Econometric Methods (Econometrics I)

Lecture 6: M-Estimation

Prof. Dr. Kai Carstensen

Kiel University

Winter Term 2023/2024

Outline of this lecture

1. Nonlinear Least Squares Estimation
2. M-Estimation
3. Identification and Consistency
4. Asymptotic Normality
5. Estimating the Asymptotic Variance
6. Application to Nonlinear Least Squares
7. Inference
8. Optimization Methods

Reference: Wooldridge, Chapter 12.

1. Nonlinear Least Squares Estimation

Examples

So far, we have studied estimation of linear models, in particular the structural model

$$\mathrm{E}(y|\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}.$$

However, there are many examples where the conditional expectation is nonlinear:

- Example 1: nonnegativity ($y > 0$) may be modeled as

$$\mathrm{E}(y|\mathbf{x}) = \exp(\mathbf{x}\boldsymbol{\beta})$$

- Example 2: if y is interpreted as a probability ($0 \leq y \leq 1$), a candidate model is

$$\mathrm{E}(y|\mathbf{x}) = \frac{\exp(\mathbf{x}\boldsymbol{\beta})}{1 + \exp(\mathbf{x}\boldsymbol{\beta})}$$

Nonlinear conditional expectation models

These are examples of the general model (denoting the true parameter vector as $\boldsymbol{\theta}_o$)

$$\mathrm{E}(y|\mathbf{x}) = m(\mathbf{x}, \boldsymbol{\theta}_o).$$

It can be written as

$$y = m(\mathbf{x}, \boldsymbol{\theta}_o) + u, \qquad \mathrm{E}(u|\mathbf{x}) = 0$$

It seems straightforward to estimate this model by nonlinear least squares (NLS), i.e., find the estimator $\hat{\boldsymbol{\theta}}$ that solves

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} N^{-1} \sum_{i=1}^{N} [y_i - m(\mathbf{x}_i, \boldsymbol{\theta})]^2 = \min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} N^{-1} \sum_{i=1}^{N} u_i^2.$$

But can we hope that this estimator is consistent, $\hat{\boldsymbol{\theta}} \xrightarrow{\text{p}} \boldsymbol{\theta}$?

Identification

A necessary condition is identification.

For nonlinear models this requires that the true parameter vector $\boldsymbol{\theta}_o$ be the unique solution to the population analogue

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathrm{E}\left\{[y-m(\mathbf{x},\boldsymbol{\theta})]^2\right\}.$$

The idea is simple: If we find a law of large number which implies

$$N^{-1}\sum_{i=1}^{N}[y_i-m(\mathbf{x}_i,\boldsymbol{\theta})]^2 \xrightarrow{\mathrm{p}} \mathrm{E}\left\{[y-m(\mathbf{x},\boldsymbol{\theta})]^2\right\},$$

then we may hope that the estimator $\hat{\boldsymbol{\theta}}$ which minimizes the sample moment converges towards the population parameter $\boldsymbol{\theta}_o$ which minimizes the population moment.

Is the NLS estimator identified? We need two straightforward assumptions:

Assumption NLS.1: For some $\boldsymbol{\theta}_o \in \boldsymbol{\Theta}$, $\mathrm{E}(y|\mathbf{x}) = m(\mathbf{x}, \boldsymbol{\theta}_o)$.

Assumption NLS.2: $\mathrm{E}\left\{[m(\mathbf{x}, \boldsymbol{\theta}_o) - m(\mathbf{x}, \boldsymbol{\theta})]^2\right\} > 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\theta} \neq \boldsymbol{\theta}_o$.

These assumptions are sufficient to guarantee that $\boldsymbol{\theta}_o$ uniquely minimizes

$$\mathrm{E}\left\{[y - m(\mathbf{x}, \boldsymbol{\theta})]^2\right\}.$$

To see this, start from

$$\begin{aligned}[y - m(\mathbf{x}, \boldsymbol{\theta})]^2 &= [y - m(\mathbf{x}, \boldsymbol{\theta}_o) + m(\mathbf{x}, \boldsymbol{\theta}_o) - m(\mathbf{x}, \boldsymbol{\theta})]^2 \\ &= [y - m(\mathbf{x}, \boldsymbol{\theta}_o)]^2 + 2u[m(\mathbf{x}, \boldsymbol{\theta}_o) - m(\mathbf{x}, \boldsymbol{\theta})] + [m(\mathbf{x}, \boldsymbol{\theta}_o) - m(\mathbf{x}, \boldsymbol{\theta})]^2.\end{aligned}$$

Taking expectations and using NLS.1 which implies $\mathrm{E}\{u|\mathbf{x}\} = 0$, we find

$$\begin{aligned}\mathrm{E}\{u[m(\mathbf{x}, \boldsymbol{\theta}_o) - m(\mathbf{x}, \boldsymbol{\theta})]\} &= \mathrm{E}\left(\mathrm{E}\{u[m(\mathbf{x}, \boldsymbol{\theta}_o) - m(\mathbf{x}, \boldsymbol{\theta})]|\mathbf{x}\}\right) \\ &= \mathrm{E}\left(\mathrm{E}\{u|\mathbf{x}\}[m(\mathbf{x}, \boldsymbol{\theta}_o) - m(\mathbf{x}, \boldsymbol{\theta})]\right) = \mathrm{E}(0) = 0.\end{aligned}$$

Hence,

$$\mathrm{E}\left\{[y - m(\mathbf{x}, \boldsymbol{\theta})]^2\right\} = \mathrm{E}\left\{[y - m(\mathbf{x}, \boldsymbol{\theta}_o)]^2\right\} + \mathrm{E}\left\{[m(\mathbf{x}, \boldsymbol{\theta}_o) - m(\mathbf{x}, \boldsymbol{\theta})]^2\right\}.$$

Therefore the choice of $\boldsymbol{\theta}$ which minimizes

$$\mathrm{E}\left\{[y - m(\mathbf{x}, \boldsymbol{\theta})]^2\right\} = \mathrm{E}\left\{[y - m(\mathbf{x}, \boldsymbol{\theta}_o)]^2\right\} + \mathrm{E}\left\{[m(\mathbf{x}, \boldsymbol{\theta}_o) - m(\mathbf{x}, \boldsymbol{\theta})]^2\right\}$$

is the one which yields

$$\mathrm{E}\left\{[m(\mathbf{x}, \boldsymbol{\theta}_o) - m(\mathbf{x}, \boldsymbol{\theta})]^2\right\} = 0.$$

By NLS.2, this is $\boldsymbol{\theta}_o$ because for any other choice of $\boldsymbol{\theta}$,

$$\mathrm{E}\left\{[m(\mathbf{x}, \boldsymbol{\theta}_o) - m(\mathbf{x}, \boldsymbol{\theta})]^2\right\} > 0.$$

Consistency

Results on consistency and asymptotic normality are presented below for the more general case of M-estimation

Interpretation

If the nonlinear model is based on the conditional mean function

$$\mathrm{E}(y|\mathbf{x}) = m(\mathbf{x}, \boldsymbol{\theta}_o)$$

interpretation is straightforward. In particular, the marginal effect of a change in x_k is

$$\frac{\partial\, \mathrm{E}(y|\mathbf{x})}{\partial x_k} = \frac{\partial m(\mathbf{x}, \boldsymbol{\theta}_o)}{\partial x_k}.$$

2. M-Estimation

Typically, we would minimize the squared distance between y and $m(\mathbf{x}, \boldsymbol{\theta})$. But sometimes different loss functions are used.

