Nonlinear Least Squares Theory

CHUNG-MING KUAN

Department of Finance & CRETA

March 9, 2010

Lecture Outline

Nonlinear Specifications

Given the dependent variable y, consider the nonlinear specification:

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

where $\mathbf{x}$ is $\ell \times 1$, $\boldsymbol{\beta}$ is $k \times 1$, and f is a given function. There are many choices of f. A flexible model is to transform one (or several) x by the Box-Cox transform of x:

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

which yields $x - 1$ when $\gamma = 1$, $1 - 1/x$ when $\gamma = -1$, and a value close to $\ln x$ when $\gamma \to 0$.

- The CES (constant elasticity of substitution) production function:

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

where $\alpha > 0$, $0 < \delta < 1$ and $\gamma \geq -1$, which yields:

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

- The translog (transcendental logarithmic) production function:

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

which is linear in parameters; in this case, the OLS method suffices.

Nonlinear Time Series Models

- An exponential autoregressive (EXPAR) model:

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

- A self-exciting threshold autoregressive (SETAR) model:

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

where $1 \leq d \leq p$ is the delay parameter, and c is the threshold parameter. Alternatively,

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

with $a_j + \delta_j = b_j$.

- Replacing the indicator function in SETAR model with a "smooth" function h we obtain the smooth threshold autoregressive (STAR) model:

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

where h is a distribution function, e.g.,

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

with c the threshold value and s a scale parameter. The STAR model admits smooth transition between different regimes, and it behaves like a SETAR model when $(y_{t-d} - c)/s$ is large.

Artificial Neural Networks

A 3-layer neural network can be expressed as

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

which contains p input units, q hidden units, and one output unit. The functions h and g are known as activation functions, and the parameters in these functions are connection weights.

- h is typically an S-shaped function; two leading choices are the logistic function $h(x) = 1/(1 + e^{-x})$ and the hyperbolic tangent function

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

- The function g may be the identity function or the same as h.

Artificial neural networks are designed to mimic the behavior of biological neural systems and have the following properties.

- Universal approximation: Neural network is capable of approximating any Borel-measurable function to any degree of accuracy, provided that q is sufficiently large. In this sense, neural networks can be understood as a series expansion, with hidden units functions as the basis functions.
- Parsimonious model: To achieve a given degree of approximation accuracy, neural networks are simpler than the polynomial and trigonometric expansions, in the sense that the number of hidden units q can grow at a much slower rate.

The NLS Estimator

- The NLS criterion function:

$$Q_T(\beta) = \frac{1}{T}[\mathbf{y} - \mathbf{f}(\mathbf{x}_1, \ldots, \mathbf{x}_T; \beta)]'[\mathbf{y} - \mathbf{f}(\mathbf{x}_1, \ldots, \mathbf{x}_T; \beta)]$$

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

- The first order condition contains k nonlinear equations with k unknowns:

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

where $\nabla_\beta \mathbf{f}(\mathbf{x}_1, \ldots, \mathbf{x}_T; \beta)$ is a $k \times T$ matrix. A solution to the first order condition is the NLS estimator $\hat{\beta}_T$.

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

- While [ID-2] ensures that $\hat{\boldsymbol{\beta}}_T$ is a minimum of $Q_T(\boldsymbol{\beta})$, it does not guarantee the uniqueness of this solution. For a given data set, there may exist multiple, local minima of $Q_T(\boldsymbol{\beta})$.
- For linear regressions, $\mathbf{f}(\boldsymbol{\beta}) = \mathbf{X}\boldsymbol{\beta}$ so that $\nabla_{\boldsymbol{\beta}} \mathbf{f}(\boldsymbol{\beta}) = \mathbf{X}'$ and $\nabla^2_{\boldsymbol{\beta}} \mathbf{f}(\boldsymbol{\beta}) = \mathbf{0}$. It follows that $\nabla^2_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta}) = 2(\mathbf{X}'\mathbf{X})/T$, which is positive definite if, and only if, $\mathbf{X}$ has full column rank. Note that in linear regression, the identification condition does not depend on $\boldsymbol{\beta}$.

Nonlinear Optimization Algorithms

An NLS estimate is usually computed using a numerical method. In particular, an iterative algorithm 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.

