M3S3/S4 STATISTICAL THEORY II

IMPROVING INEFFICIENT ESTIMATORS: THE ONE-STEP ESTIMATOR

Objective : to produce a consistent estimator with asymptotic variance equal to the inverse Fisher information

$$I(\boldsymbol{\theta_0})^{-1}$$

as this is the best possible variance we can achieve for consistent estimators.

Suppose that $\widehat{\boldsymbol{\theta}}^{(0)}$ is a (consistent) estimator of $\boldsymbol{\theta}$ with asymptotic variance $\Sigma^{(0)}$ where

$$\Sigma^{(0)} - I(\boldsymbol{\theta_0})^{-1} \text{ is positive definite} \qquad \therefore \qquad \Sigma^{(0)} \geq I(\boldsymbol{\theta_0})^{-1}$$

or

$$x^{\mathsf{T}}(\Sigma^{(0)} - I(\boldsymbol{\theta_0})^{-1})x > 0 \qquad \forall\, x \in R^d$$

so that $\widehat{\boldsymbol{\theta}}^{(0)}$ is inefficient. This estimator can be improved by two iterative procedures that each define a sequence of estimators:

- **Newton's Method** For $k = 0, 1, \ldots$, let

$$\widehat{\boldsymbol{\theta}}^{(k+1)} = \widehat{\boldsymbol{\theta}}^{(k)} - \left(\ddot{\boldsymbol{l}}_n(\widehat{\boldsymbol{\theta}}^{(k)})\right)^{-1} \dot{\boldsymbol{l}}_n(\widehat{\boldsymbol{\theta}}^{(k)})$$

- **Method of Scoring** For $k = 0, 1, \ldots$, let

$$\widehat{\boldsymbol{\theta}}^{(k+1)} = \widehat{\boldsymbol{\theta}}^{(k)} + \left(I(\widehat{\boldsymbol{\theta}}^{(k)})\right)^{-1} \frac{1}{n}\, \dot{\boldsymbol{l}}_n(\widehat{\boldsymbol{\theta}}^{(k)})$$

Recall that

$$-\frac{1}{n}\ddot{\boldsymbol{l}}_n(\widehat{\boldsymbol{\theta}}^{(k)}) \xrightarrow{p} I(\boldsymbol{\theta})$$

which explains the connection between the two approaches. The sequence of estimators will have increasingly better properties.

The following theorem proves that only **one** iterative step is required to match the asymptotic efficiency of solutions to the likelihood equations, which, from a previous Theorem (2.1) have been shown to have asymptotic variance equal to the Cramér-Rao information bound. The method of proof is as follows

1. Find a consistent but possibly inefficient estimator $\widetilde{\boldsymbol{\theta}}_n$
2. Form the one-step Newton or Scoring Estimator using the formulae

$$\begin{aligned} \widehat{\boldsymbol{\theta}}^{(1)} &= \widetilde{\boldsymbol{\theta}}_n - \left(\ddot{\boldsymbol{l}}_n(\widetilde{\boldsymbol{\theta}}_n)\right)^{-1} \dot{\boldsymbol{l}}_n(\widetilde{\boldsymbol{\theta}}_n) \\ \widehat{\boldsymbol{\theta}}^{\star} &= \widetilde{\boldsymbol{\theta}}_n + \left(I(\widetilde{\boldsymbol{\theta}}_n)\right)^{-1} \frac{1}{n}\, \dot{\boldsymbol{l}}_n(\widetilde{\boldsymbol{\theta}}_n) \end{aligned}$$

3. Show that these estimators have the same asymptotic properties as solutions to the likelihood equations. That is, under regularity conditions, if $\widehat{\boldsymbol{\theta}}_n$ satisfies

$$\dot{\boldsymbol{l}}_n(\widehat{\boldsymbol{\theta}}_n) = \mathbf{0} \qquad \text{(LE)}$$

then, by Theorem 2.1

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta_0}) \xrightarrow{\mathfrak{L}} N(0, I(\boldsymbol{\theta_0})^{-1})$$

and Theorem 2.2 shows that $\widehat{\boldsymbol{\theta}}^{(1)}$ and $\widehat{\boldsymbol{\theta}}^{\star}$ also have these properties.

Theorem 2.2 ***The Efficiency of One-Step Estimators***

Let $\widetilde{\boldsymbol{\theta}}_n$, $n = 1, 2, \ldots$, be a (strongly) consistent sequence of estimators of $\boldsymbol{\theta} \in \Theta$ with true value equal to $\boldsymbol{\theta_0}$. Suppose that