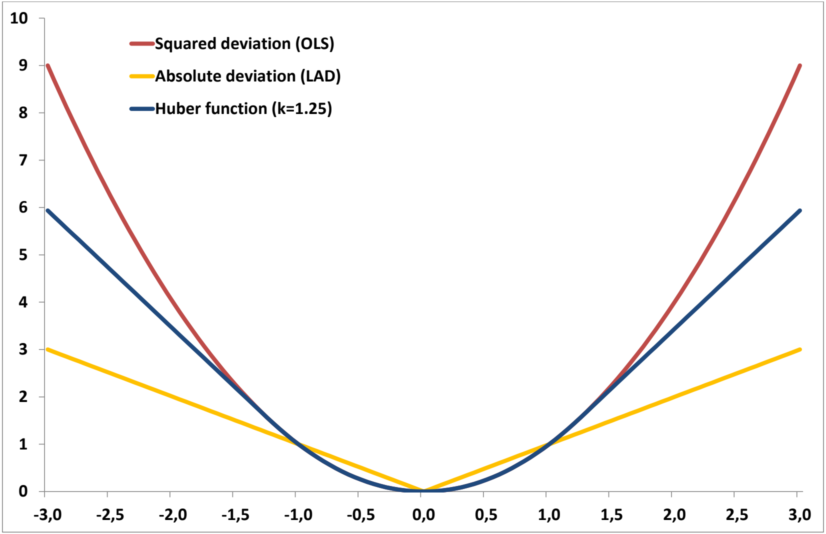
Here are some examples:

- The LAD estimator corresponds to

$$\min_{\boldsymbol{\beta} \in \boldsymbol{\Theta}} \mathrm{E}(|y - \mathbf{x}\boldsymbol{\beta}|)$$

- In its unscaled version, the Huber estimator corresponds to

$$\min_{\boldsymbol{\beta} \in \boldsymbol{\Theta}} \mathrm{E}[q(y - \mathbf{x}\boldsymbol{\beta})] = \begin{cases} 2k(y - \mathbf{x}\boldsymbol{\beta}) - k^2 & \text{if } y - \mathbf{x}\boldsymbol{\beta} > k \\ (y - \mathbf{x}\boldsymbol{\beta})^2 & \text{if } -k \leq y - \mathbf{x}\boldsymbol{\beta} \leq k \\ -2k(y - \mathbf{x}\boldsymbol{\beta}) - k^2 & \text{if } y - \mathbf{x}\boldsymbol{\beta} < -k \end{cases}$$


10
9
8
7
6
5
4
3
2
1
0
-3,0
-2,5
-2,0
-1,5
-1,0
-0,5
0,0
0,5
1,0
1,5
2,0
2,5
3,0
Squared deviation (OLS)
Absolute deviation (LAD)
Huber function (k=1.25)

The M-estimator

In general, we want to find the estimator $\hat{\boldsymbol{\theta}}$ that minimizes the sample function

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} N^{-1} \sum_{i=1}^{N} q(\mathbf{w}_i, \boldsymbol{\theta}),$$

where $\mathbf{w} \equiv (\mathbf{x}, y)$.

The aim is to estimate the population parameter $\boldsymbol{\theta}_o$ that minimizes the population function

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathrm{E}\left[q(\mathbf{w}, \boldsymbol{\theta})\right].$$

This is called M-estimation because we minimize a certain sample function. (Note that it implies maximization of $-q(\mathbf{w}_i, \boldsymbol{\theta})$ and is thus very general.)

The score and the first order condition

Recall that the M-estimator $\hat{\boldsymbol{\theta}}$ solves the minimization problem

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} N^{-1}\sum_{i=1}^{N} q(\mathbf{w}_i,\boldsymbol{\theta}).$$

Defining the score of the objective function $q(\mathbf{w},\boldsymbol{\theta})$ as

$$\mathbf{s}(\mathbf{w},\boldsymbol{\theta}) = \nabla'_{\boldsymbol{\theta}} q(\mathbf{w},\boldsymbol{\theta}) = \left[\frac{\partial q(\mathbf{w},\boldsymbol{\theta})}{\partial\theta_1},\ldots,\frac{\partial q(\mathbf{w},\boldsymbol{\theta})}{\partial\theta_p}\right]'$$

we can write the first order conditions for a minimum as

$$\sum_{i=1}^{N}\mathbf{s}(\mathbf{w}_i,\hat{\boldsymbol{\theta}}) = \mathbf{0}.$$

Since an explicit solution for $\hat{\boldsymbol{\theta}}$ is often impossible, we consider numerical solution methods below. For the following let us assume, we can find a $\hat{\boldsymbol{\theta}}$ such that the FOC is satisfied.

3. Identification and Consistency

Identification

For M-estimation, the identification requires that $\boldsymbol{\theta}_o$ be the unique solution to

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathrm{E}\left[q(\mathbf{w}, \boldsymbol{\theta})\right].$$

Hence, it requires

$$\mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta}_o)] < \mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta})]$$

for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\boldsymbol{\theta} \neq \boldsymbol{\theta}_o$.

Uniform convergence

While we will not go into the details here, uniform convergence in probability is needed to prove consistency. This is stronger than pointwise convergence we typically use (weak law of large numbers).

To this end, we have to require that (for all the conditions, see theorem 12.1 on p. 403)

- the parameter space $\boldsymbol{\Theta}$ is a compact subset of $\mathbb{R}^P$, where P is the dimension of $\boldsymbol{\theta}$,
- $q(\mathbf{w}, \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$, and
- $|q(\mathbf{w}, \boldsymbol{\theta})|$ is bounded across $\boldsymbol{\theta}$.

Then, uniform convergence holds:

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left| N^{-1} \sum_{i=1}^{N} q(\mathbf{w}_i, \boldsymbol{\theta}) - \mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta})] \right| \xrightarrow{\mathrm{p}} 0.$$

Consistency

Condition 1: $\boldsymbol{\theta}_o$ is identified.

Condition 2: Uniform convergence of $N^{-1}\sum_{i=1}^{N} q(\mathbf{w}_i, \boldsymbol{\theta})$ towards $\mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta})]$ holds.

Then the M-estimator $\hat{\boldsymbol{\theta}}$, which solves the problem

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} N^{-1} \sum_{i=1}^{N} q(\mathbf{w}_i, \boldsymbol{\theta})$$

is consistent,

$$\hat{\boldsymbol{\theta}} \xrightarrow{\mathrm{p}} \boldsymbol{\theta}_o.$$

4. Asymptotic Normality

Mean value theorem (*)

To find the asymptotic distribution, we need to use the mean value theorem.

Let us start with a scalar version cited from C. Feng et al. (2013) The Mean Value Theorem and Taylor's Expansion in Statistics, The American Statistician, 67:4, 245-248, found in the web here.

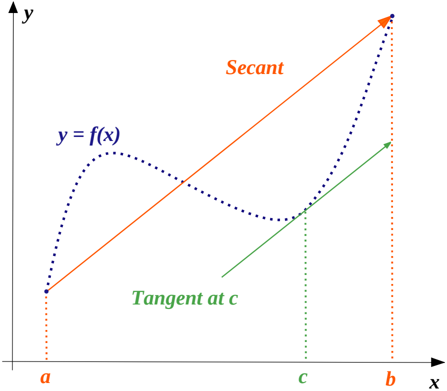
Suppose that O is an open interval and $f : O \to \mathbb{R}$ is a differentiable function. Then for any $[a, b] \subset O$, there exists a $c \in (a, b)$ such that

$$f(b) - f(a) = f'(c)\,(b - a) \qquad \text{and thus} \qquad f(b) = f(a) + f'(c)\,(b - a).$$

Intuition (stolen from Wikidepia): Write the mean value theorem as

$$\frac{f(b) - f(a)}{(b - a)} = f'(c).$$

Then it just says that there exists some c between a and b such that the secant joining the endpoints of the interval $[a, b]$ is parallel to the tangent at c.



The vector version (for a scalar function of a a vector of arguments) requires the definition of the gradient (= row vector of partial derivatives).