A generic, iterative algorithm is

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

That is, the $(i+1)^{\text{th}}$ iterated value $\boldsymbol{\beta}^{(i+1)}$ is obtained from $\boldsymbol{\beta}^{(i)}$ with an adjustment term $s^{(i)}\mathbf{d}^{(i)}$, where $\mathbf{d}^{(i)}$ characterizes the direction of change in the parameter space and $s^{(i)}$ controls the amount of change. Note that an iterative algorithm can only locate a local optimum.

Gradient Method

The first-order Taylor expansion of $Q(\beta)$ about $\beta^\dagger$ is

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

Replacing β with $\beta^{(i+1)}$ and $\beta^\dagger$ with $\beta^{(i)}$,

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

Setting $\mathbf{d}^{(i)} = -\mathbf{g}^{(i)}$, where $\mathbf{g}^{(i)}$ is $\nabla_\beta Q_T(\beta)$ evaluated at $\beta^{(i)}$, we have

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

where $\mathbf{g}^{(i)\prime)}\mathbf{g}^{(i)} \geq 0$. This leads to:

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

Steepest Descent Algorithm

To maximize the step length, note that

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

Let $\mathbf{H}^{(i)} = \nabla^2_{\boldsymbol{\beta}} Q_T(\boldsymbol{\beta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(i)}}$. By Taylor's expansion of g, we have

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

Thus, $0 = \mathbf{g}^{(i+1)\prime}\mathbf{g}^{(i)} \approx \mathbf{g}^{(i)\prime}\mathbf{g}^{(i)} - s^{(i)}\mathbf{g}^{(i)\prime}\mathbf{H}^{(i)}\mathbf{g}^{(i)}$, or equivalently,

$$s^{(i)} = \frac{\mathbf{g}^{(i)\prime}\mathbf{g}^{(i)}}{\mathbf{g}^{(i)\prime}\mathbf{H}^{(i)}\mathbf{g}^{(i)}} \geq 0,$$

when $\mathbf{H}^{(i)}$ is p.d. We obtain the steepest descent algorithm:

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

Newton Method

The Newton method takes into account the second order derivatives. Consider the second-order Taylor expansion of $Q(\beta)$ around some $\beta^\dagger$:

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

The first order condition of $Q_T(\beta)$ is $\mathbf{g}^\dagger + \mathbf{H}^\dagger(\beta - \beta^\dagger) \approx \mathbf{0}$, so that

$$\beta \approx \beta^\dagger - (\mathbf{H}^\dagger)^{-1}\mathbf{g}^\dagger.$$

This suggests the following Newton-Raphson algorithm:

$$\beta^{(i+1)} = \beta^{(i)} - \left(\mathbf{H}^{(i)}\right)^{-1}\mathbf{g}^{(i)},$$

with the step length 1 and the direction vector $-\left(\mathbf{H}^{(i)}\right)^{-1}\mathbf{g}^{(i)}$.

From Taylor's expansion it is easy to see that

$$Q_T\big(\boldsymbol{\beta}^{(i+1)}\big) - Q_T\big(\boldsymbol{\beta}^{(i)}\big) \approx -\frac{1}{2}\,\mathbf{g}^{(i)\prime}\big(\mathbf{H}^{(i)}\big)^{-1}\mathbf{g}^{(i)} \leq 0,$$

provided that $\mathbf{H}^{(i)}$ is p.s.d. Thus, the Newton-Raphson algorithm usually results in a decrease of Q_T.

When Q_T is (locally) quadratic, the second-order expansion is exact, so that $\boldsymbol{\beta} = \boldsymbol{\beta}^\dagger - (\mathbf{H}^\dagger)^{-1}\mathbf{g}^\dagger$ must be a minimum of $Q_T(\boldsymbol{\beta})$. This immediately suggests that the Newton-Raphson algorithm can reach the minimum in a single step. Yet, there are two drawbacks.

- The Hessian matrix need not be positive definite.
- The Hessian matrix must be inverted at each iteration step.

Gauss-Newton Algorithm

Letting $\boldsymbol{\Xi}(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}}\mathbf{f}(\boldsymbol{\beta})$ we have

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

Ignoring the first term, an approximation to $\mathbf{H}(\boldsymbol{\beta})$ is $2\boldsymbol{\Xi}(\boldsymbol{\beta})'\boldsymbol{\Xi}(\boldsymbol{\beta})/T$, which requires only the first order derivatives and is guaranteed to be p.s.d. The Gauss-Newton algorithm utilizes this approximation as

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

Note that the adjustment term can be obtained as the OLS estimate of regressing $\mathbf{y} - \mathbf{f}\left(\boldsymbol{\beta}^{(i)}\right)$ on $\boldsymbol{\Xi}\left(\boldsymbol{\beta}^{(i)}\right)$; this is known as the Gauss-Newton regression.

