Chapter 8

Nonlinear Least Squares Theory

For real world data, it is hard to believe that linear specifications are "universal" in characterizing all economic relationships. A straightforward extension of linear specifications is to consider specifications that are nonlinear in parameters. For example, the function $\alpha+\beta x^{\gamma}$ offers more flexibility than the simple linear function $\alpha+\beta x$. Although such an extension is quite natural, it also creates various difficulties. 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, more and more empirical evidences show that many economic relationships are in fact nonlinear. Examples include nonlinear production functions, regime switching in output series, and time series models that can capture asymmetric dynamic patterns. In this chapter, we concentrate on the estimation of and hypothesis testing for nonlinear specifications. For more discussion of 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 $e(\boldsymbol{\beta})$ denotes the error of the specification. Note that for a nonlinear specification, the number of explanatory variables ℓ need not be the same as the number of parameters k. This formulation includes the linear specification as a special case with $f(\boldsymbol{x}; \boldsymbol{\beta}) = \boldsymbol{x}'\boldsymbol{\beta}$ and $\ell = k$. Clearly, nonlinear functions that can be expressed in a linear form should be treated as linear specifications. For example, a specification involving a structural change is nonlinear in parameters:

$$y_t = \begin{cases} \alpha + \beta x_t + e_t, & t \le t^*, \\ (\alpha + \delta) + \beta x_t + e_t, & t > t^*, \end{cases}$$

but it is equivalent to the linear specification:

$$y_t = \alpha + \delta D_t + \beta x_t + e_t,$$

where $D_t = 0$ if $t \le t^*$ and $D_t = 1$ if $t > t^*$. Our discussion in this chapter focuses on the specifications that cannot be expressed as linear functions.

There are numerous nonlinear specifications considered in empirical applications. A flexible nonlinear specification is

$$y_t = \alpha + \beta \frac{x_t^\gamma - 1}{\gamma} + e_t,$$

where $(x_t^\gamma - 1)/\gamma$ is the so-called Box-Cox transform of x_t, which yields different functions, depending on the value γ. For example, the Box-Cox transform yields $x_t - 1$ when $\gamma = 1$, $1 - 1/x_t$ when $\gamma = -1$, and a value close to $\ln x_t$ when γ approaches zero. This function is thus more flexible than, e.g., the linear specification $\alpha + \beta x$ and nonlinear specification $\alpha + \beta x^\gamma$. Note that the Box-Cox transformation is often applied to positively valued variables.

In the study of firm behavior, the celebrated CES (constant elasticity of substitution) production function suggests characterizing 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 $\gamma \ge -1$. The elasticity of substitution for a CES production function is

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

where MP denotes marginal product. This 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:

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

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

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

For example, the *exponential autoregressive* (EXPAR) specification takes the following form:

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

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

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

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

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 above can be written as

$$y_t = a_0 + \sum_{j=1}^{p} a_j y_{t-j} + \Big(\Delta_0 + \sum_{j=1}^{p} \Delta_j y_{t-j}\Big) \mathbf{1}_{\{y_{t-d} > c\}} + e_t,$$

where $a_j + \Delta_j = b_j$, and $\mathbf{1}$ denotes the indicator function. To avoid abrupt changes of parameters, one may replace the indicator function with a "smooth" function h so as to allow for smoother transitions of structures. It is typical to choose the function h as a distribution function, e.g.,

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

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

$$y_t = a_0 + \sum_{j=1}^{p} a_j y_{t-j} + \Big(\Delta_0 + \sum_{j=1}^{p} \Delta_j y_{t-j}\Big)\, h(y_{t-d}; c, \delta) + e_t.$$

Clearly, this specification behaves similarly to a SETAR specification when $|(y_{t-d}-c)/\delta|$ is very large. For more nonlinear time series models and their motivations we refer to Tong (1990).

