be orthogonal to the tangent space at the optimum, this stopping rule amounts to checking whether the first order condition is satisfied approximately. In some cases, an algorithm may never meet its pre-set convergence criterion and hence keeps on iterating. To circumvent this difficulty, an optimization program usually sets a maximum number for iterations so that the program terminates automatically once the number of iterations reaches this upper bound.

8.3 Asymptotic Properties of the NLS Estimators

Analyzing the asymptotic properties of the NLS estimator is more difficult because this estimator does not have a closed form. We shall establish consistency and asymptotic normality of the NLS estimator using an approach different from that of Chapter 6.

8.3.1 Consistency

When the NLS objective function $Q_T(\boldsymbol{\beta})$ is close to $\mathbb{E}[Q_T(\boldsymbol{\beta})]$ for all $\boldsymbol{\beta}$, it is reasonable to expect that the minimizer of $Q_T(\boldsymbol{\beta})$, i.e., the NLS estimator $\hat{\boldsymbol{\beta}}_T$, is also close to a minimum

of $\mathbb{E}[Q_T(\boldsymbol{\beta})]$. Given that Q_T is nonlinear in $\boldsymbol{\beta}$, the closeness between $Q_T(\boldsymbol{\beta})$ and $\mathbb{E}[Q_T(\boldsymbol{\beta})]$ can be justified by invoking suitable ULLN, as discussed in Section 5.6.

To illustrate how consistency can be obtained, we consider a special case. Suppose that $\mathbb{E}[Q_T(\boldsymbol{\beta})]$ is a continuous function on the compact parameter space Θ_1 such that $\boldsymbol{\beta}_o$ is its unique, global minimum. The NLS estimator $\hat{\boldsymbol{\beta}}_T$ is such that

$$Q_T(\hat{\boldsymbol{\beta}}_T) = \inf_{\Theta_1} Q_T(\boldsymbol{\beta}).$$

Suppose also that Q_T obeys a SULLN That is, there is a set $\Omega_0 \subseteq \Omega$ such that $\mathbb{P}(\Omega_0) = 1$ and

$$\sup_{\boldsymbol{\beta}\in\Theta_1} \left|Q_T(\boldsymbol{\beta}) - \mathbb{E}[Q_T(\boldsymbol{\beta})]\right| \to 0,$$

for all $\omega \in \Omega_0$. Set

$$\epsilon = \inf_{\boldsymbol{\beta}\in B^c\cap\Theta_1} \big(\mathbb{E}[Q_T(\boldsymbol{\beta})] - \mathbb{E}[Q_T(\boldsymbol{\beta}_o)]\big),$$

where B is an open neighborhood of $\boldsymbol{\beta}_o$. Then for $\omega \in \Omega_0$, we can choose T sufficiently large such that

$$\mathbb{E}[Q_T(\hat{\boldsymbol{\beta}}_T)] - Q_T(\hat{\boldsymbol{\beta}}_T) < \frac{\epsilon}{2},$$

and that

$$Q_T(\hat{\boldsymbol{\beta}}_T) - \mathbb{E}[Q_T(\boldsymbol{\beta}_o)] \le Q_T(\boldsymbol{\beta}_o) - \mathbb{E}[Q_T(\boldsymbol{\beta}_o)] < \frac{\epsilon}{2},$$

because the NLS estimator $\hat{\boldsymbol{\beta}}_T$ minimizes $Q_T(\boldsymbol{\beta})$. It follows that for $\omega \in \Omega_0$,

$$\begin{aligned}
&\mathbb{E}[Q_T(\hat{\boldsymbol{\beta}}_T)] - \mathbb{E}[Q_T(\boldsymbol{\beta}_o)] \\
&\qquad \le \mathbb{E}[Q_T(\hat{\boldsymbol{\beta}}_T)] - Q_T(\hat{\boldsymbol{\beta}}_T) + Q_T(\hat{\boldsymbol{\beta}}_T) - \mathbb{E}[Q_T(\boldsymbol{\beta}_o)] \\
&\qquad < \epsilon,
\end{aligned}$$