Suppose G is an open convex subset of $\mathbb{R}^p$ and $f : G \to \mathbb{R}$ is a differentiable function of p arguments with gradient vector

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \left[\frac{\partial f(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_p} \right].$$

Then for any vectors $\mathbf{a}, \mathbf{b} \in G$, there exists a $\mathbf{c}$ between $\mathbf{a}$ and $\mathbf{b}$ such that

$$f(\mathbf{b}) - f(\mathbf{a}) = \nabla_{\mathbf{x}} f(\mathbf{c})\ (\mathbf{b} - \mathbf{a})$$

and thus

$$f(\mathbf{b}) = f(\mathbf{a}) + \nabla_{\mathbf{x}} f(\mathbf{c})\ (\mathbf{b} - \mathbf{a})$$

A "matrix" version for a K-dimensional vector function of a vector of P arguments

$$\mathbf{f}(\mathbf{x}) = \left[f^{(1)}(\mathbf{x}), \ldots, f^{(K)}(\mathbf{x})\right]'$$

can be constructed from the vector version.

Suppose G is an open convex subset of $\mathbb{R}^p$ and $f^{(k)} : G \to \mathbb{R}$ is a differentiable function of p arguments with gradient vector

$$\nabla_{\mathbf{x}} f^{(k)}(\mathbf{x}) = \left[\frac{\partial f^{(k)}(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial f^{(k)}(\mathbf{x})}{\partial x_p}\right], \qquad k = 1, \ldots, K.$$

Then for any vectors $\mathbf{a}, \mathbf{b} \in G$, there exists a $\mathbf{c}^{(k)}$ between $\mathbf{a}$ and $\mathbf{b}$ such that

$$f^{(k)}(\mathbf{b}) = f^{(k)}(\mathbf{a}) + \nabla_{\mathbf{x}} f^{(k)}(\mathbf{c}^{(k)})\,(\mathbf{b} - \mathbf{a}), \qquad k = 1, \ldots, K.$$

Collecting all K equations yields

$$\begin{pmatrix} f^{(1)}(\mathbf{b}) \\ \vdots \\ f^{(K)}(\mathbf{b}) \end{pmatrix} = \begin{pmatrix} f^{(1)}(\mathbf{a}) \\ \vdots \\ f^{(K)}(\mathbf{a}) \end{pmatrix} + \begin{pmatrix} \nabla_{\mathbf{x}} f^{(1)}(\mathbf{c}^{(1)}) \\ \vdots \\ \nabla_{\mathbf{x}} f^{(K)}(\mathbf{c}^{(K)}) \end{pmatrix} (\mathbf{b} - \mathbf{a}).$$

The equation

$$\begin{pmatrix} f^{(1)}(\mathbf{b}) \\ \vdots \\ f^{(K)}(\mathbf{b}) \end{pmatrix} = \begin{pmatrix} f^{(1)}(\mathbf{a}) \\ \vdots \\ f^{(K)}(\mathbf{a}) \end{pmatrix} + \begin{pmatrix} \nabla_{\mathbf{x}} f^{(1)}(\mathbf{c}^{(1)}) \\ \vdots \\ \nabla_{\mathbf{x}} f^{(K)}(\mathbf{c}^{(K)}) \end{pmatrix} (\mathbf{b} - \mathbf{a})$$

can be written more compactly as

$$\mathbf{f}(\mathbf{b}) = \mathbf{f}(\mathbf{a}) + \mathbf{H}\ (\mathbf{b} - \mathbf{a}),$$

where $\mathbf{H}$ is the matrix of gradients, each evaluated at different points $\mathbf{c}^{(1)}, \ldots, \mathbf{c}^{(K)}$.

Application of the mean value theorem

Suppose that $q(\mathbf{w}_i, \boldsymbol{\theta})$ is twice continuously differentiable implying that each element in $\mathbf{s}(\mathbf{w}_i, \hat{\boldsymbol{\theta}})$ is once continuously differentiable.

Denote the kth element of the vector $\mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta})$ by $s^{(k)}(\mathbf{w}_i, \boldsymbol{\theta})$. Then the mean value theorem implies

$$s^{(k)}(\mathbf{w}_i, \hat{\boldsymbol{\theta}}) = s^{(k)}(\mathbf{w}_i, \boldsymbol{\theta}_o) + \nabla_{\boldsymbol{\theta}} s^{(k)}(\mathbf{w}_i, \bar{\boldsymbol{\theta}}^{(k)})\ (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o), k = 1, \dots, P,$$

where $\bar{\boldsymbol{\theta}}^{(k)}$ is between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_o$. Stacking all P equations yields

$$\mathbf{s}(\mathbf{w}_i, \hat{\boldsymbol{\theta}}) = \mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}_o) + \ddot{\mathbf{H}}_i\ (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o).$$

For the sample this implies

$$\sum_{i=1}^{N} \mathbf{s}(\mathbf{w}_i, \hat{\boldsymbol{\theta}}) = \sum_{i=1}^{N} \mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}_o) + \left(\sum_{i=1}^{N} \ddot{\mathbf{H}}_i \right)\ (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o).$$

The Hessian

The Hessian is the matrix of second derivatives of the objective function $q(\mathbf{w}, \boldsymbol{\theta})$

$$\mathbf{H}(\mathbf{w}, \boldsymbol{\theta}) \equiv \frac{\partial^2 q(\mathbf{w}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \equiv \nabla^2_{\boldsymbol{\theta}} q(\mathbf{w}, \boldsymbol{\theta}) \equiv \begin{pmatrix} \frac{\partial^2 q(\mathbf{w},\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_1} & \frac{\partial^2 q(\mathbf{w},\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 q(\mathbf{w},\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_P} \\ \frac{\partial^2 q(\mathbf{w},\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 q(\mathbf{w},\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_2} & \cdots & \frac{\partial^2 q(\mathbf{w},\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_P} \\ \vdots & \vdots & \ddots & \\ \frac{\partial^2 q(\mathbf{w},\boldsymbol{\theta})}{\partial \theta_P \partial \theta_1} & \frac{\partial^2 q(\mathbf{w},\boldsymbol{\theta})}{\partial \theta_P \partial \theta_2} & \cdots & \frac{\partial^2 q(\mathbf{w},\boldsymbol{\theta})}{\partial \theta_P \partial \theta_P} \end{pmatrix}$$

This is equivalent to the matrix of first derivatives of the score

$$\mathbf{H}(\mathbf{w}, \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \mathbf{s}(\mathbf{w}, \boldsymbol{\theta}) = \begin{pmatrix} \nabla_{\boldsymbol{\theta}} s^{(1)}(\mathbf{w}, \boldsymbol{\theta}) \\ \vdots \\ \nabla_{\boldsymbol{\theta}} s^{(P)}(\mathbf{w}, \boldsymbol{\theta}) \end{pmatrix}$$

With respect to the application of the mean value above, we have

$$\ddot{\mathbf{H}}_i = \begin{pmatrix} \nabla_{\boldsymbol{\theta}} s^{(1)}(\mathbf{w}_i, \bar{\boldsymbol{\theta}}^{(1)}) \\ \vdots \\ \nabla_{\boldsymbol{\theta}} s^{(P)}(\mathbf{w}_i, \bar{\boldsymbol{\theta}}^{(P)}) \end{pmatrix},$$

where each vector $\bar{\boldsymbol{\theta}}^{(k)}$ is between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_o$.

Convergence of $\ddot{\mathbf{H}}_i$

Recall that $\hat{\boldsymbol{\theta}} \xrightarrow{\text{p}} \boldsymbol{\theta}_o$. While we do not know the $\bar{\boldsymbol{\theta}}^{(k)}$'s which satisfy the mean value theorem, we know that they are between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_o$. Hence, they also converge towards $\boldsymbol{\theta}_o$,

$$\bar{\boldsymbol{\theta}}^{(k)} \xrightarrow{\text{p}} \boldsymbol{\theta}_o, \qquad \text{for all } k = 1, \dots, P.$$

This implies (by Lemma 12.1 on p. 405) for the sample average

$$N^{-1} \sum_{i=1}^{N} \ddot{\mathbf{H}}_i \xrightarrow{\text{p}} \mathrm{E}[\mathbf{H}(\mathbf{w}, \boldsymbol{\theta}_o)] \equiv \mathbf{A}_o.$$

Thus, for later use by Slutzky's theorem

$$\left(N^{-1} \sum_{i=1}^{N} \ddot{\mathbf{H}}_i \right)^{-1} \xrightarrow{\text{p}} \mathbf{A}_o^{-1},$$

as long as the expected value of the Hessian is invertible.