Another well known nonlinear specification is the so-called *artificial neural network* which has been widely used in cognitive science, engineering, biology and linguistics. A 3-layer neural network can be expressed as

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

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 $h_i = h(\gamma_{i0} + \sum_{j=1}^{p} \gamma_{ij} x_j)$, 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," the parameters in these functions are "connection weights." That is, the input values simultaneously activate q hidden units, and these hidden-unit activations in turn determine the output value. The output value is 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. For example, 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. Although the class of neural networks is highly nonlinear in parameters, it possesses two appealing properties. First, a neural network is capable of approximating any Borel-measurable function to any degree of accuracy, provided that the number of hidden units q is sufficiently large. Second, to achieve a given degree of approximation accuracy, neural networks are relatively more parsimonious than, e.g., the polynomial and trignometric expansions. For more details of artificial neural networks and their relationships to econometrics we refer to Kuan and White (1994).

8.2 The Method of Nonlinear Least Squares

Formally, 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 dentoes the parameter space, a 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 Nonlinear Least Squares 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 method of *nonlinear least squares* (NLS) suggests to minimize the following NLS criterion function with respect to $\boldsymbol{\beta}$:

$$\begin{aligned} Q_T(\boldsymbol{\beta}) &= \frac{1}{T}[\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T; \boldsymbol{\beta})]'[\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T; \boldsymbol{\beta})] \\ &= \frac{1}{T}\sum_{t=1}^{T}[y_t - f(\boldsymbol{x}_t; \boldsymbol{\beta})]^2. \end{aligned} \tag{8.2}$$

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

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

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

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 this minimization problem is some $\bar{\boldsymbol{\beta}} \in \Theta_1$ that solves the first order condition: $\nabla_{\boldsymbol{\beta}} Q_T(\bar{\boldsymbol{\beta}}) = 0$, and satisfies the second order condition: $\nabla^2_{\boldsymbol{\beta}} Q_T(\bar{\boldsymbol{\beta}})$ is positive definite. We thus impose the following identification requirement; cf. [ID-1] for linear specifications.

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

While [ID-2] ensures that a minimum of $Q_T(\boldsymbol{\beta})$ can be found, it does not guarantee the uniqueness of this solution. For a a given data set, there may exist multiple solutions to the NLS minimization problem such that each solution is a local minimum of $Q_T(\boldsymbol{\beta})$. This result is stated below; cf. Theorem 3.1.

Theorem 8.1 *Given the specification* (8.1)*, suppose that* [ID-2] *holds. Then, there exists a solution that minimizes the NLS criterion function* (8.2)*.*

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, in effect, analogous to [ID-1] for the OLS method. Comparing to the OLS method, the NLS minimization problem may not have a closed-form solution because the first order condition is a system of nonlinear functions in general; see also Exercise 8.1.

The minimizer of $Q_T(\boldsymbol{\beta})$ is known as the NLS estimator and will be denoted as $\hat{\boldsymbol{\beta}}_T$. 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 $\Xi(\boldsymbol{\beta})$. Then by the first order condition,

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

That is, the residual vector is orthogonal to every column vector of $\Xi(\hat{\boldsymbol{\beta}}_T)$. Geometrically, $\boldsymbol{f}(\boldsymbol{\beta})$ defines a surface on Θ_1, and for any $\boldsymbol{\beta}$ in Θ_1, $\Xi(\boldsymbol{\beta})$ is a k-dimensional linear subspace tangent at the point $\boldsymbol{f}(\boldsymbol{\beta})$. Thus, $\boldsymbol{y}$ is orthogonally projected onto this surface at $\boldsymbol{f}(\hat{\boldsymbol{\beta}}_T)$ so 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 projections 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: The marginal response to the change of the i th regressor is $\partial f(\boldsymbol{x}_t; \boldsymbol{\beta}) / \partial x_{ti}$. Thus, one should be careful in interpreting the estimation results because a parameter in a nonlinear specification is not necessarily the marginal response to the change of a regressor.