for all T sufficiently large. In view of the definition of ϵ, $\hat{\boldsymbol{\beta}}_T$ is such that $\mathbb{E}[Q_T(\hat{\boldsymbol{\beta}}_T)]$ is closer to $\mathbb{E}[Q_T(\boldsymbol{\beta}_o)]$ with probability one and hence can not be outside the neighborhood B of $\boldsymbol{\beta}_o$. As B is arbitrary, $\hat{\boldsymbol{\beta}}_T$ must be converging to $\boldsymbol{\beta}_o$ almost surely. Convergence in probability of $\hat{\boldsymbol{\beta}}_T$ to $\boldsymbol{\beta}_o$ can be established using a similar argument; see e.g., Amemiya (1985) and Exercise 8.4.

The preceding discussion illustrates what matters for consistency is the effect of a SULLN (WULLN). Recall from Theorem 5.34 that, to ensure a SULLN (WULLN), Q_T

should obey a SLLN (WLLN) for each $\boldsymbol{\beta} \in \Theta_1$ and also satisfy a Lipschitz-type continuity condition:

$$|Q_T(\boldsymbol{\beta}) - Q_T(\boldsymbol{\beta}^\dagger)| \leq C_T \|\boldsymbol{\beta} - \boldsymbol{\beta}^\dagger\| \quad \text{a.s.},$$

with C_T bounded almost surely (in probability). If the parameter space Θ_1 is compact and convex, we have from the mean-value theorem and the Cauchy-Schwartz inequality that

$$|Q_T(\boldsymbol{\beta}) - Q_T(\boldsymbol{\beta}^\dagger)| \leq \|\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}^\ddagger)\| \, \|\boldsymbol{\beta} - \boldsymbol{\beta}^\dagger\| \quad \text{a.s.},$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^\dagger$ are in Θ_1 and $\boldsymbol{\beta}^\ddagger$ is the mean value of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^\dagger$, in the sense that $|\boldsymbol{\beta} - \boldsymbol{\beta}^\dagger| < |\boldsymbol{\beta}^\ddagger - \boldsymbol{\beta}^\dagger|$. Hence, the Lipschitz-type condition would hold for

$$C_T = \sup_{\boldsymbol{\beta} \in \Theta_1} \nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}).$$

By (8.5), $Q_T(\boldsymbol{\beta}) = T^{-1} \sum_{t=1}^T \big[y_t^2 - 2 y_t f(\boldsymbol{x}_t; \boldsymbol{\beta}) + f(\boldsymbol{x}_t; \boldsymbol{\beta})^2\big]$, so that

$$\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}) = -\frac{2}{T} \sum_{t=1}^T \nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t; \boldsymbol{\beta})[y_t - f(\boldsymbol{x}_t; \boldsymbol{\beta})].$$

Hence, $\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta})$ cannot be almost surely bounded in general. (It would be bounded if, for example, y_t are bounded random variables and both f and $\nabla_{\boldsymbol{\beta}} f$ are bounded functions. These conditions are much too restrictive, however.) On the other hand, it is practically more plausible that $\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta})$ is bounded in probability. It is the case when, for example, $\mathbb{E} \, |\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta})|$ is bounded uniformly in $\boldsymbol{\beta}$. As such, we shall restrict our discussion below to WULLN and weak consistency of $\hat{\boldsymbol{\beta}}_T$.

The discussion above motivates the conditions given below.

[C1] $\{(y_t \ \boldsymbol{w}_t')'\}$ is a sequence of random vectors, and $\boldsymbol{x}_t$ is vector containing some elements of $\mathcal{Y}^{t-1}$ and $\mathcal{W}^t$.

(i) The sequences $\{y_t^2\}$, $\{y_t f(\boldsymbol{x}_t; \boldsymbol{\beta})\}$ and $\{f(\boldsymbol{x}_t; \boldsymbol{\beta})^2\}$ all obey a WLLN for each $\boldsymbol{\beta}$ in Θ_1, where Θ_1 is compact and convex.

