1.5 The Maximum Likelihood Estimator and the Information Matrix

We have now talked about how to construct likelihoods in a variety of settings, now we can use those constructions to formalize how we make inferences about model parameters.

parameter estimation, hypothesis tests, confidence intervals.

We often restrict attention to likelihoods that are continuously differentiable wrt $\underline{\theta}$.

In this case, Recall the score function

$$S(\underline{\theta}) = S(\boldsymbol{Y}, \boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial \ell(\underline{\theta})}{\partial \theta_1} \\ \vdots \\ \frac{\partial \ell(\underline{\theta})}{\partial \theta_b} \end{pmatrix} = \begin{pmatrix} \frac{\partial \log L(\underline{\theta}|\underline{Y})}{\partial \theta_1} \\ \vdots \\ \frac{\partial \log L(\underline{\theta}|\underline{Y})}{\partial \theta_b} \end{pmatrix}$$

This function is random because it depends on the data $\underline{Y}$.

?!

Generally, the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ is the value of $\boldsymbol{\theta}$ where the maximum (over the parameter space Θ) of $L(\boldsymbol{\theta}|\boldsymbol{Y})$ is attained.

$$\hat{\underline{\theta}}_{MLE} = \underset{\underline{\theta}}{\arg\max}\, L(\underline{\theta}|\underline{Y}) \iff L(\hat{\underline{\theta}}_{MLE}|\underline{Y}) \geq L(\underline{\theta}|\underline{Y}) \quad \forall \underline{\theta} \in \Theta.$$

Under the assumption that the log-likelihood is continuously differentiable, then

$$S(\hat{\underline{\theta}}_{MLE}) = 0.$$

But not always (?!).

Example (Exponential threshold model): Suppose that $Y_1, \ldots, Y_n$ are iid from the exponential distribution with a threshold parameter μ,

strict.

$$f(y;\mu) = \begin{cases} \exp\{-(y-\mu)\} & \mu \boxed{<} y < \infty \\ 0 & \text{otherwise,} \end{cases}$$

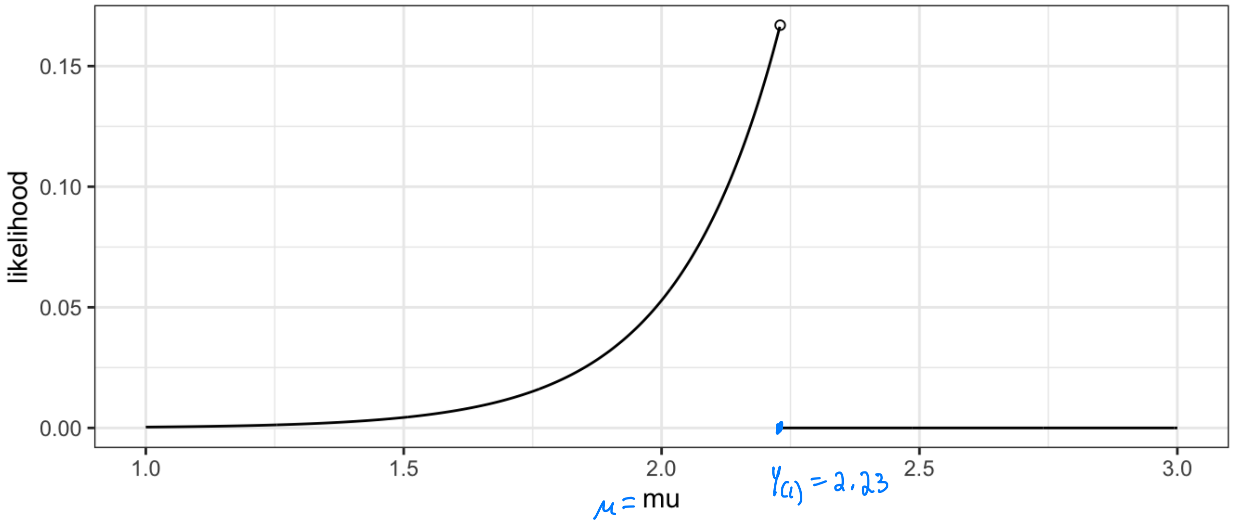
for $\infty < \mu < \infty$.

$$L(\mu | \underline{Y}) = \prod_{i=1}^{n} f(Y_i;\mu) = \prod_{i=1}^{n} \exp(-(Y_i - \mu))\, \mathbb{I}(\mu < Y_i).$$

$$= \exp(-n\bar{Y})\exp(n\mu) \underbrace{\prod_{i=1}^{n} \mathbb{I}(\mu < Y_i)}_{\mathbb{I}(\mu < Y_{(1)}).}.$$

$\Rightarrow$ the likelihood $= 0$ for any value of $\mu \geq Y_{(1)}$, $Y_{(1)} = \min\{Y_1, \ldots, Y_n\}$.