8.2.2 Nonlinear Optimization Algorithms

When a solution to the first order condition of the NLS minimization problem cannot be obtained analytically, the NLS estimates must be computed using numerical methods. To optimizing a nonlinear function, an *iterative algorithm* starts from some initial value of the argument in that function and then repeatedly calculates next available value according to a particular rule until an optimum is reached approximately. It should be noted that when there are multiple optima, an iterative algorithm may not be able to locate the global optimum. In fact, it is more common that an algorithm gets stuck at a local optimum, except in some special cases, e.g., when optimizing a globally concave (convex) function. In the literature, several new methods, such as the *simulated annealing algorithm*, have been proposed to find the global solution. These methods have not yet been standard because they are typically difficult to implement and computationally very intensive. We will therefore confine ourselves to those commonly used "local" methods.

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

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

so that the $(i+1)$ 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 resulted from 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.

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

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

Note that this approximation is valid when $\boldsymbol{\beta}^{(i+1)}$ 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 $\boldsymbol{\beta}^{(i)}$. If $\boldsymbol{d}^{(i)} = -\boldsymbol{g}^{(i)}$,

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

As $\boldsymbol{g}^{(i)\prime)}\boldsymbol{g}^{(i)}$ is non-negative, we can find a positive and small enough s such that Q_T is decreasing. Clearly, when $\boldsymbol{\beta}^{(i)}$ is already a minimum of Q_T, $\boldsymbol{g}^{(i)}$ is zero so that no further adjustment is possible. This suggests the following algorithm:

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

Choosing $\boldsymbol{d}^{(i)} = \boldsymbol{g}^{(i)}$ leads to:

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

which can be used to search for a maximum of Q_T.

Given the search direction, one may want to choose $s^{(i)}$ such that the next value of the objective function $Q_T\big(\boldsymbol{\beta}^{(i+1)}\big)$ is a minimum. This suggests that the first order condition below should hold:

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

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

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

Then 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,

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

The step length $s^{(i)}$ is non-negative whenever $\boldsymbol{H}^{(i)}$ is positive definite. The algorithm derived above now reads

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} - \frac{\boldsymbol{g}^{(i)\prime}\boldsymbol{g}^{(i)}}{\boldsymbol{g}^{(i)\prime}\boldsymbol{H}^{(i)}\boldsymbol{g}^{(i)}}\,\boldsymbol{g}^{(i)},$$

which is known as the *steepest descent algorithm.* If $H^{(i)}$ is not positive definite, $s^{(i)}$ may 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. An alternative is to 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. From this expansion, the first order condition of $Q_T(\boldsymbol{\beta})$ may be expressed as

$$\boldsymbol{g}^{\dagger} + \boldsymbol{H}^{\dagger}(\boldsymbol{\beta} - \boldsymbol{\beta}^{\dagger}) \approx \boldsymbol{0},$$

so that $\boldsymbol{\beta} \approx \boldsymbol{\beta}^{\dagger} - (\boldsymbol{H}^{\dagger})^{-1}\boldsymbol{g}^{\dagger}$. This suggests the following algorithm:

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

where the step length is 1, and the direction vector is $-\big(\boldsymbol{H}^{(i)}\big)^{-1}\boldsymbol{g}^{(i)}$. This is also known as the *Newton-Raphson algorithm.* This algorithm is more difficult to implement because it involves matrix inversion at each iteration step.

From Taylor's expansion we can also 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; this happens when, e.g., Q is concave at $\boldsymbol{\beta}^i$. When Q_T is (locally) quadratic with the local minimum $\boldsymbol{\beta}^*$, the second-order expansion about $\boldsymbol{\beta}^*$ is exact, and hence

$$\boldsymbol{\beta} = \boldsymbol{\beta}^* - \boldsymbol{H}(\boldsymbol{\beta}^*)^{-1}\boldsymbol{g}(\boldsymbol{\beta}^*).$$

In this case, the Newton-Raphson algorithm can reach the minimum in a single step. Alternatively, we may also add a step length to the Newton-Raphson algorithm:

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

where $s^{(i)}$ may be found by minimizing $Q\big(\boldsymbol{\beta}^{(i+1)}\big)$. In practice, it is more typical to choose $s^{(i)}$ such that $Q\big(\boldsymbol{\beta}^{(i)}\big)$ is decreasing at each iteration.