(ii) y_t, $f(\boldsymbol{x}_t; \boldsymbol{\beta})$ and $\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t; \boldsymbol{\beta})$ all have bounded second moment uniformly in $\boldsymbol{\beta}$.

[C2] There exists a unique parameter vector $\boldsymbol{\beta}_o$ such that $\mathbb{E}(y_t \mid \mathcal{Y}^{t-1}, \mathcal{W}^t) = f(\boldsymbol{x}_t; \boldsymbol{\beta}_o)$.

Condition [C1] is analogous to [B1] so that stochastic regressors are allowed. [C1](i) regulates that each components of $Q_T(\boldsymbol{\beta})$ obey a standard WLLN. [C1](ii) implies

$$\mathbb{E} \, |\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta})| \leq \frac{2}{T} \sum_{t=1}^T \Big(\|\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t; \boldsymbol{\beta})\|_2 \|y_t\|_2 + \|\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t; \boldsymbol{\beta})\|_2 \|f(\boldsymbol{x}_t; \boldsymbol{\beta})\|_2 \Big) \leq \Delta,$$

for some Δ not depending on $\boldsymbol{\beta}$. This in turn implies $\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta})$ is bounded in probability (uniformly in $\boldsymbol{\beta}$) by Markov's inequality. Condition [C2] is analogous to [B2] and requires $f(\boldsymbol{x}_t;\boldsymbol{\beta})$ been a correct specification of the conditional mean function. Thus, $\boldsymbol{\beta}_o$ globally minimizes $\mathbb{E}[Q_T(\boldsymbol{\beta})]$ because the conditional mean must minimize mean-squared errors.

Theorem 8.1 *Given the nonlinear specification (8.1), suppose that* [C1] *and* [C2] *hold. Then,* $\hat{\boldsymbol{\beta}}_T \xrightarrow{\mathbb{P}} \boldsymbol{\beta}_o$.

Theorem 8.1 is not satisfactory because it only deals with the convergence to the global minimum. As noted in Section 8.2.2, an iterative algorithm is not guaranteed to find a global minimum of the NLS objective function. Hence, it is more reasonable to expect the NLS estimator converges to some local minimum of $\mathbb{E}[Q_T(\boldsymbol{\beta})]$. To avoid much technicality, we shall not discuss local consistency result but assert only that the NLS estimator converges in probability to a local minimum $\boldsymbol{\beta}^*$. In this case, $f(\boldsymbol{x};\boldsymbol{\beta}^*)$ should be understood as an approximation to the conditional mean function.

8.3.2 Asymptotic Normality

Given that the NLS estimator $\hat{\boldsymbol{\beta}}_T$ is weakly consistent for some $\boldsymbol{\beta}^*$, we will sketch a proof that, with more regularity conditions, the suitably normalized NLS estimator is asymptotically distributed as a normal random vector.

By the mean-value expansion of $\nabla_{\boldsymbol{\beta}} Q_T(\hat{\boldsymbol{\beta}}_T)$ about $\boldsymbol{\beta}^*$,

$$\nabla_{\boldsymbol{\beta}} Q_T(\hat{\boldsymbol{\beta}}_T) = \nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}^*) + \nabla^2_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}_T^\dagger)(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*), \tag{8.12}$$

where $\boldsymbol{\beta}_T^\dagger$ is a mean value of $\hat{\boldsymbol{\beta}}_T$ and $\boldsymbol{\beta}^*$. Clearly, the left-hand side of (8.12) is zero because the NLS estimator solves the first order condition. When $\nabla^2_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}_T^\dagger)$ is invertible, we have

$$\begin{aligned}\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*) &= -[\nabla^2_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}_T^\dagger)]^{-1}\sqrt{T}\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}^*) \\ &= -\boldsymbol{H}_T(\boldsymbol{\beta}^*)^{-1}\sqrt{T}\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}^*) + o_{\mathbb{P}}(1),\end{aligned} \tag{8.13}$$

where $\boldsymbol{H}_T(\boldsymbol{\beta}) = \mathbb{E}[\nabla^2_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta})]$.

