Lecture 8: Other asymptotically efficient estimators and pseudo MLE

One-Step MLE

Let $s_n(\gamma)$ be the score function.
Let $\widehat{\theta}_n^{(0)}$ be an estimator of θ that may not be asymptotically efficient.
The *one-step* MLE is the first iteration in computing an MLE (or RLE) using the Newton-Raphson method with $\widehat{\theta}_n^{(0)}$ as the initial value,

$$\widehat{\theta}_n^{(1)} = \widehat{\theta}_n^{(0)} - [\nabla s_n(\widehat{\theta}_n^{(0)})]^{-1} s_n(\widehat{\theta}_n^{(0)})$$

Without any further iteration, $\widehat{\theta}_n^{(1)}$ is asymptotically efficient under some conditions.

Theorem 4.19

Assume that the conditions in Theorem 4.16 hold and that $\widehat{\theta}_n^{(0)}$ is $\sqrt{n}$-consistent for θ (Definition 2.10).
(i) The one-step MLE $\widehat{\theta}_n^{(1)}$ is asymptotically efficient.
(ii) The one-step MLE obtained by replacing $\nabla s_n(\gamma)$ with its expected value, $-I_n(\gamma)$ (the Fisher-scoring method), is asymptotically efficient.

Proof

Since $\widehat{\theta}_n^{(0)}$ is $\sqrt{n}$-consistent, we can focus on the event $\widehat{\theta}_n^{(0)} \in A_\varepsilon = \{\gamma : \|\gamma - \theta\| \le \varepsilon\}$ for a sufficiently small ε such that $A_\varepsilon \subset \Theta$. From the mean-value theorem,

$$s_n(\widehat{\theta}_n^{(0)}) = s_n(\theta) + \left[\int_0^1 \nabla s_n\big(\theta + t(\widehat{\theta}_n^{(0)} - \theta)\big)\, dt\right] (\widehat{\theta}_n^{(0)} - \theta).$$

Substituting this into the formular for $\widehat{\theta}_n^{(1)}$, we obtain that

$$\widehat{\theta}_n^{(1)} - \theta = -[\nabla s_n(\widehat{\theta}_n^{(0)})]^{-1} s_n(\theta) + [I_k - G_n(\widehat{\theta}_n^{(0)})](\widehat{\theta}_n^{(0)} - \theta),$$

where

$$G_n(\widehat{\theta}_n^{(0)}) = [\nabla s_n(\widehat{\theta}_n^{(0)})]^{-1} \int_0^1 \nabla s_n\big(\theta + t(\widehat{\theta}_n^{(0)} - \theta)\big)\, dt.$$

From the proof of Theorem 4.17,

$$\|[I_n(\theta)]^{1/2}[\nabla s_n(\widehat{\theta}_n^{(0)})]^{-1}[I_n(\theta)]^{1/2} + I_k\| \to_p 0.$$

Proof (continued)

Using an argument similar to those in the proof of Theorem 4.17, we can show that

$$\|G_n(\widehat{\theta}_n^{(0)}) - I_k\| \to_p 0.$$

These results and the fact that $\sqrt{n}(\widehat{\theta}_n^{(0)} - \theta) = O_p(1)$ imply

$$\sqrt{n}(\widehat{\theta}_n^{(1)} - \theta) = \sqrt{n}[I_n(\theta)]^{-1} s_n(\theta) + o_p(1).$$

This proves (i).
The proof for (ii) is similar.

Example 4.40

Let $X_1, ..., X_n$ be i.i.d. from the Weibull distribution $W(\theta, 1)$, where $\theta > 0$ is unknown.
Note that

$$s_n(\theta) = \frac{n}{\theta} + \sum_{i=1}^{n} \log X_i - \sum_{i=1}^{n} X_i^{\theta} \log X_i$$

Example 40 (continued)

Then

$$\nabla s_n(\theta) = -\frac{n}{\theta^2} - \sum_{i=1}^{n} X_i^{\theta}(\log X_i)^2.$$

Hence, the one-step MLE of θ is

$$\widehat{\theta}_n^{(1)} = \widehat{\theta}_n^{(0)} \left[1 + \frac{n + \widehat{\theta}_n^{(0)}(\sum_{i=1}^{n} \log X_i - \sum_{i=1}^{n} X_i^{\widehat{\theta}_n^{(0)}} \log X_i)}{n + (\widehat{\theta}_n^{(0)})^2 \sum_{i=1}^{n} X_i^{\widehat{\theta}_n^{(0)}} (\log X_i)^2} \right].$$

Usually one can use a moment estimator (§3.5.2) as the initial estimator $\widehat{\theta}_n^{(0)}$.
In this example, a moment estimator of θ is the solution of $\bar{X} = \Gamma(\theta^{-1} + 1)$.