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta_0}) \xrightarrow{\mathfrak{L}} N(0, \Sigma(\boldsymbol{\theta_0}))$$

with $\Sigma(\boldsymbol{\theta_0})$ finite. Then, under the conditions of Wald's Theorem on the strong consistency of the MLE, and conditions A0-A4 of theorem 2.1 that ensure the asymptotic behaviour of the MLE (or, at least, consistent solutions to the likelihood equations), $\widehat{\boldsymbol{\theta}}_n$, the two estimators

$$\widehat{\boldsymbol{\theta}}^{(1)} = \widetilde{\boldsymbol{\theta}}_n - \left(\ddot{\boldsymbol{l}}_n(\widetilde{\boldsymbol{\theta}}_n)\right)^{-1} \dot{\boldsymbol{l}}_n(\widetilde{\boldsymbol{\theta}}_n) \tag{N}$$

and

$$\widehat{\boldsymbol{\theta}}^{\star} = \widetilde{\boldsymbol{\theta}}_n + \left(I(\widetilde{\boldsymbol{\theta}}_n)\right)^{-1} \frac{1}{n}\, \dot{\boldsymbol{l}}_n(\widetilde{\boldsymbol{\theta}}_n) \tag{S}$$

are ***asymptotically equivalent*** *to the MLE, so that*

$$\widehat{\boldsymbol{\theta}}^{(1)} - \widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \mathbf{0}$$

and

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}^{(1)} - \boldsymbol{\theta_0}) \xrightarrow{\mathfrak{L}} N(0, I(\boldsymbol{\theta_0})^{-1})$$

with identical results for $\widehat{\boldsymbol{\theta}}^{\star}$.

Proof. Suppose that $\widehat{\boldsymbol{\theta}}_n$ is a (strongly) consistent of estimators that satisfy

$$\dot{\boldsymbol{l}}_n(\widehat{\boldsymbol{\theta}}_n) = \mathbf{0} \tag{LE}$$

then, by Theorem 2.1

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta_0}) \xrightarrow{\mathfrak{L}} N(0, I(\boldsymbol{\theta_0})^{-1})$$

Note: At no stage in the estimation will we actually have to find the numerical value of $\widehat{\boldsymbol{\theta}}_n$; we merely rely on its existence and asymptotic properties, both of which are guaranteed by the conditions of Theorem 2.1.

Now, using a **Mean-Value Theorem** first-order expansion of $\dot{\boldsymbol{l}}_n$ about $\widehat{\boldsymbol{\theta}}_n$ yields the following equation:

$$\dot{\boldsymbol{l}}_n(\widetilde{\boldsymbol{\theta}}_n) = \dot{\boldsymbol{l}}_n(\widehat{\boldsymbol{\theta}}_n) + \left\{\int_0^1 \ddot{\boldsymbol{l}}_n(\widehat{\boldsymbol{\theta}}_n + v(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n))\, dv\right\}(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n) = \left\{\int_0^1 \ddot{\boldsymbol{l}}_n(\widehat{\boldsymbol{\theta}}_n + v(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n))\, dv\right\}(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n). \tag{1}$$

as, by assumption, $\dot{\boldsymbol{l}}_n(\widehat{\boldsymbol{\theta}}_n) = \mathbf{0}$. In this equation, the left hand side is a $d \times 1$ vector, the term in the integrand is a $d \times d$ matrix.

Then, from the definition of $\widehat{\boldsymbol{\theta}}^{(1)}$, it follows that

$$(\widehat{\boldsymbol{\theta}}^{(1)} - \widehat{\boldsymbol{\theta}}_n) = (\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n) - \left(\ddot{\boldsymbol{l}}_n(\widetilde{\boldsymbol{\theta}}_n)\right)^{-1} \dot{\boldsymbol{l}}_n(\widetilde{\boldsymbol{\theta}}_n)$$

so that, by equation (1),

$$\begin{aligned}
\sqrt{n}(\widehat{\boldsymbol{\theta}}^{(1)} - \widehat{\boldsymbol{\theta}}_n) &= \sqrt{n}\left[(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n) - \left(\ddot{\boldsymbol{l}}_n(\widetilde{\boldsymbol{\theta}}_n)\right)^{-1} \dot{\boldsymbol{l}}_n(\widetilde{\boldsymbol{\theta}}_n)\right] \\
&= \sqrt{n}\left[(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n) - \left(\ddot{\boldsymbol{l}}_n(\widetilde{\boldsymbol{\theta}}_n)\right)^{-1} \left\{\int_0^1 \ddot{\boldsymbol{l}}_n(\widehat{\boldsymbol{\theta}}_n + v(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n))\, dv\right\}(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n)\right] \\
&= \left[\mathbf{1}_d - \left(\ddot{\boldsymbol{l}}_n(\widetilde{\boldsymbol{\theta}}_n)\right)^{-1} \left\{\int_0^1 \ddot{\boldsymbol{l}}_n(\widehat{\boldsymbol{\theta}}_n + v(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n))\, dv\right\}\right] \sqrt{n}(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n) \qquad (2)
\end{aligned}$$

Recall that both $\widehat{\boldsymbol{\theta}}_n$ and $\widetilde{\boldsymbol{\theta}}_n$ are consistent by assumption

$$\widehat{\boldsymbol{\theta}}_n \xrightarrow{a.s.} \boldsymbol{\theta}_{\mathbf{0}} \qquad \widetilde{\boldsymbol{\theta}}_n \xrightarrow{a.s.} \boldsymbol{\theta}_{\mathbf{0}}$$

and, this implies that

$$\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n \xrightarrow{a.s.} \mathbf{0}.$$