To see the second equality in (8.13), let vec denote the operator such that for the matrix $\boldsymbol{A}$, $\mathrm{vec}(\boldsymbol{A})$ is the vector that stacks all the column vectors of $\boldsymbol{A}$. By the triangle inequality,

$$\begin{aligned}&\big\|\mathrm{vec}\big[\nabla^2_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}_T^\dagger)\big] - \mathrm{vec}\big[\boldsymbol{H}_T(\boldsymbol{\beta}^*)\big]\big\| \\ &\quad\le \big\|\mathrm{vec}\big[\nabla^2_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}_T^\dagger)\big] - \mathrm{vec}\big[\boldsymbol{H}_T(\boldsymbol{\beta}_T^\dagger)\big]\big\| + \big\|\mathrm{vec}\big[\boldsymbol{H}_T(\boldsymbol{\beta}_T^\dagger)\big] - \mathrm{vec}\big[\boldsymbol{H}_T(\boldsymbol{\beta}^*)\big]\big\|.\end{aligned}$$

The first term on the right-hand side converges to zero in probability, provided that $\nabla^2_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta})$ also obeys a WULLN. As $\boldsymbol{\beta}_T^\dagger$ is a mean value of $\hat{\boldsymbol{\beta}}_T$ and $\boldsymbol{\beta}^*$, weak consistency of

$\hat{\boldsymbol{\beta}}_T$ implies $\boldsymbol{\beta}_T^\dagger$ also converges in probability to $\boldsymbol{\beta}^*$. This shows that, when $\boldsymbol{H}_T(\boldsymbol{\beta})$ is continuous in $\boldsymbol{\beta}$, the second term also converges to zero in probability. Consequently, $\nabla^2_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}_T^\dagger)$ is essentially close to $\boldsymbol{H}_T(\boldsymbol{\beta}^*)$.

By (8.13), $\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*)$ and $-\boldsymbol{H}_T(\boldsymbol{\beta}^*)^{-1}\sqrt{T}\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}^*)$ are asymptotically equivalent. We shall derive the limiting distribution of the latter. Under suitable regularity conditions,

$$\sqrt{T}\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}^*) = -\frac{2}{\sqrt{T}}\sum_{t=1}^T \nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\boldsymbol{\beta}^*)[y_t - f(\boldsymbol{x}_t;\boldsymbol{\beta}^*)]$$

obeys a CLT, i.e., $(\boldsymbol{V}_T^*)^{-1/2}\sqrt{T}\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}^*) \xrightarrow{D} \mathcal{N}(\boldsymbol{0},\, \boldsymbol{I}_k)$, where

$$\boldsymbol{V}_T^* = \operatorname{var}\left(\frac{2}{\sqrt{T}}\sum_{t=1}^T \nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\boldsymbol{\beta}^*)[y_t - f(\boldsymbol{x}_t;\boldsymbol{\beta}^*)]\right).$$

Then for $\boldsymbol{D}_T^* = \boldsymbol{H}_T(\boldsymbol{\beta}^*)^{-1}\boldsymbol{V}_T^*\boldsymbol{H}_T(\boldsymbol{\beta}^*)^{-1}$, we obtain the following asymptotic normality result:

$$(\boldsymbol{D}_T^*)^{-1/2}\boldsymbol{H}_T(\boldsymbol{\beta}^*)^{-1}\sqrt{T}\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}^*) \xrightarrow{D} \mathcal{N}(\boldsymbol{0},\, \boldsymbol{I}_k).$$

By (8.13),

$$(\boldsymbol{D}_T^*)^{-1/2}\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*) \xrightarrow{D} \mathcal{N}(\boldsymbol{0},\, \boldsymbol{I}_k). \tag{8.14}$$

When $\boldsymbol{D}_T^*$ is replaced by its consistent estimator $\widehat{\boldsymbol{D}}_T$, we have

$$\widehat{\boldsymbol{D}}_T^{-1/2}\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*) \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k). \tag{8.15}$$