Asymptotic distribution of the scores

Since we have a random sample, we may apply a CLT to the scores if $\mathrm{E}[\mathbf{s}(\mathbf{w}, \boldsymbol{\theta}_o)] = \mathbf{0}$. To show this, recall that $\mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta})]$ has a minimum at $\boldsymbol{\theta}_o$. Hence,

$$\nabla_{\boldsymbol{\theta}} \mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta})]|_{\boldsymbol{\theta}=\boldsymbol{\theta}_o} = \mathbf{0}.$$

If the derivative and expectations operators can be interchanged, we obtain

$$\mathrm{E}[\nabla_{\boldsymbol{\theta}} q(\mathbf{w}, \boldsymbol{\theta}_o)] = \mathrm{E}[\mathbf{s}(\mathbf{w}, \boldsymbol{\theta}_o)] = \mathbf{0}.$$

Hence, the CLT yields

$$N^{-1/2} \sum_{i=1}^{N} \mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}_o) \xrightarrow{\mathrm{d}} \mathrm{Normal}(\mathbf{0}, \mathbf{B}_o),$$

where

$$\mathbf{B}_o = \mathrm{E}[\mathbf{s}(\mathbf{w}, \boldsymbol{\theta}_o)\mathbf{s}(\mathbf{w}, \boldsymbol{\theta}_o)'] = \mathrm{Var}[\mathbf{s}(\mathbf{w}, \boldsymbol{\theta}_o).]$$

Asymptotic distribution of the M-estimator

Recall the first order condition

$$\sum_{i=1}^{N} \mathbf{s}(\mathbf{w}_i, \hat{\boldsymbol{\theta}}) = \mathbf{0}$$

and the mean value expansion

$$\sum_{i=1}^{N} \mathbf{s}(\mathbf{w}_i, \hat{\boldsymbol{\theta}}) = \sum_{i=1}^{N} \mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}_o) + \left(\sum_{i=1}^{N} \ddot{\mathbf{H}}_i\right) (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o).$$

Combining these equations and multiplying by $N^{-1/2}$ yields

$$\mathbf{0} = N^{-1/2} \sum_{i=1}^{N} \mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}_o) + \left(N^{-1} \sum_{i=1}^{N} \ddot{\mathbf{H}}_i\right) N^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)$$

and thus

$$N^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) = \left(N^{-1} \sum_{i=1}^{N} \ddot{\mathbf{H}}_i\right)^{-1} \left[-N^{-1/2} \sum_{i=1}^{N} \mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}_o)\right].$$

Since

$$-N^{-1/2}\sum_{i=1}^{N}\mathbf{s}(\mathbf{w}_i,\boldsymbol{\theta}_o)\xrightarrow{\mathrm{d}}\mathrm{Normal}(\mathbf{0},\mathbf{B}_o)$$

and

$$\left(N^{-1}\sum_{i=1}^{N}\ddot{\mathbf{H}}_i\right)^{-1}\xrightarrow{\mathrm{p}}\mathbf{A}_o^{-1}$$

by Cramer's theorem (or the asymptotic equivalence lemma), we obtain

$$N^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_o)=\left(N^{-1}\sum_{i=1}^{N}\ddot{\mathbf{H}}_i\right)^{-1}\left[-N^{-1/2}\sum_{i=1}^{N}\mathbf{s}(\mathbf{w}_i,\boldsymbol{\theta}_o)\right]\xrightarrow{\mathrm{d}}\mathrm{Normal}(\mathbf{0},\mathbf{V}_o),$$

where

$\mathbf{V}_o=\mathbf{A}_o^{-1}\mathbf{B}_o\mathbf{A}_o^{-1}$. **Note: $\mathbf{V}_o$ is differently defined than in the textbook.**

5. Estimating the Asymptotic Variance

Estimation of $\mathbf{A}_o$

To estimate

$$\mathrm{E}[\mathbf{H}(\mathbf{w}, \boldsymbol{\theta}_o)] \equiv \mathbf{A}_o$$

a straightforward and consistent estimator uses the sample average evaluated at the consistent parameter estimator $\hat{\boldsymbol{\theta}}$,

$$N^{-1} \sum_{i=1}^{N} \mathbf{H}(\mathbf{w}_i, \hat{\boldsymbol{\theta}}) \equiv N^{-1} \sum_{i=1}^{N} \hat{\mathbf{H}}_i \xrightarrow{\mathrm{p}} \mathbf{A}_o.$$

While this estimator is consistent, it is sometimes difficult to calculate the second derivatives of the objective function.

When we know enough about the structure of the model, a popular alternative is to find the expectation of the Hessian conditional on $\mathbf{x}$ (recall that $\mathbf{w} \equiv (\mathbf{x}, y)$),

$$\mathbf{A}(\mathbf{x}, \boldsymbol{\theta}_o) \equiv \mathrm{E}[\mathbf{H}(\mathbf{w}, \boldsymbol{\theta}_o)|\mathbf{x}],$$

which is only a function of $\mathbf{x}$ (and not of y). By the law of iterated expectations,

$$\mathrm{E}[\mathbf{A}(\mathbf{x}, \boldsymbol{\theta}_o)] = \mathrm{E}\{\mathrm{E}[\mathbf{H}(\mathbf{w}, \boldsymbol{\theta}_o)|\mathbf{x}]\} = \mathrm{E}[\mathbf{H}(\mathbf{w}, \boldsymbol{\theta}_o)] = \mathbf{A}_o,$$

which can be consistently estimated as

$$N^{-1}\sum_{i=1}^{N}\mathbf{A}(\mathbf{x}_i, \hat{\boldsymbol{\theta}}) \equiv N^{-1}\sum_{i=1}^{N}\hat{\mathbf{A}}_i \xrightarrow{\mathrm{p}} \mathbf{A}_o.$$

In some leading cases, including NLS and certain ML problems, $\mathbf{A}(\mathbf{x}, \boldsymbol{\theta}_o)$ depends only of the first derivatives on the conditional mean function and is thus easier to compute compared to the first approach.

Estimation of $\mathbf{B}_o$

To estimate

$$\mathbf{B}_o = \mathrm{E}[\mathbf{s}(\mathbf{w}, \boldsymbol{\theta}_o)\mathbf{s}(\mathbf{w}, \boldsymbol{\theta}_o)']$$

a straightforward and consistent estimator uses the sample average evaluated at the consistent parameter estimator $\hat{\boldsymbol{\theta}}$,

$$N^{-1}\sum_{i=1}^{N}\mathbf{s}(\mathbf{w}_i, \hat{\boldsymbol{\theta}})\mathbf{s}(\mathbf{w}_i, \hat{\boldsymbol{\theta}})' \equiv N^{-1}\sum_{i=1}^{N}\hat{\mathbf{s}}_i\hat{\mathbf{s}}_i' \xrightarrow{p} \mathbf{B}_o.$$

Estimation of the asymptotic variance

Combining the estimators for $\mathbf{A}_o$ and $\mathbf{B}_o$ yields a consistent estimator of $\text{Avar}[\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)]$:

$$\widehat{\text{Avar}}[\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)] = \hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}.$$

To obtain asymptotic standard errors for the M-estimator $\hat{\boldsymbol{\theta}}$, use

$$\hat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}} \equiv \widehat{\text{Avar}}(\hat{\boldsymbol{\theta}}) = \hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}/N.$$

Depending on which estimator of $\mathbf{A}_o$ is chosen, $\hat{\mathbf{V}}$ can be expressed as

$$\left(\sum_{i=1}^{N} \hat{\mathbf{H}}_i\right)^{-1} \left(\sum_{i=1}^{N} \hat{\mathbf{s}}_i \hat{\mathbf{s}}_i'\right) \left(\sum_{i=1}^{N} \hat{\mathbf{H}}_i\right)^{-1}$$

or

$$\left(\sum_{i=1}^{N} \hat{\mathbf{A}}_i\right)^{-1} \left(\sum_{i=1}^{N} \hat{\mathbf{s}}_i \hat{\mathbf{s}}_i'\right) \left(\sum_{i=1}^{N} \hat{\mathbf{A}}_i\right)^{-1}.$$

The first one needs the fewest assumptions to be consistent and is thus a fully robust variance matrix estimator.