Results similar to that in Theorem 4.19 can be obtained in the GLM.

One-way random effects model

Consider the one-way random effects model

$$X_{ij} = \mu + A_i + e_{ij}, \quad j = 1, ..., n, i = 1, ..., m,$$

where $\mu \in \mathscr{R}$, A_i's are iid as $N(0, \sigma_a^2)$, e_{ij}'s are iid as $N(0, \sigma^2)$, σ_a^2 and σ^2 are unknown, and A_i's and e_{ij}'s are independent.

It can be shown that the MLE of μ is $\bar{X}_{\cdot\cdot} = (nm)^{-1} \sum_{i=1}^m \sum_{j=1}^n X_{ij}$, which is normally distributed with mean μ and variance $m^{-1}(\sigma_a^2 + n^{-1}\sigma^2)$.

The MLE of σ^2 is

$$\widehat{\sigma}^2 = S_E/[m(n-1)], \quad S_E = \sum_{i=1}^m \sum_{j=1}^n (X_{ij} - \bar{X}_{i\cdot})^2, \ \ \bar{X}_{i\cdot} = \frac{1}{n} \sum_{j=1}^n X_{ij},$$

and the MLE of σ_a^2 is $\widehat{\sigma}_a^2 I_{[0,\infty)}(\widehat{\sigma}_a^2)$, where

$$\widehat{\sigma}_a^2 = S_A/[n(m-1)] - S_E/[nm(n-1)], \quad S_A = n \sum_{i=1}^m (\bar{X}_{i\cdot} - \bar{X}_{\cdot\cdot})^2.$$

$\widehat{\sigma}_a^2$ is an ANOVA type estimator, which may be negative.

One-way random effects model

We now show that as long as $nm \to \infty$, $P(\hat{\sigma}_a^2 \leq 0) \to 0$.
Since S_E/σ^2 has the chi-square distribution $\chi^2_{m(n-1)}$,
$S_E/[m(n-1)] \to_p \sigma^2$ as $nm \to \infty$ (either $n \to \infty$ or $m \to \infty$).
Since $S_A/(\sigma^2 + n\sigma_a^2)$ has the chi-square distribution χ^2_{m-1},
$S_A/[n(m-1)] \sim (\sigma_a^2 + n^{-1}\sigma^2)W_{m-1}/(m-1)$, where W_{m-1} is a random variable having the chi-square distribution χ^2_{m-1}.

Case 1: $m \to \infty$ and $n \to \infty$.
$S_E/[nm(n-1)] \to_p 0$ and $(\sigma_a^2 + n^{-1}\sigma^2)W_{m-1}/(m-1) \to_p \sigma_a^2 > 0$.
Hence, $\hat{\sigma}_a^2 \to_p \sigma_a^2 > 0$, which implies $P(\hat{\sigma}_a^2 \leq 0) \to 0$.

Case 2: $m \to \infty$ but n is fixed.
$(\sigma_a^2 + n^{-1}\sigma^2)W_{m-1}/(m-1) \to_p (\sigma_a^2 + n^{-1}\sigma^2)$ and
$S_E/[nm(n-1)] \to_p n^{-1}\sigma^2$, which implies $\hat{\sigma}_a^2 \to_p \sigma_a^2 > 0$.

Case 3: $n \to \infty$ but m is fixed.
$(\sigma_a^2 + n^{-1}\sigma^2)W_{m-1}/(m-1) \to_d \sigma_a^2 W_{m-1}/(m-1)$ and
$S_E/[nm(n-1)] \to_p 0$. By Slutsky's theorem,
$\hat{\sigma}_a^2 \to_d \sigma_a^2 W_{m-1}/(m-1) \geq 0$.
Hence, $P(\hat{\sigma}_a^2 \leq 0) \to 0$.

One-way random effects model

Thus, the asymptotic distributions of MLE's are the same as those of $\bar{X}_{\cdot\cdot}$, $\widehat{\sigma}_a^2$, and $\widehat{\sigma}^2$.