Consistent estimation of $\boldsymbol{D}_T^*$ is completely analogous to that for linear regression; see Chapter 6.3. First observe that $\boldsymbol{H}_T(\boldsymbol{\beta}^*)$ is

$$\begin{aligned}\boldsymbol{H}_T(\boldsymbol{\beta}^*) = \frac{2}{T}\sum_{t=1}^T \mathbb{E}\big(\big[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\boldsymbol{\beta}^*)\big]\big[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\boldsymbol{\beta}^*)\big]'\big) \\ - \frac{2}{T}\sum_{t=1}^T \mathbb{E}\big(\nabla^2_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\boldsymbol{\beta}^*)\big[y_t - f(\boldsymbol{x}_t;\boldsymbol{\beta}^*)\big]\big),\end{aligned}$$

which can be consistently estimated by its sample counterpart:

$$\widehat{\boldsymbol{H}}_T = \frac{2}{T}\sum_{t=1}^T \big[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\hat{\boldsymbol{\beta}}_T)\big]\big[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\hat{\boldsymbol{\beta}}_T)\big]' - \frac{2}{T}\sum_{t=1}^T \nabla^2_{\boldsymbol{\beta}}\big[f(\boldsymbol{x}_t;\hat{\boldsymbol{\beta}}_T)\hat{e}_t\big].$$

When $\epsilon_t = y_t - f(\boldsymbol{x}_t;\boldsymbol{\beta}^*)$ are uncorrelated with $\nabla^2_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\boldsymbol{\beta}^*)$, $\boldsymbol{H}_T(\boldsymbol{\beta}^*)$ depends only on the expectation of the outer products of $\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\boldsymbol{\beta}^*)$. In this case, $\widehat{\boldsymbol{H}}_T$ simplifies to

$$\widehat{\boldsymbol{H}}_T = \frac{2}{T}\sum_{t=1}^T \big[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\hat{\boldsymbol{\beta}}_T)\big]\big[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\hat{\boldsymbol{\beta}}_T)\big]'.$$

This estimator is analogous to estimating $\boldsymbol{M}_{xx}$ by $\sum_{t=1}^T \boldsymbol{x}_t\boldsymbol{x}_t'/T$ in linear regressions.

If $\boldsymbol{\beta}^* = \boldsymbol{\beta}_o$ so that $f(\boldsymbol{x}_t;\boldsymbol{\beta}_o)$ is the conditional mean of y_t, $\boldsymbol{V}_T^*$ is

$$\boldsymbol{V}_T^o = \frac{4}{T}\sum_{t=1}^T \mathbb{E}\left(\epsilon_t^2\left[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\boldsymbol{\beta}_o)\right]\left[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\boldsymbol{\beta}_o)\right]'\right).$$

When there is conditional homoskedasticity: $\mathbb{E}(\epsilon_t^2|\mathcal{Y}^{t-1},\mathcal{W}^t) = \sigma_o^2$, $\boldsymbol{V}_T^o$ simplifies to

$$\boldsymbol{V}_T^o = \sigma_o^2\frac{4}{T}\sum_{t=1}^T \mathbb{E}\left(\left[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\boldsymbol{\beta}_o)\right]\left[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\boldsymbol{\beta}_o)\right]'\right),$$

which can be consistently estimated by

$$\widehat{\boldsymbol{V}}_T = \hat\sigma_T^2\frac{4}{T}\sum_{t=1}^T\left[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\hat{\boldsymbol{\beta}}_T)\right]\left[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\hat{\boldsymbol{\beta}}_T)\right]',$$

where $\hat\sigma_T^2 = \sum_{t=1}^T \hat e_t^2/T$ is a consistent estimator for σ_o^2. In this case,

$$\widehat{\boldsymbol{D}}_T = \hat\sigma_T^2\left(\frac{1}{T}\sum_{t=1}^T\left[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\hat{\boldsymbol{\beta}}_T)\right]\left[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\hat{\boldsymbol{\beta}}_T)\right]'\right)^{-1}.$$

This estimator is analogous to the standard OLS variance matrix estimator $\hat\sigma_T^2(\boldsymbol{X}'\boldsymbol{X}/T)^{-1}$ for linear regressions.