The second one is usually only valid when some feature of the conditional distribution of y given $\mathbf{x}$ is correctly specified. Hence, it needs more assumptions and is called a semirobust variance matrix estimator.

Generalized information matrix equality

In many contexts like NLS (see below) and maximum likelihood (see next lecture) the generalized information matrix equality holds under additional conditions:

$$\mathrm{E}[\mathbf{s}(\mathbf{w},\boldsymbol{\theta}_o)\mathbf{s}(\mathbf{w},\boldsymbol{\theta}_o)'] = \sigma_o^2\,\mathrm{E}[\mathbf{H}(\mathbf{w},\boldsymbol{\theta}_o)], \qquad \sigma_o^2 > 0,$$

and thus

$$\mathbf{B}_o = \sigma_o^2\mathbf{A}_o.$$

This implies the simplification

$$\mathbf{V}_{\hat{\boldsymbol{\theta}}} \equiv \mathrm{Avar}(\hat{\boldsymbol{\theta}}) = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}/N = \sigma_o^2\mathbf{A}^{-1}/N$$

which can be estimated as

$$\hat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}} = \hat{\sigma}_o^2\left(\sum_{i=1}^{N}\hat{\mathbf{H}}_i\right)^{-1} \qquad \text{or} \qquad \hat{\mathbf{V}}_{\hat{\boldsymbol{\theta}}} = \hat{\sigma}_o^2\left(\sum_{i=1}^{N}\hat{\mathbf{A}}_i\right)^{-1}.$$

6. Application to Nonlinear Least Squares

NLS setup

Let us come back to the nonlinear conditional mean function

$$\mathrm{E}(y|\mathbf{x}) = m(\mathbf{x}, \boldsymbol{\theta}_o),$$

which can be written as

$$y = m(\mathbf{x}, \boldsymbol{\theta}_o) + u, \qquad \text{with} \qquad \mathrm{E}(u|\mathbf{x}) = 0.$$

We aim to minimize

$$N^{-1} \sum_{i=1}^{N} [y_i - m(\mathbf{x}_i, \boldsymbol{\theta})]^2/2,$$

where division by 2 is just a convenient normalization. In terms of the M-estimator, the objective function specializes to

$$q(\mathbf{w}, \boldsymbol{\theta}) = [y - m(\mathbf{x}, \boldsymbol{\theta})]^2/2.$$

NLS score

The NLS score simplifies to

$$\mathbf{s}(\mathbf{w},\boldsymbol{\theta}) = \nabla'_{\boldsymbol{\theta}} q(\mathbf{w},\boldsymbol{\theta}) = \frac{\partial q(\mathbf{w},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -[y - m(\mathbf{x},\boldsymbol{\theta})]\frac{\partial m(\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -u\frac{\partial m(\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

Here it is simple to prove that the expectation of the score is zero. The conditional expectation is zero,

$$\mathrm{E}\left[-u\frac{\partial m(\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \middle| \mathbf{x} \right] = -\mathrm{E}\left[u \middle| \mathbf{x} \right] \frac{\partial m(\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0,$$

and by the law of iterated expectations also the unconditional expectation.

NLS Hessian

The NLS Hessian simplifies to

$$\begin{aligned}
\mathbf{H}(\mathbf{w},\boldsymbol{\theta}) &= \frac{\partial \mathbf{s}(\mathbf{w},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} \\
&= -\frac{\partial}{\partial \boldsymbol{\theta}'}\left[\frac{\partial m(\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\; u\right] \\
&= -u\frac{\partial^2 m(\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} - \frac{\partial m(\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial [y - m(\mathbf{x},\boldsymbol{\theta}_o)]}{\partial \boldsymbol{\theta}'} \\
&= -u\frac{\partial^2 m(\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} + \frac{\partial m(\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial m(\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}
\end{aligned}$$

Note that correct specification of the conditional mean function implies

$$\mathrm{E}\left\{u\frac{\partial^2 m(\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right\} = \mathrm{E}\left\{E[u|\mathbf{x}]\frac{\partial^2 m(\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right\} = 0,$$

and thus

$$\mathrm{E}\left[\mathbf{H}(\mathbf{w},\boldsymbol{\theta})\right] = \mathrm{E}\left[\frac{\partial m(\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial m(\mathbf{x},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right].$$

NLS variance matrix

Now let us find NLS-specific expressions for $\mathbf{A}_o$ and $\mathbf{B}_o$ of the variance of the M-estimator, $\mathbf{V}_o = \mathbf{A}_o^{-1}\mathbf{B}_o\mathbf{A}_o^{-1}$.

The matrix $\mathbf{A}_o$ turns out to be

$$\mathbf{A}_o \equiv \mathrm{E}[\mathbf{H}(\mathbf{w}, \boldsymbol{\theta}_o)] = \mathrm{E}\left[\frac{\partial m(\mathbf{x}, \boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta}} \frac{\partial m(\mathbf{x}, \boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta}'}\right].$$

For the matrix $\mathbf{B}_o$ we obtain

$$\mathbf{B}_o = \mathrm{E}[\mathbf{s}(\mathbf{w}, \boldsymbol{\theta}_o)\mathbf{s}(\mathbf{w}, \boldsymbol{\theta}_o)'] = \mathrm{E}\left[u^2 \frac{\partial m(\mathbf{x}, \boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta}} \frac{\partial m(\mathbf{x}, \boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta}'}\right].$$

Homoscedasticity

Under the

Assumption NLS.3: $\mathrm{Var}(y|\mathbf{x}) = \mathrm{Var}(u|\mathbf{x}) = \sigma_o^2$

the variance matrix simplifies considerably because then

$$\begin{aligned}\mathrm{E}\left[u^2 \frac{\partial m(\mathbf{x}, \boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta}} \frac{\partial m(\mathbf{x}, \boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta}'} \middle| \mathbf{x} \right] &= \mathrm{E}[u^2|\mathbf{x}] \; \frac{\partial m(\mathbf{x}, \boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta}} \frac{\partial m(\mathbf{x}, \boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta}'} \\ &= \sigma_o^2 \frac{\partial m(\mathbf{x}, \boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta}} \frac{\partial m(\mathbf{x}, \boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta}'}\end{aligned}$$

and thus, by the law of iterated expectations,

$$\mathbf{B}_o = \sigma_o^2 \, \mathrm{E}\left[\frac{\partial m(\mathbf{x}, \boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta}} \frac{\partial m(\mathbf{x}, \boldsymbol{\theta}_o)}{\partial \boldsymbol{\theta}'} \right] = \sigma_o^2 \mathbf{A}_o.$$

Hence, here the generalized information matrix equality holds and we obtain

$$\mathbf{V}_o = \sigma_o^2 \mathbf{A}_o^{-1}.$$

7. Inference

Confidence intervals for θ

Confidence intervals are obtained from the asymptotic distribution of $\hat{\boldsymbol{\theta}}$.

Nothing new here.

But keep in mind that in general there are three choices to estimate the variance matrix from which the standard errors are constructed:

- a fully robust estimator based on the second derivatives of the objective function,
- a semirobust estimator based on the conditional expectation of the second derivatives (in the NLS case we saw that this requires that the conditional mean function be correctly specified), and
- a nonrobust estimator based on the simplification due to the generalized information matrix equality (in the NLS case we saw that this requires homoscedasticity).

