Section 8.3. Efficiency

Suppose we want to estimate a real-valued function of θ, say $g(\theta)$, $g(\cdot) : R^k \rightarrow R$. Assume that g has continuous partial derivatives. Consider estimating $g(\theta_0)$ by $g(\hat{\theta}(\boldsymbol{X}_n))$, where $\hat{\theta}(\boldsymbol{X}_n)$ is the MLE.

Assuming that the 8 regularity conditions hold, we know that $\sqrt{n}(\hat{\theta}(\boldsymbol{X}_n) - \theta_0) \xrightarrow{D} N(0, I^{-1}(\theta_0))$. By the multivariate delta method, we know that

$$\sqrt{n}(g(\hat{\theta}(\boldsymbol{X}_n)) - g(\theta_0)) \xrightarrow{D} N(0, a(\theta_0)' I^{-1}(\theta_0) a(\theta_0))$$

where $a(\theta_0)$ is the gradient of $g(\theta)$ at θ_0.

By the information inequality, we know that among all unbiased estimators, $\delta(\boldsymbol{X}_n)$, of $g(\theta)$,

$$Var_{\theta_0}[\delta(\boldsymbol{X}_n)] = E_{\theta_0}[(\delta(\boldsymbol{X}_n) - g(\theta_0))^2] \geq a(\theta_0)' I_n^{-1}(\theta_0) a(\theta_0) \quad (1)$$

where $I_n(\theta_0) = nI(\theta_0)$.

Consider an estimator, $T(\boldsymbol{X}_n)$, of $g(\theta_0)$ which is asymptotically normal and asymptotically unbiased, i.e.,

$$\sqrt{n}(T(\boldsymbol{X}_n) - g(\theta_0)) \xrightarrow{D} N(0, V(\theta_0)) \quad (2)$$

It turns out that under some additional regularity conditions $\underline{\text{on } T(\boldsymbol{X}_n)}$, we can show that

$$V(\theta_0) \geq a(\theta_0)' I^{-1}(\theta_0) a(\theta_0) \quad (3)$$

Definition: A $\underline{\text{regular}}$ estimator $T(\boldsymbol{X}_n)$ of $g(\theta_0)$ which satisfies (2) with $V(\theta_0) = a(\theta_0)' I^{-1}(\theta_0) a(\theta_0)$ is said to be *asymptotically efficient.*

If $g(\hat{\theta}(\boldsymbol{X}_n))$ is regular, then we know that it is asymptotically efficient.

Remarks about Lower Bounds (1) and (3)

- (1) is attained only under exceptional circumstances (i.e., usually need completeness), while (3) is obtained under quite general regularity conditions.
- The UMVUE tends to be unique, while asymptotically efficient estimators are not. If $T(\boldsymbol{X}_n)$ is asymptotically efficient, then so is $T(\boldsymbol{X}_n) + R_n$, provided $\sqrt{n} R_n \xrightarrow{P} 0$.
- In (1), the estimator must be unbiased, whereas in (3), the estimator must be consistent and asymptotically unbiased.

- $V(\theta_0)$ in (3) is an asymptotic variance, whereas (1) refers to the actual variance of $\delta(\boldsymbol{X}_n)$.

For a long time, it was believed that the regularity conditions needed to make (3) hold involved regularity conditions on the density $p(x;\theta)$. This belief was exploded by Hodges.

Hodges' Example of Super-Efficiency

Suppose that $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ where the X_i's are i.i.d. $Normal(\mu_0, 1)$. We can show that $I(\mu_0) = 1$ for all μ_0. Consider the following estimator of μ_0,

$$\hat{\mu}(\boldsymbol{X}_n) = \begin{cases} \bar{\boldsymbol{X}}_n & |\bar{\boldsymbol{X}}_n| \geq n^{-1/4} \\ 0 & |\bar{\boldsymbol{X}}_n| < n^{-1/4} \end{cases}$$

Hodges showed that:

$$\sqrt{n}(\hat{\mu}(\boldsymbol{X}_n) - \mu_0) \rightarrow_D \begin{cases} N(0,1) \text{ if } \mu_0 \neq 0 \\ 0 \text{ if } \mu_0 = 0 \end{cases}$$

The latter variance makes the asymptotic distribution of the MLE inadmissible.

Remarks about Super-Efficiency

- The problem here is not due to irregularities of the density function, but due to the partialness of our estimator to $\mu_0 = 0$.
- This example shows that no regularity conditions on the density can prevent an estimator from violating (3). This possibility can only be avoided by placing restrictions on the sequence of estimators.
- LeCam (1953) showed that for any sequence of estimators satisfying (2), the set of points in Θ violating (3) has Lebesgue measure zero.
- Superefficiency shows that "parametric" models can be useful and justified when checking them in appropriate ways.