When there is conditional heteroskedasticity such that $\mathbb{E}(\epsilon_t^2|\mathcal{Y}^{t-1},\mathcal{W}^t)$ are functions of the elements of $\mathcal{Y}^{t-1}$ and $\mathcal{W}^t$, $\boldsymbol{V}_T^o$ can be consistently estimated by

$$\widehat{\boldsymbol{V}}_T = \frac{4}{T}\sum_{t=1}^T \hat e_t^2\left[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\hat{\boldsymbol{\beta}}_T)\right]\left[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\hat{\boldsymbol{\beta}}_T)\right]'.$$

Consequently,

$$\widehat{\boldsymbol{D}}_T = \left(\frac{1}{T}\sum_{t=1}^T\left[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\hat{\boldsymbol{\beta}}_T)\right]\left[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\hat{\boldsymbol{\beta}}_T)\right]'\right)^{-1}\hat{\boldsymbol{V}}_T \left(\frac{1}{T}\sum_{t=1}^T\left[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\hat{\boldsymbol{\beta}}_T)\right]\left[\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\hat{\boldsymbol{\beta}}_T)\right]'\right)^{-1},$$

which is White's heteroskedasticity-consistent covariance matrix estimator for nonlinear regressions. If $\{\epsilon_t\}$ is not a martingale difference sequence with respect to $\mathcal{Y}^{t-1}$ and $\mathcal{W}^t$, $\boldsymbol{V}_T^*$ can still be consistently estimated using a Newey-West type estimator; see Exercise 8.7.

8.4 Large Sample Tests

We again consider testing linear restrictions of parameters: $\boldsymbol{R}\boldsymbol{\beta}^* = \boldsymbol{r}$, where $\boldsymbol{R}$ is a $q \times k$ matrix and $\boldsymbol{r}$ is a $q \times 1$ vector of pre-specified constants. More generally, one may want to test for nonlinear restrictions $\boldsymbol{r}(\boldsymbol{\beta}^*) = \boldsymbol{0}$, where $\boldsymbol{r}$ is now a $\mathbb{R}^q$-valued nonlinear function. By linearizing $\boldsymbol{r}$, the testing principles for linear restrictions carry over to this case; we omit the details.

The Wald test now evaluates the difference between the NLS estimates and the hypothetical values. By the asymptotic normality result (8.15), we have under the null hypothesis that

$$\widehat{\boldsymbol{\Gamma}}_T^{-1/2}\sqrt{T}\boldsymbol{R}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*) = \widehat{\boldsymbol{\Gamma}}_T^{-1/2}\sqrt{T}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r}) \xrightarrow{D} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_q),$$

where $\widehat{\boldsymbol{\Gamma}}_T = \boldsymbol{R}\widehat{\boldsymbol{D}}_T\boldsymbol{R}'$, and $\widehat{\boldsymbol{D}}_T$ is a consistent estimator for $\boldsymbol{D}_T^*$. It follows that the Wald statistic is

$$\mathcal{W}_T = T(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})\widehat{\boldsymbol{\Gamma}}_T^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}_T - \boldsymbol{r})' \xrightarrow{D} \chi^2(q),$$

which is analogous to the Wald statistic based on the OLS estimator in linear regressions.

Remark: A well known problem with the Wald test for nonlinear hypotheses is that the statistic is not invariant with respect to the expression of $\boldsymbol{r}(\boldsymbol{\beta}^*) = \boldsymbol{0}$. For example, the Wald tests perform quite differently against two equivalent hypotheses: $\beta_1\beta_2 = 1$ and $\beta_1 = 1/\beta_2$. See, e.g., Gregory & Veall (1985) and Phillips & Park (1988).

Exercises

8.1 Suppose that $Q_T(\boldsymbol{\beta})$ is quadratic in $\boldsymbol{\beta}$:

$$Q_T(\boldsymbol{\beta}) = a + \boldsymbol{b}'\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{C}\boldsymbol{\beta},$$

where a is a scalar, $\boldsymbol{b}$ a vector and $\boldsymbol{C}$ a symmetric, positive definite matrix. Find the first order condition of minimizing $Q_T(\boldsymbol{\beta})$ and the resulting solution. Is the OLS criterion function (3.2) quadratic in $\boldsymbol{\beta}$?