Since $S_E/\sigma^2 \sim \chi^2_{m(n-1)}$, as $nm \to \infty$ (either $n \to \infty$ or $m \to \infty$),

$$\sqrt{nm}\left(\widehat{\sigma}^2 - \sigma^2\right) \to_d N(0, 2\sigma^4).$$

For $\widehat{\sigma}_a^2$, we need to consider the three cases previously discussed.

Case 1: $m \to \infty$ and $n \to \infty$. In this case,

$$\sqrt{m}\left[\frac{S_E}{nm(n-1)} - \frac{\sigma^2}{n}\right] \to_p 0 \quad \text{and} \quad \sqrt{m}\left(\frac{W_{m-1}}{m-1} - 1\right) \to_d N(0,2).$$

Since $S_A/[n(m-1)] \sim (\sigma_a^2 + n^{-1}\sigma^2)W_{m-1}/(m-1)$,

$$\sqrt{m}(\widehat{\sigma}_a^2 - \sigma_a^2) = \sqrt{m}\left[\frac{S_A}{n(m-1)} - \left(\sigma_a^2 + \frac{\sigma^2}{n}\right) + \frac{\sigma^2}{n} - \frac{S_E}{nm(n-1)}\right]$$

has the same asymptotic distribution as that of

$$\sqrt{m}\left(\sigma_a^2 + \frac{\sigma^2}{n}\right)\left(\frac{W_{m-1}}{m-1} - 1\right).$$

One-way random effects model

Thus,

$$\sqrt{m}(\widehat{\sigma}_a^2 - \sigma_a^2) \to_d N(0, 2\sigma_a^4).$$

Case 2: $m \to \infty$ but n is fixed.
In this case,

$$\sqrt{m}\left[\frac{S_E}{nm(n-1)} - \frac{\sigma^2}{n}\right] \to_d N(0, 2\sigma^4 n^{-3}).$$

From the argument in the previous case and the fact that S_A and S_E are independent, we obtain that

$$\sqrt{m}(\widehat{\sigma}_a^2 - \sigma_a^2) \to_d N\left(0, 2(\sigma_a^2 + n^{-1}\sigma^2)^2 + 2\sigma^4 n^{-3}\right).$$

Case 3: $n \to \infty$ but m is fixed.
In this case, $S_E/[nm(n-1)] - \sigma^2/n \to_p 0$ and

$$\left(\sigma_a^2 + \frac{\sigma^2}{n}\right)\left(\frac{W_{m-1}}{m-1} - 1\right) \to_d \sigma_a^2\left(\frac{W_{m-1}}{m-1} - 1\right).$$

Therefore,

$$\widehat{\sigma}_a^2 - \sigma_a^2 \to_d \sigma_a^2\left(\frac{W_{m-1}}{m-1} - 1\right).$$

Pseudo MLE

Let $X_1, \ldots, X_n$ be a random sample from a pdf in a family indexed by two parameters θ and π with likelihood $\ell(\theta, \pi)$.

The method of pseudo MLE may be viewed as follows.

- Based on the sample, an estimate $\widehat{\pi}$ of π is obtained using some technique other than MLE.
- The pseudo MLE of θ is then obtained by maximizing the likelihood $\ell(\theta, \widehat{\pi})$.

Discussion

- π is viewed as a nuisance parameter.
- Pseudo MLE consists of replacing π by an estimate and solving a reduced system of likelihood equations, which works when a higher dimensional MLE is intractable but a lower dimensional MLE is feasible.
- The consistency and asymptotic normality hold under fairly standard regularity conditions.
- The requirements on the model are slightly less stringent for pseudo MLE than for the MLE.

Lemma

Let $X_1, \ldots, X_n$ be i.i.d. from a distribution F_π, with $\pi \in \Pi$.
Let $\pi_0 \in \Pi$ be the true value of parameter, and let $\widehat{\pi}$ be a sample estimator such that $\widehat{\pi} \to_p \pi_0$.
Let $\psi(x, \pi)$ be a differentiable function of π for $\pi \in B$, an open neighborhood of π_0, and for almost all x in the sample space.
Suppose $E|\psi(X, \pi_0)| < \infty$.
If

$$\left| \frac{\partial}{\partial \pi} \psi(x, \pi) \right| \leq M(x)$$

for all $\pi \in B$, where $E[M(X)] < \infty$, then

$$\frac{1}{n} \sum_{i=1}^{n} \psi(X_i, \widehat{\pi}) \longrightarrow_p E\psi(X, \pi_0).$$

Proof. Consider the Taylor series expansion of $\frac{1}{n}\sum_{i=1}^{n} \psi(X_i, \widehat{\pi})$.