Confidence intervals for nonlinear functions of θ

Particularly in nonlinear models, it is sometimes interesting to find confidence intervals of nonlinear functions, $\mathbf{c}(\boldsymbol{\theta})$, of the parameters $\boldsymbol{\theta}$.

Using the **delta method** it is easy to find the asymptotic distribution of $\mathbf{c}(\hat{\boldsymbol{\theta}})$ (Lemma 3.9, p. 46).

Start from the P-dimensional asymptotic normal distribution

$$N^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \xrightarrow{\text{d}} \text{Normal}(0, \mathbf{V}).$$

Now let $\mathbf{c} : \boldsymbol{\Theta} \to \mathbb{R}^Q$ be a continuously differentiable function, where $Q \leq P$, and assume that $\boldsymbol{\theta}$ is in the interior of the parameter space. Define the $Q \times P$ Jacobian

$$\mathbf{C}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \mathbf{c}(\boldsymbol{\theta}).$$

Then

$$N^{1/2}[\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{c}(\boldsymbol{\theta}_o)] \xrightarrow{\text{d}} \text{Normal}(0, \mathbf{C}(\boldsymbol{\theta})\mathbf{V}\mathbf{C}(\boldsymbol{\theta})').$$

This implies that confidence intervals for nonlinear functions $\mathbf{c}(\boldsymbol{\theta})$ are based on

$$\mathbf{c}(\hat{\boldsymbol{\theta}}) \overset{a}{\sim} \text{Normal}[\mathbf{c}(\boldsymbol{\theta}_o), \mathbf{C}(\boldsymbol{\theta})\mathbf{V}\mathbf{C}(\boldsymbol{\theta})'/N].$$

In practice, $\mathbf{V}$ is estimated as usual and $\mathbf{C}(\boldsymbol{\theta})$ is replaced by $\mathbf{C}(\hat{\boldsymbol{\theta}})$.

Hence, all we need to do is to calculate the first derivatives of the nonlinear function and compute the variance estimate

$$\mathbf{C}(\hat{\boldsymbol{\theta}})\hat{\mathbf{V}}\mathbf{C}(\hat{\boldsymbol{\theta}})'/N.$$

The square roots of the main diagonal are the standard errors used to construct the confidence intervals for $\mathbf{c}(\boldsymbol{\theta})$.

Proof (*)

The proof uses the mean value expansion from above,

$$\mathbf{f}(\mathbf{b}) - \mathbf{f}(\mathbf{a}) = \mathbf{H}\ (\mathbf{b} - \mathbf{a}),$$

where we replace the general function $\mathbf{f}(\cdot)$ by $\mathbf{c}(\cdot)$ and use the points $\mathbf{b} = \hat{\boldsymbol{\theta}}$ and $\mathbf{a} = \boldsymbol{\theta}_o$. In addition the matrix of derivatives $\mathbf{H}$ becomes $\ddot{\mathbf{C}}$ (the Jacobian with rows evaluated at unknown mean values between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_o$). This yields

$$\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{c}(\boldsymbol{\theta}_o) = \ddot{\mathbf{C}}\ (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o).$$

Multiplication with $\sqrt{N}$ yields

$$N^{1/2}[\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{c}(\boldsymbol{\theta}_o)] = \ddot{\mathbf{C}}\ N^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o).$$

Since the unknown mean values are trapped between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_o$, $\ddot{\mathbf{C}}$ and $\mathbf{C}(\boldsymbol{\theta})$ are asymptotically equivalent. By the asymptotic equivalence lemma, we may thus consider

$$\mathbf{C}(\boldsymbol{\theta})\ N^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o).$$

Recall that

$$N^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \xrightarrow{\text{d}} \text{Normal}(0, \mathbf{V}).$$

Hence,

$$\mathbf{C}(\boldsymbol{\theta})\ N^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) \xrightarrow{\text{d}} \text{Normal}(0, \mathbf{C}(\boldsymbol{\theta})\mathbf{V}\mathbf{C}(\boldsymbol{\theta})').$$

This is asymptotically equivalent to

$$N^{1/2}[\mathbf{c}(\hat{\boldsymbol{\theta}}) - \mathbf{c}(\boldsymbol{\theta}_o)] \xrightarrow{\text{d}} \text{Normal}(0, \mathbf{C}(\boldsymbol{\theta})\mathbf{V}\mathbf{C}(\boldsymbol{\theta})').$$

Wald tests

Given the asymptotic normal distribution of the M-estimator, Wald tests of linear hypotheses of the kind

$$H_0 : \mathbf{R}\boldsymbol{\theta} = \mathbf{r} \quad \text{against} \quad H_1 : \mathbf{R}\theta \neq \mathbf{r}$$

can be performed as in the linear regression case.

Asymptotically, the Wald statistic

$$W_N \equiv \left[\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{r}\right]' \left[\mathbf{R}(\hat{\mathbf{V}}/N)\mathbf{R}'\right]^{-1} \left[\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{r}\right]$$

is χ^2_Q distributed, where Q is the number of restrictions.

Note that the Wald test can also used to test nonlinear hypotheses $\mathbf{c}(\boldsymbol{\theta}) = \mathbf{0}$. To this end, a Wald statistic is constructed based on the asymptotic normal distribution of $\mathbf{c}(\boldsymbol{\theta})$ which yields under H_0

$$W_N \equiv \mathbf{c}(\hat{\boldsymbol{\theta}})' \left[\mathbf{C}(\hat{\boldsymbol{\theta}})(\hat{\mathbf{V}}/N)\mathbf{C}(\hat{\boldsymbol{\theta}})'\right]^{-1} \mathbf{c}(\hat{\boldsymbol{\theta}}) \overset{a}{\sim} \chi^2_Q.$$

For details, see Wooldridge (p. 46-47).

Score tests and QLR tests

Two other important testing principles:

Score test: computes the score (which should be zero without restriction) under H_0 and rejects if it is “far enough” from zero.

Quasi-likelihood ratio (QLR) test: compares the value of the objective function under the null and the alternative and rejects if the difference is “large enough”.

We come back to these testing principle in the next lecture on maximum likelihood estimation.

8. Optimization Methods

Newton-Raphson method

Nonlinear problems typically require an iterative, numerical solution because an analytic solution is not possible.

A standard approach is the Newton-Raphson method.

Let us start with the scalar case. Consider the problem to find a root of the known function $f(x)$. This is done as follows:

- Find an initial guess x_0.
- Repeat the following steps until you arrive at a point where $f(x) \approx 0$.

  1. Compute the update $x_{i+1} = x_i - f(x_i)/f'(x_i)$.
  2. Compute $f(x_{i+1})$ and check whether it is near enough to zero. If yes, stop; if no, set i to $i+1$ and go to step 1.

What is the Newton-Raphson method doing?

It starts at an initial guess x_0.

Then it finds the tangency to f at x_0.

Then it finds the root of this tangency and calls it x_1.

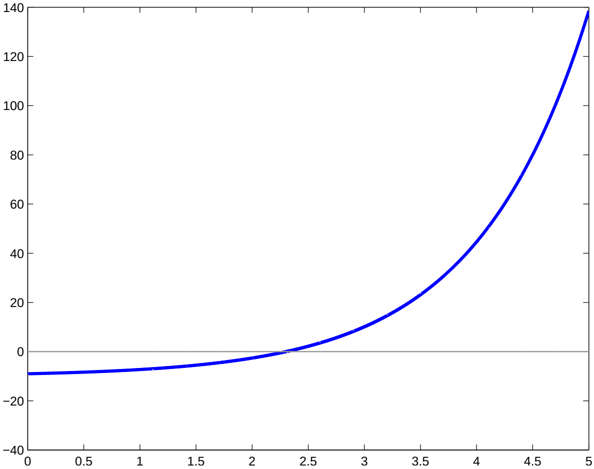
Then it finds the tangency to f at x_1.

Then it finds the root of this tangency and calls it x_2.

This is repeated until in a step x_n, $f(x_n)$ is almost zero.