8.2 Let $\hat{\epsilon}_t = y_t - \hat{y}_t$ denote the t^{th} NLS residuals. Is $\sum_{t=1}^T \hat{\epsilon}_t$ zero in general? Why or why not?

8.3 Given the nonlinear specification of the CES production function

$$\ln y = \ln \alpha - \frac{\lambda}{\gamma} \ln\left[\delta L^{-\gamma} + (1-\delta)K^{-\gamma}\right] + e,$$

find the second order Taylor expansion of $\ln y$ around $\gamma = 0$. How would you estimate this linearized function and how can you calculate the original parameters α, γ, δ and λ?

8.4 Suppose that $\mathbb{E}[Q_T(\boldsymbol{\beta})]$ is a continuous function on the compact parameter space Θ_1 such that $\boldsymbol{\beta}_o$ is its unique, global minimum. Also suppose that the NLS estimator $\hat{\boldsymbol{\beta}}_T$ is such that

$$\mathbb{E}[Q_T(\hat{\boldsymbol{\beta}}_T)] = \inf_{\Theta_1} \mathbb{E}[Q_T(\boldsymbol{\beta})].$$

Prove that when Q_T has a WULLN effect, then $\hat{\boldsymbol{\beta}}_T$ converges in probability to $\boldsymbol{\beta}_o$.

8.5 Apply Theorem 8.1 to discuss the consistency property of the OLS estimator for the linear specification $y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + e_t$.

8.6 Let $\epsilon_t = y_t - f(\boldsymbol{x}_t;\boldsymbol{\beta}_o)$. If $\{\epsilon_t\}$ is a martingale difference sequence with respect to $\mathcal{Y}^{t-1}$ and $\mathcal{W}^t$ such that $\mathbb{E}(\epsilon_t^2 \mid \mathcal{Y}^{t-1}, \mathcal{W}^t) = \sigma_o^2$, state the conditions under which $\hat{\sigma}_T^2 = \sum_{t=1}^T \hat{e}_t^2/T$ is consistent for σ_o^2.

8.7 Let $\epsilon_t = y_t - f(\boldsymbol{x}_t;\boldsymbol{\beta}^*)$, where $\boldsymbol{\beta}^*$ may not be the same as $\boldsymbol{\beta}_o$. If $\{\epsilon_t\}$ is not a martingale difference sequence with respect to $\mathcal{Y}^{t-1}$ and $\mathcal{W}^t$, give consistent estimators for $\boldsymbol{V}_T^*$ and $\boldsymbol{D}_T^*$.

References

Amemiya, Takeshi (1985). *Advanced Econometrics*, Cambridge, MA: Harvard University Press.

Bierens, Herman J. (1994). *Topics in Advanced Econometrics*, New York, NY: Cambridge University Press.

Davidson, Russell and James G. MacKinnon (1993). *Estimation and Inference in Econometrics*, New York, NY: Oxford University Press.

Gallant, A. Ronald (1987). *Nonlinear Statistical Inference*, New York, NY: John Wiley & Sons.

Gallant, A. Ronald and Halbert White (1988). *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*, Oxford, UK: Basil Blackwell.

Gregory, Allan W. and Michael R. Veall (1985). Formulating Wald tests of nonlinear restrictions, *Econometrica*, **53**, 1465–1468.

Kuan, Chung-Ming (2008). Artificial neural networks, in *New Palgrave Dictionary of Economics*, S. N. Durlauf and L. E. Blume (eds.), Palgrave Macmillan.

Kuan, Chung-Ming and Halbert White (1994). Artificial neural networks: An econometric perspective, *Econometric Reviews*, **13**, 1–91.

Phillips, P. C. B. and Joon Park (1988). On the formulation of Wald tests of nonlinear restrictions, *Econometrica*, **56**, 1065–1083.

Tong, Howell (1990). *Nonlinear Time Series, A Dynamical System Approach*, New York, NY: Oxford University Press.