Other Modifications

To maintain a correct search direction, $\mathbf{H}^{(i)}$ needs to be p.d.

- Correcting $\mathbf{H}^{(i)}$ by: $\mathbf{H}_c^{(i)} = \mathbf{H}^{(i)} + c^{(i)}\mathbf{I}$, where $c^{(i)} > 0$ is chosen to "force" $\mathbf{H}_c^{(i)}$ to be p.d.
- For $\tilde{\mathbf{H}} = \mathbf{H}^{-1}$, one may compute $\tilde{\mathbf{H}}_c^{(i)} = \tilde{\mathbf{H}}^{(i)} + c\mathbf{I}$. Such a correction is used in the Marquardt-Levenberg algorithm.
- The quasi-Newton method corrects $\tilde{\mathbf{H}}^{(i)}$ by a symmetric matrix:

$$\tilde{\mathbf{H}}^{(i+1)} = \tilde{\mathbf{H}}^{(i)} + \mathbf{C}^{(i)}.$$

This is used by the Davidon-Fletcher-Powell (DFP) algorithm and the Broydon-Fletcher-Goldfarb-Shanno (BFGS) algorithm.

Initial Values and Convergence Criteria

- Initial values: Specified by the researcher or obtained using a random number generator. Prior information, if available, should also be taken into account.
- Convergence criteria:
  - $\left\|\boldsymbol{\beta}^{(i+1)} - \boldsymbol{\beta}^{(i)}\right\| < c$, where $\|\cdot\|$ denotes the Euclidean norm,
  - $\left\|\mathbf{g}\left(\boldsymbol{\beta}^{(i)}\right)\right\| < c$, or
  - $\left|Q_T\left(\boldsymbol{\beta}^{(i+1)}\right) - Q_T\left(\boldsymbol{\beta}^{(i)}\right)\right| < c$.

  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.

Digression: Uniform Law of Large Numbers

Consider the function $q(z_t(\omega);\theta)$. It is a r.v. for a given θ and a function of θ for a given ω. Suppose $\{q(z_t;\theta)\}$ obeys a SLLN for each $\theta \in \Theta$:

$$Q_T(\omega;\theta) = \frac{1}{T}\sum_{t=1}^{T} q(z_t(\omega);\theta) \xrightarrow{\text{a.s.}} Q(\theta),$$

where $Q(\theta)$ is non-stochastic. Note that $\Omega_0^c(\theta) = \{\omega : Q_T(\omega;\theta) \not\to Q(\theta)\}$ varies with θ.

- Although $\mathbb{P}(\Omega_0^c(\theta)) = 0$, $\cup_{\theta\in\Theta}\Omega_0^c(\theta)$ is an uncountable union of non-convergence sets and may not have probability zero.
- $\cap_{\theta\in\Theta}\Omega_0(\theta)$ may occur with probability less than one.

When θ also depends on T (e.g., when θ is replaced by an estimator $\tilde{\theta}_T$), there may not exist a finite T^* such that $Q_T(\omega;\tilde{\theta}_T)$ are arbitrarily close to $Q(\omega;\tilde{\theta}_T)$ for all $T > T^*$. Thus, we need a notion of convergence that is uniform on the parameter space Θ.

We say that $Q_T(\omega;\theta)$ converges to $Q(\theta)$ uniformly in θ almost surely (in probability) if

$$\sup_{\theta\in\Theta} |Q_T(\theta) - Q(\theta)| \to 0, \quad \text{a.s. (in probability)}.$$

We also say that $q(z_t(\omega);\theta)$ obey a strong (or weak) uniform law of large numbers (SULLN or WULLN).