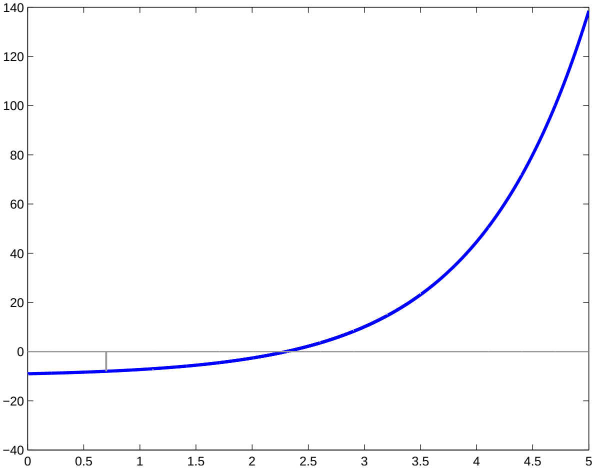
(A graphical description will be shown in class. A nice animation can be found here.)

Example: How to find the root of $f(x) = \exp(x) - 10$



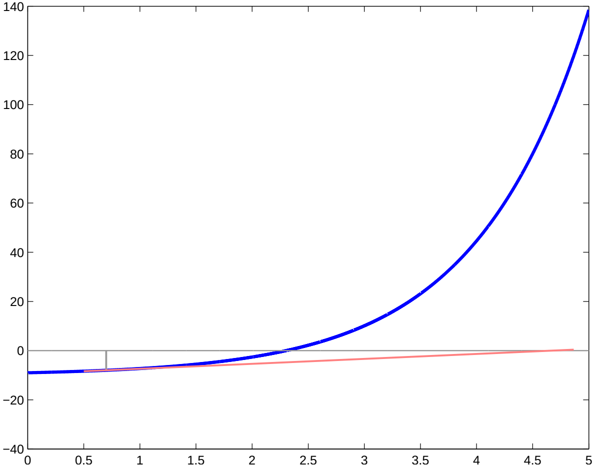
$f(x) = \exp(x) - 10$, root: $\log(10) = 2.3026$