Notation

$s(\theta, \pi) = \partial \log \ell(\theta, \pi) / \partial \theta$
$\nabla_\varphi s(\theta, \pi) = \partial s(\theta, \pi) / \partial \varphi$, $\varphi = \theta$ or π.

Asymptotic existence and consistency of pseudo MLE

Assume the conditions in Theorem 4.16.
Assume also $\widehat{\pi}$ is a consistent estimator of π_0.
As $n \to \infty$, with probability tending to 1, there exists $\widehat{\theta}$ such that

$$s(\widehat{\theta}, \widehat{\pi}) = 0 \quad \text{and} \quad \widehat{\theta} \to_p \theta_0$$

where θ_0 is the true value of θ.

Proof.

By the lemma,

$$\frac{\log \ell(\theta, \widehat{\pi}) - \log \ell(\theta_0, \widehat{\pi})}{n} \to_p E \log \frac{f_{\theta,\pi_0}(X_1)}{f_{\theta_0,\pi_0}(X_1)} < \log E \frac{f_{\theta,\pi_0}(X_1)}{f_{\theta_0,\pi_0}(X_1)} < 0,$$

which means $\ell(\theta, \widehat{\pi})$ has a local maximum in $(\theta_0 - \varepsilon, \theta_0 + \varepsilon)$.
The rest of proof is the same as that for Theorem 4.17.

In many applications, the pseudo maximum likelihood equation has a unique solution and the pseudo MLE is indeed consistent.

Asymptotic Normality of pseudo MLE

Assume an additional assumption that

$$\widehat{\pi} - \pi_0 = \frac{1}{n}\sum_{i=1}^{n} \gamma(X_i) + o_p(n^{-1/2})$$

where γ is a function satisfying $E\gamma(X_1) = 0$ and $\mathrm{Var}(\gamma(X_1)) = \Sigma_\pi$ is finite.

We can then establish the asymptotic normality of the pseudo MLE.

We consider a consistent sequence $\widehat{\theta}$.

Since $s(\widehat{\theta}, \widehat{\pi}) = 0$,

$$-s(\theta_0, \widehat{\pi}) = \nabla_\theta s(\theta_0, \widehat{\pi})(\widehat{\theta} - \theta_0) + o_p(n^{1/2})$$

By the Lemma again, we can show that

$$n^{-1}\nabla_\theta s(\theta_0, \widehat{\pi}) \to_p -\Sigma_\theta$$

where Σ_θ is the Fisher information about θ when π_0 is known. Then

$$n^{-1/2}\Sigma_\theta^{-1} s(\theta_0, \widehat{\pi}) = \sqrt{n}(\widehat{\theta} - \theta_0) + o_p(1)$$

We need to derive the asymptotic normality of $s(\theta_0, \widehat{\pi})$.

$$n^{-1/2} s(\theta_0, \widehat{\pi}) = n^{-1/2}[s(\theta_0, \pi_0) + s(\theta_0, \widehat{\pi}) - s(\theta_0, \pi_0)]$$
$$= n^{-1/2} s(\theta_0, \pi_0) + n^{-1/2} \nabla_\pi s(\theta_0, \pi_0)(\widehat{\pi} - \pi_0) + o_p(1)$$

By the SLLN, $n^{-1}\nabla_\pi s(\theta_0, \pi_0) \to_{a.s.} E[n^{-1}\nabla_\pi s(\theta_0, \pi_0)] = -\Sigma_{\theta\pi}$.

Also,

$$s(\theta_0, \pi_0) = \sum_{i=1}^n \zeta(X_i), \qquad \zeta(X_i) = \frac{\nabla_{\theta_0} f_{\theta_0, \pi_0}(X_i)}{f_{\theta, \pi_0}(X_i)}$$

with $E\zeta(X_i) = 0$ and $\mathrm{Var}(\zeta(X_i)) = \Sigma_\theta$.