A algorithm that avoids computing the second-order derivatives is the so-called *Gauss-Newton algorithm.* When $Q_T(\boldsymbol{\beta})$ is the NLS criterion function,

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

where $\Xi(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}}\boldsymbol{f}(\boldsymbol{\beta})$. It is therefore convenient to ignore the first term on the right-hand side and approximate $\boldsymbol{H}(\boldsymbol{\beta})$ by $2\Xi(\boldsymbol{\beta})'\Xi(\boldsymbol{\beta})/T$. There are some advantages of this approximation. First, only the first-order derivatives need to be computed. Second, this approximation is guaranteed to be positive definite under [ID-2]. The resulting algorithm is

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

Observe that the adjustment term can be obtained as the OLS estimator of regressing $\boldsymbol{y} - \boldsymbol{f}\big(\boldsymbol{\beta}^{(i)}\big)$ on $\Xi\big(\boldsymbol{\beta}^{(i)}\big)$; this regression is thus known as the *Gauss-Newton regression.* The iterated $\boldsymbol{\beta}$ values can be easily computed by performing the Gauss-Newton regression repeatedly. The performance of this algorithm may be quite different from the

Newton-Raphson algorithm because it utilizes only 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 so-called *Marquardt-Levenberg algorithm*.

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

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

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 by random number generation; 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 a 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. $\|\boldsymbol{\beta}^{(i+1)} - \boldsymbol{\beta}^{(i)}\| < c$, where $\| \cdot \|$ denotes the Euclidean norm,

2. $\|\boldsymbol{g}(\boldsymbol{\beta}^{(i)})\| < c$, or

3. $|Q_T(\boldsymbol{\beta}^{(i+1)}) - Q_T(\boldsymbol{\beta}^{(i)})| < c$.

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

8.3.1 Consistency

As the NLS estimator does not have an analytic form in general, a different approach is thus needed to establish NLS consistency. Intuitively, 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 $\boldsymbol{\beta}$, a ULLN must be invoked to justify the closeness between $Q_T(\boldsymbol{\beta})$ and $\mathbb{E}[Q_T(\boldsymbol{\beta})]$, 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 has a SULLN effect, i.e., there is a set $\Omega_0 \subseteq \Omega$ such that $\mathbb{P}(\Omega_0) = 1$ and

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

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

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

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

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

for all T sufficiently large. This shows that, comparing to all $\boldsymbol{\beta}$ outside the neighborhood B of $\boldsymbol{\beta}_o$, $\hat{\boldsymbol{\beta}}_T$ will eventually render $\mathbb{E}[Q_T(\boldsymbol{\beta})]$ closer to $\mathbb{E}[Q_T(\boldsymbol{\beta}_o)]$ with probability one. Thus, $\hat{\boldsymbol{\beta}}_T$ must be in B for large T. As B is arbitrary, $\hat{\boldsymbol{\beta}}_T$ must converge 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 shows 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

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

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

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

Observe that in the NLS context,

$$Q_T(\boldsymbol{\beta}) = \frac{1}{T}\sum_{t=1}^{T}\big(y_t^2 - 2y_t f(\boldsymbol{x}_t;\boldsymbol{\beta}) + f(\boldsymbol{x}_t;\boldsymbol{\beta})^2\big),$$

and

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

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.) 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 $\hat{\boldsymbol{\beta}}_T$.

To proceed we assume that the identification requirement [ID-2] holds with probability one. The discussion above motivates the additional 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

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

for some Δ which does not depend 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 minimizes mean-squared errors.

Theorem 8.2 *Given the nonlinear specification (8.1), suppose that* [C1] *and* [C2] *hold. Then,* $\hat{\boldsymbol{\beta}}_T \xrightarrow{\mathbb{P}} \boldsymbol{\beta}_o$.