Example: How to find the root of $f(x) = \exp(x) - 10$



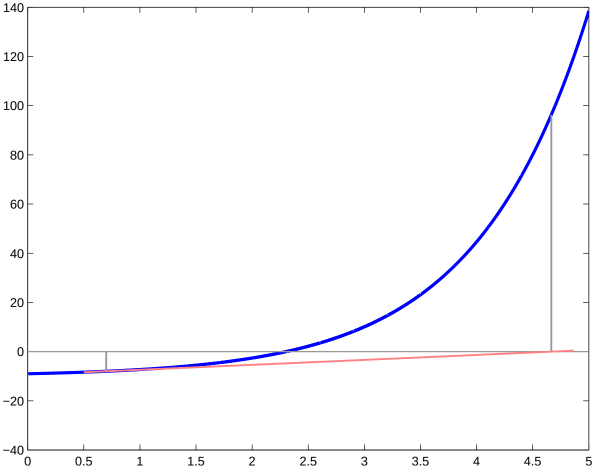
$x_0 = 0.70$

Example: How to find the root of $f(x) = \exp(x) - 10$



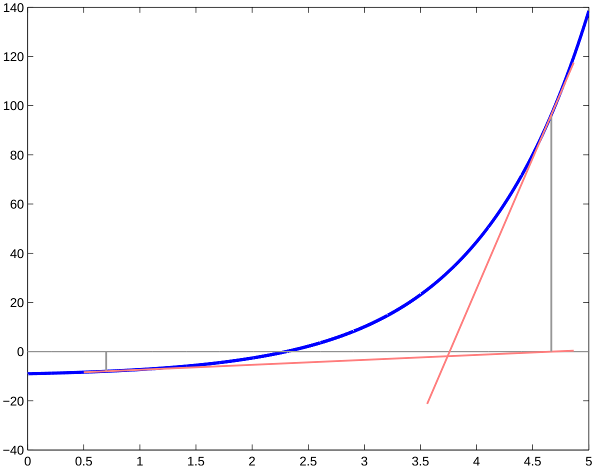
$x_0 = 0.70, \; x_1 = 4.67$

Example: How to find the root of $f(x) = \exp(x) - 10$



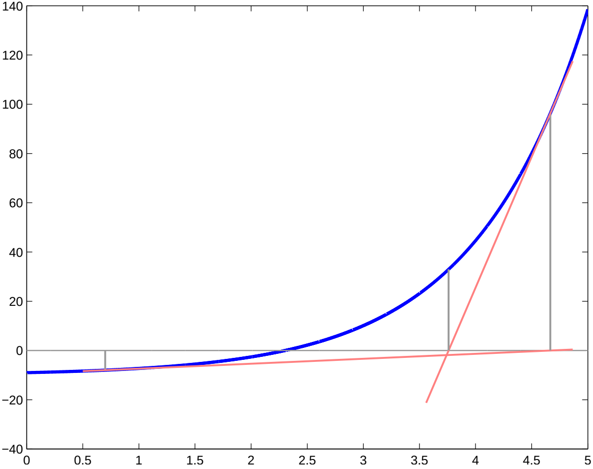
$x_1 = 4.67$

Example: How to find the root of $f(x) = \exp(x) - 10$



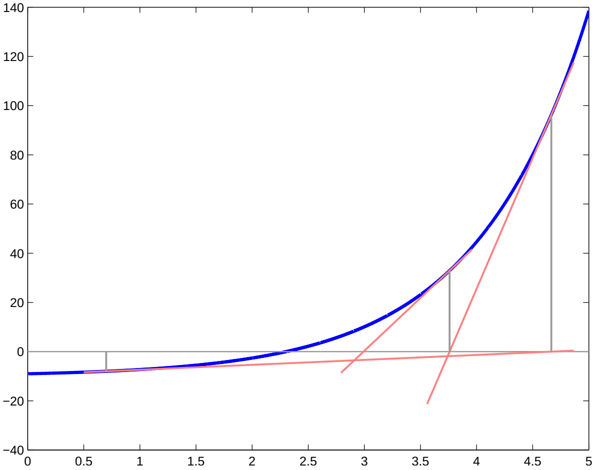
$x_1 = 4.67, \; x_2 = 3.76$

Example: How to find the root of $f(x) = \exp(x) - 10$



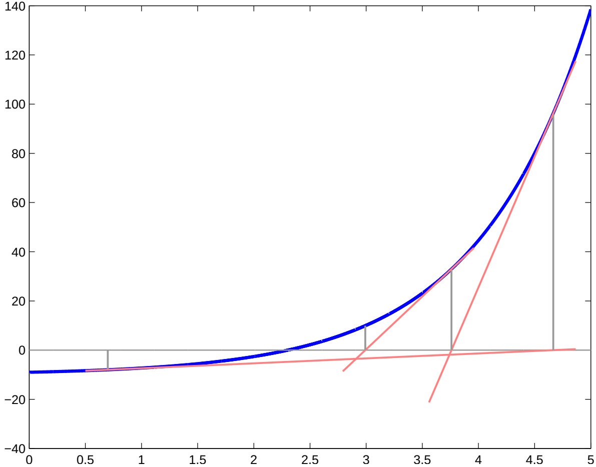
$x_2 = 3.76$

Example: How to find the root of $f(x) = \exp(x) - 10$



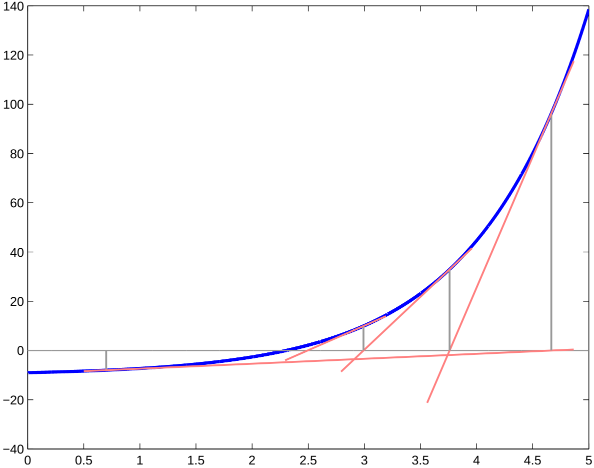
$x_2 = 3.76,\ x_3 = 2.99$

Example: How to find the root of $f(x) = \exp(x) - 10$



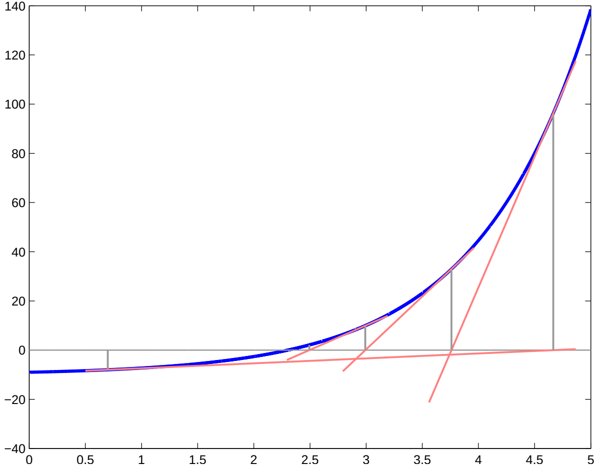
$x_3 = 2.99$

Example: How to find the root of $f(x) = \exp(x) - 10$



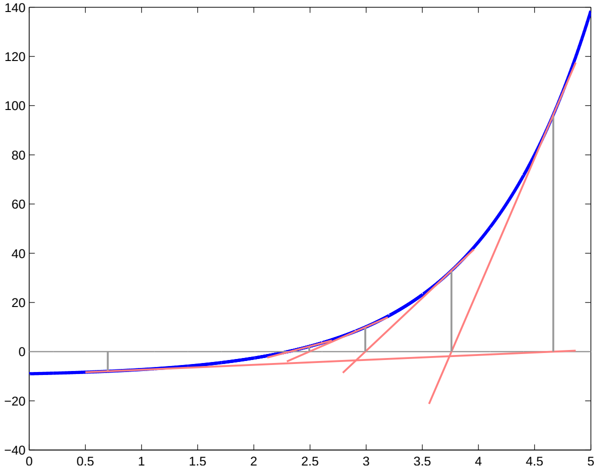
$x_3 = 2.99,\ x_4 = 2.49$

Example: How to find the root of $f(x) = \exp(x) - 10$



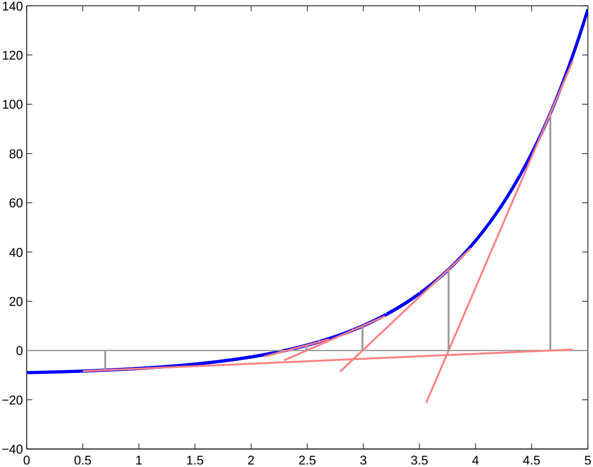
$x_4 = 2.49$

Example: How to find the root of $f(x) = \exp(x) - 10$



$x_4 = 2.49,\ x_5 = 2.32$

Example: How to find the root of $f(x) = \exp(x) - 10$



$x_5 = 2.31$, root: $\log(10) = 2.3026$

Using Newton-Raphson to find the M-estimator

Recall that the M-estimator is the coefficient vector $\hat{\boldsymbol{\theta}}$ that is the root of the FOC

$$\sum_{i=1}^{N} \mathbf{s}(\mathbf{w}_i, \hat{\boldsymbol{\theta}}) = \mathbf{0}.$$

The multivariate generalization of the Newton-Raphson method implemented for our problem thus is

$$\begin{array}{ccccccc} x_{g+1} & = & x_g & - & [f'(x_g)]^{-1} & & f(x_g) \\ \downarrow & & \downarrow & & \downarrow & & \downarrow \\ \boldsymbol{\theta}^{\{g+1\}} & = & \boldsymbol{\theta}^{\{g\}} & - & \left[\sum_{i=1}^{N} \mathbf{H}(\mathbf{w}_i, \boldsymbol{\theta}^{\{g\}})\right]^{-1} & & \left[\sum_{i=1}^{N} \mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}^{\{g\}})\right] \end{array}$$

Discussion

The iteration

$$\boldsymbol{\theta}^{\{g+1\}} = \boldsymbol{\theta}^{\{g\}} - \left[\sum_{i=1}^{N} \mathbf{H}(\mathbf{w}_i, \boldsymbol{\theta}^{\{g\}})\right]^{-1} \left[\sum_{i=1}^{N} \mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}^{\{g\}})\right]$$

will typically find the roots of the scores if the objective function has a unique minimum. But there are some caveats:

- Results may depend on the starting values. If the iteration does not converge for a certain set of starting values, try different ones.
- We need second derivatives which are sometimes difficult to find. While many software packages can compute numerical approximations, this is less accurate and can become computationally very demanding.
- The Hessian may not be invertible at certain points of the parameter space.

How to speed up convergence

Sometimes the iteration

$$\boldsymbol{\theta}^{\{g+1\}} = \boldsymbol{\theta}^{\{g\}} - \left[\sum_{i=1}^{N} \mathbf{H}(\mathbf{w}_i, \boldsymbol{\theta}^{\{g\}})\right]^{-1} \left[\sum_{i=1}^{N} \mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}^{\{g\}})\right]$$

yields too large a change in the parameters. Instead, one may try

$$\boldsymbol{\theta}^{\{g+1\}} = \boldsymbol{\theta}^{\{g\}} - r \;\left[\sum_{i=1}^{N} \mathbf{H}(\mathbf{w}_i, \boldsymbol{\theta}^{\{g\}})\right]^{-1} \left[\sum_{i=1}^{N} \mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}^{\{g\}})\right],$$

where we take different r's (e.g., a grid of values between 0 and 1) and choose that step length that yields the smallest sample objective function $\sum_{i=1}^{N} q(\mathbf{w}_i, \boldsymbol{\theta}^{\{g+1\}})$.

Alternative procedures

The generalized Gauss-Newton method uses the conditional expectation $\mathbf{A}(\mathbf{x}_i, \boldsymbol{\theta}^{\{g\}})$ instead of the Hessian

$$\boldsymbol{\theta}^{\{g+1\}} = \boldsymbol{\theta}^{\{g\}} - r \left[\sum_{i=1}^{N} \mathbf{A}(\mathbf{x}_i, \boldsymbol{\theta}^{\{g\}})\right]^{-1} \left[\sum_{i=1}^{N} \mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}^{\{g\}})\right],$$

This works particularly well when $\mathbf{A}(\mathbf{x}, \boldsymbol{\theta}_o)$ can be obtained in closed form.

The Berndt, Hall, Hall, and Hausman (BHHH) algorithm uses the outer product of the score instead of the Hessian:

$$\boldsymbol{\theta}^{\{g+1\}} = \boldsymbol{\theta}^{\{g\}} - r \left[\sum_{i=1}^{N} \mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}^{\{g\}})\mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}^{\{g\}})'\right]^{-1} \left[\sum_{i=1}^{N} \mathbf{s}(\mathbf{w}_i, \boldsymbol{\theta}^{\{g\}})\right].$$

This circumvents calculating second derivatives and yields a quadratic form that is generally invertible. However, there is evidence that it does not work well in all cases.

Possible strategy: Whenever one algorithm is too difficult to implement or too time-consuming to converge, try another one.