Example: Let z_t be i.i.d. with zero mean and

$$q_T(z_t(\omega);\theta) = z_t(\omega) + \begin{cases} T\theta, & 0 \le \theta \le \frac{1}{2T}, \\ 1 - T\theta, & \frac{1}{2T} < \theta \le \frac{1}{T}, \\ 0, & \frac{1}{T} < \theta < \infty. \end{cases}$$

Observe that for $\theta \ge 1/T$ and $\theta = 0$,

$$Q_T(\omega;\theta) = \frac{1}{T}\sum_{t=1}^{T} q_T(z_t;\theta) = \frac{1}{T}\sum_{t=1}^{T} z_t \xrightarrow{\text{a.s.}} 0,$$

by Kolmogorov's SLLN. For a given θ, we can choose T large enough such that $Q_T(\omega;\theta) \xrightarrow{\text{a.s.}} 0$, where 0 is the pointwise limit. Yet for $\Theta = [0,\infty)$,

$$\sup_{\theta\in\Theta} |Q_T(\omega;\theta)| = |\bar{z}_T + 1/2| \xrightarrow{\text{a.s.}} 1/2,$$

so that the uniform limit is different from the pointwise limit.

What is the extra condition needed to ensure SULLN if we already have, for each $\boldsymbol{\theta} \in \Theta$,

$$Q_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}[q_{Tt}(\mathbf{z}_t;\boldsymbol{\theta}) - \mathbb{E}(q_{Tt}(\mathbf{z}_t;\boldsymbol{\theta}))] \xrightarrow{\text{a.s.}} 0.$$

Suppose $Q_T(\boldsymbol{\theta})$ satisfies a Lipschitz-type condition: for $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^\dagger$ in Θ,

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

where $|C_T| \le \Delta$ a.s. and Δ does not depend on $\boldsymbol{\theta}$. Then,

$$\sup_{\boldsymbol{\theta}\in\Theta} |Q_T(\boldsymbol{\theta})| \le \sup_{\boldsymbol{\theta}\in\Theta} |Q_T(\boldsymbol{\theta}) - Q_T(\boldsymbol{\theta}^\dagger)| + |Q_T(\boldsymbol{\theta}^\dagger)|.$$

Given $\epsilon > 0$, we can choose $\boldsymbol{\theta}^\dagger$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^\dagger\| < \epsilon/(2\Delta)$. Then,

$$\sup_{\boldsymbol{\theta}\in\Theta} |Q_T(\boldsymbol{\theta}) - Q_T(\boldsymbol{\theta}^\dagger)| \leq C_T \frac{\epsilon}{2\Delta} \leq \frac{\epsilon}{2},$$

uniformly in T. Also, by pointwise convergence of Q_T, $|Q_T(\boldsymbol{\theta}^\dagger)| < \epsilon/2$ for large T. Consequently, for all T sufficiently large,

$$\sup_{\boldsymbol{\theta}\in\Theta} |Q_T(\boldsymbol{\theta})| \leq \epsilon.$$

This shows that pointwise convergence and a Lipschitz condition on Q_T together suffice for a SULLN or WULLN.

Consistency

The NLS criterion function is $Q_T(\beta) = T^{-1}\sum_{t=1}^{T}[y_t - f(\mathbf{x}_t;\beta)]^2$, and its minimizer is the NLS estimator $\hat{\beta}_T$. Suppose $\mathbb{E}[Q_T(\beta)]$ is continuous on Θ_1 such that β_o is its unique, global minimum. If $Q_T(\beta)$ is close to $\mathbb{E}[Q_T(\beta)]$, we would expect $\hat{\beta}_T$ close to β_o.

To see this, assuming that Q_T obeys a SULLN:

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

for all $\omega \in \Omega_0$ and $\mathbb{P}(\Omega_0) = 1$. Set

$$\epsilon = \inf_{\beta\in B^c\cap\Theta_1}\left(\mathbb{E}[Q_T(\beta)] - \mathbb{E}[Q_T(\beta_o)]\right),$$

for an open neighborhood B of β_o.

For $\omega \in \Omega_0$, we have for large T, $\mathbb{E}[Q_T(\hat{\beta}_T)] - Q_T(\hat{\beta}_T) < \frac{\epsilon}{2}$, and

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

because the NLS estimator $\hat{\beta}_T$ minimizes $Q_T(\beta)$. It follows that

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

for all T sufficiently large. As $\hat{\beta}_T$ is such that $\mathbb{E}[Q_T(\hat{\beta}_T)]$ is closer to $\mathbb{E}[Q_T(\beta_o)]$ with probability one, it can not be outside the neighborhood B of β_o. As B is arbitrary, $\hat{\beta}_T$ must be converging to β_o almost surely.

Q: How do we ensure a SULLN or WULLN?