Regular Estimator

When we first discussed estimators we wanted to rule out partial estimators, i.e., estimators which favored some values of the parameters over others. In an asymptotic sense, we may want our sequence of estimators to be impartial so that we rule out estimators like the one presented by Hodges. Toward this end, we may restrict ourselves to *regular* estimators. A regular sequence of estimators is one whose asymptotic distribution remains the same in shrinking neighborhoods of the true parameter value.

More formally. consider a sequence of estimators $\{T(\boldsymbol{X}_n)\}$ and a sequence of parameter values $\{\theta_n\}$ so that $\sqrt{n}(\theta_n - \theta_0)$ is bounded. $T(\boldsymbol{X}_n)$ is a regular sequence of estimators if

$$P_{\theta_n}[\sqrt{n}(T(\boldsymbol{X}_n) - g(\theta_n)) \leq a]$$

converges to the same limit as

$$P_{\theta_0}[\sqrt{n}(T(\boldsymbol{X}_n) - g(\theta_0)) \leq a]$$

for all a. For estimators that are *regular* and satisfy (2), (3) holds. When we talk about influence functions, we will formally establish this result.

Is Hodge's Estimator Regular?

Suppose that $\mu_0 = 0$. We know that the limiting distribution of $\sqrt{n}(\hat{\mu}(\boldsymbol{X}_n) - \mu_0)$ is a degenerate distribution with point mass at zero. Consider the sequence $\mu_n = \tau/\sqrt{n}$, where τ is some positive constant. Note that $\sqrt{n}(\mu_n - \mu_0) \to \tau$. Now, we can show that $\sqrt{n}(\hat{\mu}(\boldsymbol{X}_n) - \mu_n) \overset{P(\mu_n)}{\to} -\tau$. To see this, not that

$$\begin{aligned}
P_{\mu_n}[\sqrt{n}(\hat{\mu}(\boldsymbol{X}_n) - \mu_n) = -\tau] &= P_{\mu_n}[\sqrt{n}(\hat{\mu}(\boldsymbol{X}_n) - \tau/\sqrt{n}) = -\tau] \\
&= P_{\mu_n}[\hat{\mu}(\boldsymbol{X}_n) = 0] \\
&= P_{\mu_n}[|\bar{\boldsymbol{X}}_n| < n^{-1/4}] \\
&= P_{\mu_n}[-n^{-1/4} < \bar{\boldsymbol{X}}_n < n^{-1/4}] \\
&= P_{\mu_n}[\sqrt{n}(-n^{-1/4} - \mu_n) < \sqrt{n}(\bar{\boldsymbol{X}}_n - \mu_n) < \sqrt{n}(n^{-1/4} - \mu_n)] \\
&= \Phi(\sqrt{n}(n^{-1/4} - \mu_n)) - \Phi(\sqrt{n}(-n^{-1/4} - \mu_n)) \\
&= \Phi(n^{1/4} - \tau) - \Phi(-n^{1/4} - \tau) \\
&\to 1
\end{aligned}$$

So, the limiting distribution of $\sqrt{n}(\hat{\mu}(\boldsymbol{X}_n) - \mu_n)$ is also degenerate, but with point mass at $-\tau$. This is a different limiting distribution than that of $\sqrt{n}(\hat{\mu}(\boldsymbol{X}_n) - \mu_0)$. Therefore, $\hat{\mu}(\boldsymbol{X}_n)$ is not regular.

Calculating MLE's via Iterative Algorithms

To find the MLE, we usually set the score equations equal to zero and solve. The MLE usually satisfies,

$$\sum_{i=1}^{n} \psi(X_i; \hat{\theta}(\boldsymbol{X}_n)) = 0$$

When there is no closed form solution, we can use an iterative method to solve the score equations. At the kth iteration, we have a proposed solution to the above equation, $\theta^{(k)}$. We update the solution by noting that

$$0 = \sum_{i=1}^{n} \psi(X_i; \hat{\theta}(\boldsymbol{X}_n)) = \sum_{i=1}^{n} \psi(X_i; \theta^{(k)}) - nJ_n^*(\boldsymbol{X}_n)(\hat{\theta}(\boldsymbol{X}_n) - \theta^{(k)})$$

where θ_n^* falls between $\hat{\theta}(\boldsymbol{X}_n)$ and $\theta^{(k)}$.

This implies that

$$\hat{\theta}(\boldsymbol{X}_n) = \theta^{(k)} + \{nJ_n^*(\boldsymbol{X}_n)\}^{-1} \sum_{i=1}^n \psi(X_i; \theta^{(k)})$$

What should be use for $J_n(\theta_n^*)$. One proposal is to use $J_n(\theta^{(k)})$, so that

$$\theta^{(k+1)} = \theta^{(k)} + \{nJ_n(\theta^{(k)})\}^{-1} \sum_{i=1}^n \psi(X_i; \theta^{(k)})$$

If we have a closed form solution for $I(\theta)$, we can replace $J_n(\theta_n^*)$ by $I(\theta^{(k)})$, so that

$$\theta^{(k+1)} = \theta^{(k)} + \{nI(\theta^{(k)})\}^{-1} \sum_{i=1}^n \psi(X_i; \theta^{(k)})$$

This latter substitution is called *Fisher scoring*.

