1.5.2 Observed Information

The information matrix is not random, but it is also not observable from the data.

You need knowledge of the distribution to calculate it.

would be great to use $I(\hat{\underline{\theta}}_{MLE}) = E\left\{-\frac{\partial^2}{\partial\underline{\theta}\,\partial\underline{\theta}^T}\log f(Y_i;\underline{\theta})\Big|_{\underline{\theta}=\hat{\underline{\theta}}_{MLE}}\right\}$

Let $Y_1, \ldots, Y_n$ be iid with density $f_Y(y_i;\boldsymbol{\theta})$. The log likelihood is defined as

$$\log L(\underline{\theta}|\underline{Y}) = \sum_{i=1}^{n}\log f_Y(Y_i;\underline{\theta})$$

taking two derivatives and dividing by n results in

define: $\bar{I}(\underline{Y},\underline{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left\{-\frac{\partial^2}{\partial\underline{\theta}\,\partial\underline{\theta}^T}\log f(Y_i;\underline{\theta})\right\}$

average curvature contribution.

if $I(\underline{\theta}) = E\left\{-\frac{\partial^2}{\partial\underline{\theta}\,\partial\underline{\theta}^T}\log f(Y_i;\underline{\theta})\right\}$ then $\bar{I}(\underline{Y},\underline{\theta})$ would be an obvious estimator *if we know $\underline{\theta}$*!

$\Rightarrow \bar{I}(\underline{Y},\hat{\underline{\theta}}_{MLE})$ seems like a natural estimator for $I(\underline{\theta})$.

Definition: The matrix $n\bar{I}\,(Y;\hat{\boldsymbol{\theta}}_{\text{MLE}})$ is called the sample information matrix, or the *observed information matrix*.

doesn't depend on sample size.

Note: $I(\underline{\theta})$ is the expected curvature of the log-likelihood surface from one observation

the observed information matrix $n\bar{I}(\underline{y}, \hat{\underline{\theta}}_{MLE})$ is from a sample of size n and does depend on sample size.

Recall $\hat{\underline{\theta}}_{MLE} \overset{\circ}{\sim} N(\underline{\theta}, \{n I(\underline{\theta})\}^{-1})$. To get an approximate variance of $\hat{\underline{\theta}}_{MLE}$ for a sample of size n, we need that matrix to depend on n.

Why use $I(\boldsymbol{\theta}) = \mathrm{E}\left[-\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\log f(Y_1;\boldsymbol{\theta})\right]$ as the basis for an estimator, rather than $I(\boldsymbol{\theta}) = \mathrm{E}\left[\left\{\frac{\partial}{\partial\boldsymbol{\theta}^\top}\log f(Y_1;\boldsymbol{\theta})\right\}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f(Y_1;\boldsymbol{\theta})\right\}\right]$?

The hessian (curvature) @ $\hat{\theta}_{MLE}$ is readily available from optimization methods $\Rightarrow$ $n\bar{I}(\underline{y}, \hat{\theta}_{MLE})$ can be computed easily.

Alternatively could use $\bar{I}^*(\underline{y}, \underline{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\partial}{\partial\underline{\theta}^T}\log f(y_i;\underline{\theta})\right\}\left\{\frac{\partial}{\partial\underline{\theta}}\log f(y_i;\underline{\theta})\right\}$

because $E[\bar{I}^*(y,\underline{\theta})] = I(\underline{\theta})$ also.

We'll see this again in misspecified models (and how to "correct" them) — robustness vs. efficiency.

e.g. Estimating Equations

Now let's prove the asymptotic normality of the MLE (in the scalar case).

Useful facts: For $X_1, \ldots, X_n$ iid with $\operatorname{Var} X_1 = \sigma^2 < \infty$,

WLLN: $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} E[X_1]$

CLT: $\sqrt{n}\,(\bar{X}_n - EX_1) \xrightarrow{d} N(0, \sigma^2)$.

Let $Y_i \overset{iid}{\sim} f_Y(y;\theta)$ and $\hat{\theta}_{MLE}$ is such that $S(\hat{\theta}_{MLE}) = \frac{d}{d\theta}\ell(\theta)\Big|_{\theta=\hat{\theta}_{MLE}} = 0$.

Let $S(\theta) = \frac{d}{d\theta}\ell(\theta) = \sum_{i=1}^{n} \frac{d}{d\theta}\log f(Y_i;\theta)$

$= \sum_{i=1}^{n} s(Y_i;\theta)$ where $s(Y_i;\theta) = \frac{d}{d\theta}\log(Y_i;\theta)$.

We know $E[s(Y_i;\theta)] = 0$ and $\operatorname{Var}[s(Y_i;\theta)] = I(\theta)$ and $\{s(Y_i,\theta)\}_{i=1}^{n}$ are iid r.v.'s.

$\Rightarrow \sqrt{n}\left(\frac{1}{n}S(\theta) - 0\right) \xrightarrow{d} N(0, I(\theta))$ by CLT.

$\Leftrightarrow (nI(\theta))^{-1/2} S(\theta) \xrightarrow{d} Z$, $Z \sim N(0,1)$ $(*)$.

Secondly, let $J(\theta) = -\sum_{i=1}^{n} \frac{d^2 \log f_Y(Y_i;\theta)}{d\theta^2} = \underbrace{-\sum_{i=1}^{n} \frac{d}{d\theta} s(Y_i;\theta)}_{\text{sum of iid r.v.'s}}$ $+$ $E\left[-\frac{d}{d\theta}s(Y_i;\theta)\right] = I(\theta)$.

and so, $\frac{1}{n}J(\theta) \xrightarrow{P} I(\theta)$ by WLLN $\Leftrightarrow$ $nJ^{-1}(\theta) \xrightarrow{P} I(\theta)^{-1}$ $(**)$.

So far we have been considering the true value θ. Let $\ell(\theta)$ be sufficiently smooth to allow for Taylor Expansion.

$\Rightarrow 0 \overset{\text{assumption}}{=} S(\hat{\theta}_{MLE}) \overset{\text{smoothness assumption}}{\approx} S(\theta) + \frac{dS(\theta)}{d\theta}(\hat{\theta}_{MLE} - \theta) \Leftrightarrow \hat{\theta}_{MLE} - \theta \approx -\frac{1}{\frac{dS(\theta)}{d\theta}} \cdot S(\theta)$.

$= J(\theta)^{-1} S(\theta)$

Thus $\overbrace{\sqrt{n}\, I(\theta)^{1/2}(\hat{\theta}_{MLE} - \theta)}^{\text{The thing we want } \xrightarrow{d} N(0,1)} \approx \sqrt{n}\, I(\theta)^{1/2} J(\theta)^{-1} S(\theta)$

$= \{nI(\theta)\}^{1/2} J(\theta)^{-1} \{nI(\theta)\}^{1/2} \{nI(\theta)\}^{-1/2} S(\theta)$

$= I(\theta)^{1/2} \underbrace{nJ^{-1}}_{\xrightarrow{P} I(\theta)^{-1}\ (**)} I(\theta)^{1/2} \underbrace{\{nI(\theta)\}^{-1/2} S(\theta)}_{\xrightarrow{d} Z\ (*)}$

$\xrightarrow{d} N(0,1)$ by Slutsky's Theorem. //

Note: the argument to replace $I(\theta)$ by $\bar{I}(\hat{\theta}_{MLE})$ in the asymptotic result is justified by convergence in probability.

This argument is generalized to $\underline{\theta}$ by interpreting the score as a $b \times 1$ vector, $I(\underline{\theta})$ as $b \times b$ matrix, $Z \sim N_b(\underline{0}, I_b)$ dsn