Consider the artificial data set $\boldsymbol{y} = [2.47, 2.35, 2.23, 3.53, 2.36]$.



$\underline{\hat{\mu}_{MLE} = 2.23}$, right? $L(2.23 | \underline{Y}) = 0$ $\Rightarrow$ $(*)$ $L(\hat{\mu}_{MLE} | Y) \not\geq L(\mu | Y)$ $\forall \mu \in \mathbb{R}$. AND

yes.

$S(\hat{\mu}_{MLE}) \neq 0$ because $\ell(\hat{\mu}_{MLE})$ not differentiable here.

If replace $\mu < y$ with $\mu \leq y$ in $f(y;\mu)$ then $(*)$ will hold by not score equation.

to see this consider maximizing $L_h(\mu | \underline{Y}) = \left(\frac{1}{2h}\right)^n \prod_{i=1}^{n} \{F_Y(Y_i + h;\mu) - F_Y(Y_i - h;\mu)\}$ for "small enough" value of h.

Then maximize $\lim_{h \to 0^+} L_h(\mu | \underline{Y})$. $\Rightarrow$ $\hat{\mu}_{MLE} = Y_{(1)}$.

Rest of this section: assume support doesn't depend on the parameter value.

1.5.1 The Fisher Information Matrix

$Y_i \overset{iid}{\sim} f(y;\underline{\theta})$

The Fisher information matrix $\boldsymbol{I}(\boldsymbol{\theta})$ is defined as the $b \times \overset{\text{dimension of } \underline{\theta}}{b}$ matrix where

$$I_{ij}(\boldsymbol{\theta}) = E\left[\left\{\frac{\partial}{\partial\theta_i}\log f(Y_1;\underline{\theta})\right\}\left\{\frac{\partial}{\partial\theta_j}\log f(Y_1;\underline{\theta})\right\}\right]$$

Is this random? No! it's an expectation!

Notice: this is the "information" in one observation. (note Y_1).

In matrix form,

$$\boldsymbol{I}(\boldsymbol{\theta}) = E\Big[\underbrace{\left(\frac{\partial}{\partial\underline{\theta}^T}\log f(Y_1;\underline{\theta})\right)}_{\text{column vector.}}\underbrace{\left(\frac{\partial}{\partial\underline{\theta}}\log f(Y_1;\underline{\theta})\right)}_{\text{row vector}}\Big]$$

$\parallel$

Let $s(y;\underline{\theta}) = \left\{\frac{\partial}{\partial\underline{\theta}}\log f(y;\underline{\theta})\right\}^T$ ← column vector.

↑ score contribution.

Then $\underline{I}(\underline{\theta}) = E\left[s(Y_1;\underline{\theta})\,s(Y_1;\underline{\theta})^T\right]$.

↑

Again this depends on 1 observation (not n of them).

Fisher information facts:

1. The Fisher information matrix is the variance of the score contribution.

Why? $E[S(Y_1, \underline{\theta})] = 0$

Fact ① from GLM section.

Big Result

2. If regularity conditions are met,

defined wrt a single observation.

$$\underbrace{\sqrt{n}}_{\text{based on } n \text{ observations.}}(\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}) \xrightarrow{d} \text{N}_b(\underset{\uparrow \text{ unbiased}}{\mathbf{0}}, \boldsymbol{I}(\boldsymbol{\theta})^{-1}).$$

inverse Fisher information.

"approximately distributed"

$\Rightarrow$ if n is large $\sqrt{n}(\hat{\underline{\theta}}_{MLE} - \underline{\theta}) \overset{\cdot}{\sim} N(\underline{0}, I(\underline{\theta})^{-1})$

or, $\hat{\underline{\theta}}_{MLE} - \underline{\theta} \overset{\cdot}{\sim} N(\underline{0}, \frac{1}{n} I(\underline{\theta})^{-1})$

$\hat{\underline{\theta}}_{MLE} \overset{\cdot}{\sim} N(\underline{\theta}, \{nI(\underline{\theta})\}^{-1})$ $\underline{(*)}$.

We will prove this result for $b=1$. (later).

3. If $b = 1$, then any unbiased estimator must have variance greater than or equal to $\underline{\{nI(\boldsymbol{\theta})\}^{-1}}$

↰ Cramer-Rao lower bound.

If $b \neq 1$: If Σ is the asymptotic cov matrix of any other consistent estimator, then

$$\Sigma - I(\theta)^{-1}$$ is positive definite.

4. The information matrix is related to the $\underbrace{\text{curvature}}_{\text{Hessian}}$ of the log-likelihood contribution.

$$I(\underline{\theta}) = E\left[\left(\frac{\partial}{\partial \underline{\theta}^T} \log f(Y_1; \underline{\theta})\right)\left(\frac{\partial}{\partial \underline{\theta}} \log f(Y_1; \underline{\theta})\right)\right]$$

$$= E\left[-\frac{\partial^2}{\partial \underline{\theta}\, \partial \underline{\theta}^T} \log f(Y_1; \underline{\theta})\right]$$ ← assuming ℓ is twice differentiable and using fact ② from GLM section.

$$= E\left[-\frac{\partial}{\partial \underline{\theta}} s(Y_1; \underline{\theta})\right]$$ (writing another way).