With either substitution, we iterate to convergence.

Why does this work?

Caution: Be careful to make sure that you have identified the MLE as opposed to a local maximum or minimum.

Estimating the Asymptotic Variance

We have already proved under appropriate regularity conditions that

$$\sqrt{n}(\hat{\theta}(\boldsymbol{X}_n) - \theta_0) \xrightarrow{D} N(0, I^{-1}(\theta_0))$$

In order to use this result for inferential purposes, we need to be able to approximate the limiting distribution. This means we have to be able to estimate $I(\theta_0)$ and $I^{-1}(\theta_0)$. This follows easily from the results that be have already derived.. Specifically, we know that

$$J_n(\hat{\theta}(\boldsymbol{X}_n)) \xrightarrow{P} I(\theta_0) \ \ ; \ \ \{J_n(\hat{\theta}(\boldsymbol{X}_n))\}^{-1} \xrightarrow{P} I^{-1}(\theta_0)$$

and

$$I(\hat{\theta}(\boldsymbol{X}_n)) \xrightarrow{P} I(\theta_0) \ \ ; \ \ \{I(\hat{\theta}(\boldsymbol{X}_n))\}^{-1} \xrightarrow{P} I^{-1}(\theta_0)$$

Testing $H_0 : \boldsymbol{f}(\theta) = \boldsymbol{f}(\theta_0)$; Confidence Ellipses for $\boldsymbol{f}(\theta_0)$

By the multivariate delta method, we know that

$$\sqrt{n}(\boldsymbol{f}(\hat{\theta}(\boldsymbol{X}_n)) - \boldsymbol{f}(\theta_0)) \xrightarrow{D} N(0, \bigtriangledown \boldsymbol{f}(\theta_0) I^{-1}(\theta_0) \bigtriangledown \boldsymbol{f}(\theta_0)')$$

This implies that

$$n(\boldsymbol{f}(\hat{\theta}(\boldsymbol{X}_n)) - \boldsymbol{f}(\theta_0))'\{\bigtriangledown \boldsymbol{f}(\theta_0) I^{-1}(\theta_0) \bigtriangledown \boldsymbol{f}(\theta_0)'\}^{-1}(\boldsymbol{f}(\hat{\theta}(\boldsymbol{X}_n)) - \boldsymbol{f}(\theta_0)) \xrightarrow{D} \chi_q^2$$

Note that

$$n(\boldsymbol{f}(\hat{\theta}(\boldsymbol{X}_n))-\boldsymbol{f}(\theta_0))'\{\triangledown\boldsymbol{f}(\hat{\theta}(\boldsymbol{X}_n))J_n(\hat{\theta}(\boldsymbol{X}_n))^{-1}\triangledown\boldsymbol{f}(\hat{\theta}(\boldsymbol{X}_n))'\}^{-1}(\boldsymbol{f}(\hat{\theta}(\boldsymbol{X}_n))-\boldsymbol{f}(\theta_0)) \tag{4}$$

is equal to

$$n(\boldsymbol{f}(\hat{\theta}(\boldsymbol{X}_n)) - \boldsymbol{f}(\theta_0))'\{\triangledown\boldsymbol{f}(\theta_0)I^{-1}(\theta_0) \triangledown \boldsymbol{f}(\theta_0)'\}^{-1}(\boldsymbol{f}(\hat{\theta}(\boldsymbol{X}_n)) - \boldsymbol{f}(\theta_0)) + \tag{5}$$

$$\begin{aligned}\sqrt{n}(\boldsymbol{f}(\hat{\theta}(\boldsymbol{X}_n)) - \boldsymbol{f}(\theta_0))'\{\{\triangledown\boldsymbol{f}(\hat{\theta}(\boldsymbol{X}_n))J_n(\hat{\theta}(\boldsymbol{X}_n))^{-1} \triangledown \boldsymbol{f}(\hat{\theta}(\boldsymbol{X}_n))'\}^{-1} - \\ \{\triangledown\boldsymbol{f}(\theta_0)I^{-1}(\theta_0) \triangledown \boldsymbol{f}(\theta_0)'\}^{-1}\}\sqrt{n}(\boldsymbol{f}(\hat{\theta}(\boldsymbol{X}_n)) - \boldsymbol{f}(\theta_0))\end{aligned} \tag{6}$$

(5) converges in distribution to χ^2_q and (6) converges in probability to zero. By Slutsky's theorem, we know that (4) converges in distribution to χ^2_q. So (4) can be used for testing $H_0 : \boldsymbol{f}(\theta) = \boldsymbol{f}(\theta_0)$ or for forming confidence ellipses (intervals) for $\boldsymbol{f}(\theta_0)$.

Many problems can be viewed as using “incomplete” data. For such problems:

Direct solution of the score equations can be unstable - infeasible.

The EM algorithm (Dempster, Laird, and Rubin, 1977) can greatly facilitate optimization.