Define $\mathrm{Cov}(\zeta(X_i), \gamma(X_i)) = \Sigma_{cov}$.

Then

$$\mathrm{Var}(\zeta(X_i) + \Sigma_{\theta\pi}\gamma(X_i)) = \Sigma_\theta + \Sigma_{\theta\pi}\Sigma_\pi\Sigma_{\theta\pi}^\tau - 2\Sigma_{\theta\pi}\Sigma_{cov}$$

and

$$\sqrt{n}(\widehat{\theta} - \theta_0) = n^{-1/2}\Sigma_\theta^{-1} s(\theta_0, \widehat{\pi}) + o_p(1)$$
$$= n^{-1/2}\sum_{i=1}^n \Sigma_\theta^{-1}[\zeta(X_i) - \Sigma_{\theta\pi}\gamma(X_i)] + o_p(1)$$
$$\to_d N\left(0, \Sigma_\theta^{-1} + \Sigma_\theta^{-1}(\Sigma_{\theta\pi}\Sigma_\pi\Sigma_{\theta\pi}^\tau - 2\Sigma_{\theta\pi}\Sigma_{cov})\Sigma_\theta^{-1}\right)$$

Comparison

$\widehat{\theta}_{\pi_0}$: the MLE when π_0 is known.
$\widehat{\theta}_{ML}$: the MLE of θ.
$\widehat{\theta}_{PML}$: the pseudo MLE of θ.
From Theorem 4.17, $\widehat{\theta}_{\pi_0}$ is more efficient than $\widehat{\theta}_{ML}$ or $\widehat{\theta}_{PML}$; also, $\widehat{\theta}_{ML}$ is more efficient than $\widehat{\theta}_{PML}$.
In the special case where $\Sigma_{\theta\pi} = 0$, all three estimators are asymptotically equivalent.

Example: Signal plus noise model

Let $X_1, \ldots, X_n$ be i.i.d. from $Y + Z$, where $Y \sim$ Poisson(θ_0) is signal, $Z \sim Bi(N, \pi_0)$ is noise, and Y and Z are independent.
The moment estimators of π_0 and θ_0 are

$$\widehat{\pi} = \sqrt{(\bar{X} - S^2)/N} \quad \text{and} \quad \widehat{\theta} = \bar{X} - N\widehat{\pi},$$

where $\bar{X}$ and S^2 are the sample mean and variance, provided $\bar{X} \geq S^2$, which occurs with probability tending to 1 as $n \to \infty$.

Since the p.d.f. of X_i involves convolution, the MLE of (θ, π) is not so easy to compute.
The pseudo MLE can be computed with π replaced by $\widehat{\pi}$ in the p.d.f.

The asymptotic variances of the MLE, pseudo MLE and moment estimator (MME) of the signal parameter θ_0 are:

$$\begin{aligned}
\sigma^2_{\mathrm{MLE}} &= \frac{\phi_{22}}{\phi_{11}\phi_{22} - \phi_{12}^2}, \\
\sigma^2_{\mathrm{PMLE}} &= \frac{1}{\phi_{11}} + \frac{\phi_{12}^2}{\phi_{11}^2} t^2 (\Gamma_{22} - 2\Gamma_{23} + \Gamma_{33}), \\
\sigma^2_{\mathrm{MME}} &= (1 - Nt)^2 \Gamma_{22} + 2Nt(1 - Nt)\Gamma_{23} + (Nt)^2 \Gamma_{33},
\end{aligned}$$

where $t = 1/2N\pi_0$, $\Gamma_{22} = \theta_0 + N\pi_0(1 - \pi_0), \Gamma_{23} = \theta_0 + N\pi_0(1 - 2\pi_0)$, $\Gamma_{33} = \theta_0 + 2\big(\theta_0 + N\pi_0(1 - \pi_0)\big)^2 + N\pi_0(1 - \pi_0)\big(1 - 6\pi_0(1 - \pi_0)\big)^2$, $\phi_{11} = \Sigma_{\theta_0}$, $\phi_{12} = \Sigma_{\theta_0 \pi_0}$, and ϕ_{22} is the last diagonal element of the Fisher information matrix about (θ, p).
It is not easy to compare these expressions analytically.
For a specific range of parameters, we could find $\sigma^2_{\mathrm{PMLE}} < \sigma^2_{\mathrm{MME}}$.