If Θ_1 is compact and convex, we have from the mean-value theorem and the Cauchy-Schwartz inequality that

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

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

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

with $\nabla_\beta Q_T(\beta) = -2 \sum_{t=1}^T \nabla_\beta f(\mathbf{x}_t; \beta)[y_t - f(\mathbf{x}_t; \beta)]/T$. Note that $\nabla_\beta Q_T(\beta)$ may be bounded in probability, but it may not be bounded in an almost sure sense. (Why?)

We impose the following conditions.

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

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

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

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

Theorem 8.1

Given the nonlinear specification: $y = f(\mathbf{x};\boldsymbol{\beta}) + e(\boldsymbol{\beta})$, suppose that [C1] and [C2] hold. Then, $\hat{\boldsymbol{\beta}}_T \xrightarrow{\mathbb{P}} \boldsymbol{\beta}_o$.

Remark: Theorem 8.1 is not satisfactory because it only deals with the convergence to the global minimum. Yet, 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 converging to some local minimum of $\mathbb{E}[Q_T(\beta)]$. Therefore, we shall, in what follows, assert only that the NLS estimator converges in probability to a local minimum β^* of $\mathbb{E}[Q_T(\beta)]$. In this case, $f(\mathbf{x};\beta^*)$ is, at most, an approximation to the conditional mean function.

Asymptotic Normality

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

$$0 = \nabla_\beta Q_T(\hat{\beta}_T) = \nabla_\beta Q_T(\beta^*) + \nabla_\beta^2 Q_T(\beta_T^\dagger)(\hat{\beta}_T - \beta^*),$$

where $\beta_T^\dagger$ is a mean value of $\hat{\beta}_T$ and β^*. Thus, when $\nabla_\beta^2 Q_T(\beta_T^\dagger)$ is invertible, we have

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

where $\mathbf{H}_T(\beta) = \mathbb{E}[\nabla_\beta^2 Q_T(\beta)]$. That is, $\sqrt{T}(\hat{\beta}_T - \beta^*)$ and $-\mathbf{H}_T(\beta^*)^{-1}\sqrt{T}\nabla_\beta Q_T(\beta^*)$ are asymptotically equivalent.

Under suitable conditions,

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

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

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

Then for $\mathbf{D}_T^* = \mathbf{H}_T(\boldsymbol{\beta}^*)^{-1}\mathbf{V}_T^*\mathbf{H}_T(\boldsymbol{\beta}^*)^{-1}$,

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

By asymptotic equivalence,

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

When $\mathbf{D}_T^*$ is replaced by a consistent estimator $\widehat{\mathbf{D}}_T$,

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

Note that

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

which can be consistently estimated by its sample counterpart:

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

When $\epsilon_t = y_t - f(\mathbf{x}_t; \boldsymbol{\beta}^*)$ are uncorrelated with $\nabla^2_{\boldsymbol{\beta}} f(\mathbf{x}_t; \boldsymbol{\beta}^*)$, $\mathbf{H}_T(\boldsymbol{\beta}^*)$ depends only on the expectation of the outer product of $\nabla_{\boldsymbol{\beta}} f(\mathbf{x}_t; \boldsymbol{\beta}^*)$ so that $\widehat{\mathbf{H}}_T$ may be simplified as

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

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

If $\{\epsilon_t\}$ is not a martingale difference sequence with respect to $\mathcal{Y}^{t-1}$ and $\mathcal{W}^t$, $\mathbf{V}_T^*$ can be consistently estimated using a Newey-West type estimator. This is more likely in practice as the NLS estimator typically converges to a local optimum $\boldsymbol{\beta}^*$.

Wald Tests

- Hypothesis: $\mathbf{R}\boldsymbol{\beta}^* = \mathbf{r}$, where $\mathbf{R}$ is a $q \times k$ selection matrix and $\mathbf{r}$ is a $q \times 1$ vector of pre-specified constants.
- By the asymptotic normality result, we have under the null that

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

  where $\widehat{\boldsymbol{\Gamma}}_T = \mathbf{R}\widehat{\mathbf{D}}_T\mathbf{R}'$, and $\widehat{\mathbf{D}}_T$ is a consistent estimator for $\mathbf{D}_T^*$.
- The Wald statistic is

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

- For nonlinear restrictions $\mathbf{r}(\boldsymbol{\beta}^*) = \mathbf{0}$, the Wald test is not invariant with respect to the form of $\mathbf{r}(\boldsymbol{\beta}) = \mathbf{0}$.