Theorem 8.2 is not completely satisfactory because it is concerned 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 that the NLS estimator only converges to some local minimum of $\mathbb{E}[Q_T(\boldsymbol{\beta})]$. A simple proof of such local consistency result is not yet available. We therefore omit the details and assert only that the NLS estimator converges in probability to a local minimum $\boldsymbol{\beta}^*$. Note that $f(\boldsymbol{x}; \boldsymbol{\beta}^*)$ is, at most, 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.

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

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

where $\boldsymbol{\beta}^\dagger_T$ is a mean value of $\hat{\boldsymbol{\beta}}_T$ and $\boldsymbol{\beta}^*$. Clearly, the left-hand side is zero because $\hat{\boldsymbol{\beta}}_T$ is the NLS estimator and hence solves the first order condition. By [ID-2], the Hessian matrix is invertible, so that

$$\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*) = -[\nabla^2_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}^\dagger_T)]^{-1} \sqrt{T} \nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}^*).$$

The asymptotic distribution of $\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*)$ is therefore the same as that of the right-hand side.

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

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

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 $\boldsymbol{\beta}^{\dagger}_T$ is a mean value of $\hat{\boldsymbol{\beta}}_T$ and $\boldsymbol{\beta}^*$, weak consistency of $\hat{\boldsymbol{\beta}}_T$ implies $\boldsymbol{\beta}^{\dagger}_T$ 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, $\nabla^2_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}^{\dagger}_T)$ is essentially close to $\boldsymbol{H}_T(\boldsymbol{\beta}^*)$.

The result above shows that the normalized NLS estimator, $\sqrt{T}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}^*)$, is asymptotically equivalent to

$$-\boldsymbol{H}_T(\boldsymbol{\beta}^*)^{-1}\sqrt{T}\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}^*),$$

and hence they must have the same limiting distribution. 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., $(\boldsymbol{V}^*_T)^{-1/2}\sqrt{T}\nabla_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}^*) \xrightarrow{D} N(\boldsymbol{0},\, \boldsymbol{I}_k)$, 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 immediately obtain the following asymptotic normality result:

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

which in turn implies

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

As in linear regression, asymptotic normality of the normalized NLS estimator remains valid when $\boldsymbol{D}^*_T$ is replaced by its consistent estimator $\hat{\boldsymbol{D}}_T$:

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

Thus, finding a consistent estimator for $\boldsymbol{D}^*_T$ is important in practice.

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

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

which can be consistently estimated by its sample counterpart:

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

Let $\epsilon_t = y_t - f(\boldsymbol{x}_t;\boldsymbol{\beta}^*)$. When ϵ_t 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 product of $\nabla_{\boldsymbol{\beta}} f(\boldsymbol{x}_t;\boldsymbol{\beta}^*)$ so that $\hat{\boldsymbol{H}}_T$ simplifies to

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

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

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

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

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

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

which can be consistently estimated by

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

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,

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

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

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

so that

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

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

This 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 be consistently estimated using a Newey-West type estimator; see Exercise 8.7.

8.4 Hypothesis Testing

We again consider testing linear restrictions of parameters so that the null hypothesis is $\boldsymbol{R}\boldsymbol{\beta}_o = \boldsymbol{r}$, 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 $\boldsymbol{r}(\boldsymbol{\beta}_o) = \boldsymbol{0}$, 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.

The Wald test now evaluates the difference between the NLS estimates and the hypothetical values. When normalized NLS estimates, $T^{1/2}(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_o)$, have an asymptotic normal distribution with asymptotic covariance matrix $\boldsymbol{D}_T$, we have under the null hypothesis

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

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

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

which is of the same form as the Wald statistic based on the OLS estimator.

Remark: A well known problem with the Wald test for nonlinear hypotheses is that the statistic is not invariant with respect to the expressions 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 & Veal (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

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

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 8.2 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.

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

© Chung-Ming Kuan, 2007