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

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

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 $\underline{\theta}$ *

$\Rightarrow$ $\bar{I}(\underline{Y}, \hat{\theta}_{MLE})$ 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.*

Note: $I(\underset{\sim}{\theta})$ is the expected curvature of the log-likelihood surface from one observation ↗ doesn't depend on sample size

The observed information $n\bar{I}(\underset{\sim}{Y}; \hat{\theta}_{MLE})$ is from a sample of size n and does depend on sample size.

Recall $\hat{\theta}_{MLE} \overset{\cdot}{\sim} N(\underset{\sim}{\theta}, \{nI(\theta)\}^{-1})$ (*)

To get approximate variance of $\hat{\theta}_{MLE}$ for sample of size n, we need the 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 ⇒

$n\bar{I}(\underset{\sim}{Y}, \hat{\theta}_{MLE})$ can be computed easily.

Alternatively could use $\bar{I}^*(\underset{\sim}{Y}, \underset{\sim}{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left[\left\{\frac{\partial}{\partial \underset{\sim}{\theta}^T}\log f(Y_i; \underset{\sim}{\theta})\right\}\left\{\frac{\partial}{\partial \underset{\sim}{\theta}}\log f(Y_i; \underset{\sim}{\theta})\right\}\right]$

(because $E[\bar{I}^*(\underset{\sim}{Y}, \underset{\sim}{\theta})] = I(\underset{\sim}{\theta})$ also).

This is not typically used unless specification of f is less clear (model misspecification).

($\bar{I}(\underset{\sim}{Y}, \underset{\sim}{\theta})$ is more efficient)

We will see this again later.

Now let's prove the asymptotic normality of the MLE (in the $\underline{\text{scalar case}}^{b=1.}$).

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

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

$$\text{CLT}: \sqrt{n}\left(\bar{X}_n - EX_1\right) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n X_i - EX_1\right) \xrightarrow{d} N(0, \sigma^2).$$

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

Let $$S(\theta) = \frac{d}{d\theta} l(\theta) = \sum_{i=1}^n \frac{d}{d\theta} \log f(Y_i;\theta)$$
$$= \sum_{i=1}^n s(Y_i, \theta) \quad \text{where} \quad s(Y_i,\theta) = \frac{d}{d\theta}\log f(Y_i;\theta)$$

We know $E[s(Y_1,\theta)] = 0$ and $\operatorname{Var}[s(Y_1,\theta)] = I(\theta)$ and $\{s(Y_i,\theta)\}_{i=1}^n$ are iid r.v.'s.

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

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

Secondly, let $$J(\theta) = -\sum_{i=1}^n \underbrace{\frac{d^2 \log f_Y(Y_i;\theta)}{d\theta}}_{\text{sum of iid R.V.'s}} = -\sum_{i=1}^n \frac{d}{d\theta} s(Y_i,\theta). \text{ Then } E\left[-\frac{d}{d\theta}s(Y_1,\theta)\right] = I(\theta).$$

$$\Rightarrow \frac{1}{n} J(\theta) \xrightarrow{P} I(\theta) \text{ by WLLN} \Longleftrightarrow nJ^{-1}(\theta) \xrightarrow{P} I(\theta)^{-1} \quad (**)$$

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

$$0 \overset{\text{assumption.}}{=} S(\hat{\theta}_{MLE}) \approx S(\theta) + \frac{dS(\theta)}{d\theta}(\hat{\theta}_{MLE} - \theta) \Longleftrightarrow \hat{\theta}_{MLE} - \theta \approx -\frac{1}{\frac{dS(\theta)}{d\theta}} \cdot S(\theta).$$
$$= \underline{J(\theta)^{-1} S(\theta).}$$

Thus, $$\overbrace{\sqrt{n}\, I(\theta)^{1/2}(\hat{\theta}_{MLE} - \theta)}^{\text{The thing we want} \xrightarrow{d} N(0,1).} \approx \sqrt{n}\, I(\theta)^{1/2} \underline{J(\theta)^{-1} S(\theta)}$$
$$= \underbrace{\{nI(\theta)\}^{1/2} J(\theta)^{-1} \{nI(\theta)\}^{1/2}}_{I(\theta)\,\underbrace{nJ(\theta)^{-1}}_{\xrightarrow{P} I(\theta)^{-1}\ (**).}} \underbrace{\{nI(\theta)\}^{-1/2} S(\theta).}_{\xrightarrow{d} Z\ (*).}$$

$$\xrightarrow{d} N(0,1) \text{ by Slutsky's Th'm } //$$

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

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