Therefore, (under the conditions of the theorem) by the Uniform Strong Law of Large Numbers (Chapter 1)

$$\frac{1}{n}\,\ddot{\boldsymbol{l}}_n(\widetilde{\boldsymbol{\theta}}_n) \xrightarrow{a.s.} -I(\boldsymbol{\theta}_{\mathbf{0}})$$

and, as $\widehat{\boldsymbol{\theta}}_n \xrightarrow{a.s.} \boldsymbol{\theta}_{\mathbf{0}}$ and $\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n \xrightarrow{a.s.} \mathbf{0}$, it follows that for any finite scalar v,

$$\widehat{\boldsymbol{\theta}}_n + v(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n) \xrightarrow{a.s.} \boldsymbol{\theta}_{\mathbf{0}}$$

so that

$$\frac{1}{n}\left\{\int_0^1 \ddot{\boldsymbol{l}}_n(\widehat{\boldsymbol{\theta}}_n + v(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n))\;dv\right\} \xrightarrow{a.s.} \left\{\int_0^1 E_{f_{X|\boldsymbol{\theta}_{\mathbf{0}}}}[\ddot{\boldsymbol{l}}_n(\boldsymbol{\theta}_{\mathbf{0}})]\;dv\right\} = \left\{\int_0^1 1\;dv\right\} E_{f_{X|\boldsymbol{\theta}_{\mathbf{0}}}}[\ddot{\boldsymbol{l}}_n(\boldsymbol{\theta}_{\mathbf{0}})] = -I(\boldsymbol{\theta}_{\mathbf{0}}).$$

Therefore, in equation (2)

$$\left(\ddot{\boldsymbol{l}}_n(\widetilde{\boldsymbol{\theta}}_n)\right)^{-1}\left\{\int_0^1 \ddot{\boldsymbol{l}}_n(\widehat{\boldsymbol{\theta}}_n + v(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n))\;dv\right\} \xrightarrow{a.s.} I(\boldsymbol{\theta}_{\mathbf{0}})^{-1}I(\boldsymbol{\theta}_{\mathbf{0}}) = \boldsymbol{I}_d$$

and so

$$\left[\mathbf{1}_d - \left(\ddot{\boldsymbol{l}}_n(\widetilde{\boldsymbol{\theta}}_n)\right)^{-1}\left\{\int_0^1 \ddot{\boldsymbol{l}}_n(\widehat{\boldsymbol{\theta}}_n + v(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n))\;dv\right\}\right] \xrightarrow{a.s.} \mathbf{1}_d - \mathbf{1}_d = \mathbf{0} \tag{3}$$

Also in equation (2),

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n) = \sqrt{n}(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{\mathbf{0}}) - \sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{\mathbf{0}})$$

and, by assumption

$$\left.\begin{array}{lcl} \sqrt{n}(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{\mathbf{0}}) & \xrightarrow{\mathfrak{L}} & N(0,\Sigma(\boldsymbol{\theta}_{\mathbf{0}})) \\ \sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{\mathbf{0}}) & \xrightarrow{\mathfrak{L}} & N(0,I(\boldsymbol{\theta}_{\mathbf{0}})^{-1}) \end{array}\right\} \quad \Longrightarrow \quad \sqrt{n}(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n) \xrightarrow{\mathfrak{L}} \boldsymbol{Z}_0 \sim N(0,\Sigma(\boldsymbol{\theta}_{\mathbf{0}}) + I(\boldsymbol{\theta}_{\mathbf{0}})^{-1})$$

Hence, from equations (2) and (3)

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}^{(1)} - \widehat{\boldsymbol{\theta}}_n) = \left[\mathbf{1}_d - \left(\ddot{\boldsymbol{l}}_n(\widetilde{\boldsymbol{\theta}}_n)\right)^{-1}\left\{\int_0^1 \ddot{\boldsymbol{l}}_n(\widehat{\boldsymbol{\theta}}_n + v(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n))\;dv\right\}\right]\sqrt{n}(\widetilde{\boldsymbol{\theta}}_n - \widehat{\boldsymbol{\theta}}_n) \xrightarrow{p} \mathbf{0} \times \boldsymbol{Z}_0 = \mathbf{0}.$$

This result uses the fact that convergence almost surely implies convergence in probability, and Slutsky's Theorem.

Hence,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}^{(1)} - \widehat{\boldsymbol{\theta}}_n) \xrightarrow{p} \mathbf{0}$$

and the two estimators are asymptotically equivalent. But the asymptotic distribution of $\widehat{\boldsymbol{\theta}}_n$ is known, and is a non-degenerate Normal distribution, and thus it follows that

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}^{(1)} - \boldsymbol{\theta}_{\mathbf{0}}) \xrightarrow{\mathfrak{L}} N(0,I(\boldsymbol{\theta}_{\mathbf{0}})^{-1})$$

an improvement on the original estimator, $\widetilde{\boldsymbol{\theta}}_n$, where

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{\mathbf{0}}) \xrightarrow{\mathfrak{L}} N(0,\Sigma(\boldsymbol{\theta}_{\mathbf{0}})).$$

The proof for $\widehat{\boldsymbol{\theta}}^{\star}$ follows in the same fashion.

$\blacksquare$