$$\mathrm{E}[\hat{\sigma}^2_{\mathrm{MLE}}] = E\left[\frac{1}{n}\sum_{i=1}^n \frac{1}{4}(Y_{i_1} - Y_{i_2})^2\right]$$

$$\text{(iid)} \quad = \frac{1}{4} E\left[(Y_{i_1} - Y_{i_2})^2\right]$$

$$= \frac{1}{4} E\left[\left((Y_{i_1} - \mu_i) - (Y_{i_2} - \mu_i)\right)^2\right]$$

$$= \frac{1}{4} E\left[(Y_{i_1} - \mu_i)^2 - 2(Y_{i_1} - \mu_i)(Y_{i_2} - \mu_i) + (Y_{i_2} - \mu_i)^2\right]$$

$$= \frac{1}{4}\left[\sigma^2 - 0 + \sigma^2\right]$$

$$= \frac{\sigma^2}{2} \neq \sigma^2 !$$

So as $n \to \infty$, $\hat{\sigma}^2_{MLE} \xrightarrow{p} \frac{\sigma^2}{2}$ by WLLN

This seems bad.

Happens because # of nuisance parameters grows w/ n (# groups).

A reworking of the data seems more promising. Let,

$$V_i = \frac{Y_{i1} - Y_{i2}}{\sqrt{2}} \qquad \text{and} \qquad W_i = \frac{Y_{i1} + Y_{i2}}{\sqrt{2}}$$

Because Y_{ij} are Gaussian,

$V_i \sim N(0, \sigma^2)$ and $W_i \sim N(\sqrt{2}\mu_i, \sigma^2)$ Also, $V_i \perp W_i$!

$$\begin{bmatrix} V_i \\ W_i \end{bmatrix} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} Y_{i_1} \\ Y_{i_2} \end{bmatrix} \Rightarrow \mathrm{Var}\left(\begin{bmatrix} V_i \\ W_i \end{bmatrix}\right) = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} = \sigma^2\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Consider the density of V & W:

$$f_{VW}(\underline{v}, \underline{w}; \sigma^2, \mu_1, \ldots, \mu_n) = \underbrace{f_V(\underline{v}; \sigma^2)}_{\text{no nuisance parameters!}} f_W(\underline{w}; \mu_1, \ldots, \mu_n, \sigma^2)$$

$$\Rightarrow \ell(\sigma \mid \underline{V}) = -n \log\sqrt{2\pi} - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^n V_i^2$$

$$\frac{\partial \ell(\sigma \mid \underline{V})}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^n V_i^2 \Rightarrow \hat{\sigma}^2_{MMLE} = \frac{1}{n}\sum_{i=1}^n V_i^2$$

A marginal likelihood approach is simple provided you can find a statistic V whose dsn is free of the nuisance parameter!

For conditional likelihoods, we can often exploit the existence of sufficient statistics for the nuisance parameters under the assumption that the parameter of interest is known.

Let T_i be sufficient for a nuisance parameter (μ_i)

Then the conditional dsn of the data given $\underline{T} = (T_1, \dots, T_n)$ doesn't depend on the nuisance parameters!

$\Longrightarrow$ we can look for the conditional dsn of the data | T.

Example (Exponential Families): The structure of exponential families is such that it is often possible to exploit their properties to eliminate nuisance parameters. Let Y have a density of the form

$$f(y;\boldsymbol{\eta}) = h(y)\exp\left\{\sum_{i=1}^{s}\eta_i T_i(y) - A(\boldsymbol{\eta})\right\},$$

then

if $\eta = (\theta_1^T, \theta_2^T)$

These are sufficient statistics

$$f(y; \underline{\theta}_1, \underline{\theta}_2) = h(y)\exp\left\{\sum \theta_{1i} W_i + \sum \theta_{2j} V_j - A(\underline{\theta}_1, \underline{\theta}_2)\right\}$$

where
the conditional distribution of $\underline{W}$ given $\underline{V}$ is exponential family of the form

$$f(\underline{w} \mid \underline{v}; \underline{\theta}_1) = g_{\underline{v}}(\underline{w}) \exp\left(\sum_{i=1}^{r} \theta_{1i} w_i - A_{\underline{v}}(\underline{\theta}_1)\right)$$

depends only on $\underline{\theta}_1$!

Thus, exponential families often provide an automatic procedure for finding $\boldsymbol{W}$ and $\boldsymbol{V}$.

Example (Logistic Regression): For binary Y_i, the standard logistics regression model is

$$P(Y_i = 1) = p_i(\boldsymbol{x}_i, \boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}$$

and the likelihood is

$$L(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{X}) = \prod_{i=1}^{n} p_i(x_i, \beta)^{Y_i} \{1 - p_i(x_i, \beta)\}^{1-Y_i}$$

$$= \prod_{i=1}^{n} \left\{ \frac{\exp(x_i^T \beta)}{1+\exp(x_i^T\beta)} \right\}^{Y_i} \left\{ \frac{1}{1+\exp(x_i^T\beta)} \right\}^{1-Y_i}$$

$$= \frac{\exp\left(\sum_{i=1}^n Y_i (x_i^T \beta)\right)}{\prod_{i=1}^n \left(1 + \exp(x_i^T \beta)\right)}$$

$$= c(x, \beta) \exp\left(\sum_{j=1}^{p} \beta_j \sum_{i=1}^{n} x_{ij} Y_i \right)$$

$\Rightarrow T_j = \sum_{i=1}^n x_{ij} Y_i \quad j = 1, \ldots, p$ are sufficient for β_j's.

Suppose $\theta_1 = \beta_k$ is the parameter of interest (e.g. treatment variable) and others are nuisance parameters.

$\Rightarrow W_1 = T_k = \sum_{i=1}^n x_{ik} Y_i$ and $\underset{\sim}{V} = (T_1, \ldots, T_{k-1}, T_{k+1}, \ldots, T_p)^T$

$\frac{\text{joint}}{\text{marginal}}$

and the conditional density $P(W = t_k \mid \underset{\sim}{V} = \underset{\sim}{v}) = P(T_k = t_k \mid T_1 = t_1, \ldots, T_{k-1} = t_{k-1}, T_{k+1} = t_{k-1}, \ldots, T_p = t_p)$

$$\vdots$$

$$= \frac{c(t_1, \ldots, t_p) \exp(\beta_k t_k)}{\sum_u c(t_1, \ldots, t_{k-1}, u, t_{k+1}, \ldots, t_p) \exp(\beta_k u)}$$

depends only on β_k!

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 will 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.

$$\hat{\underline{\theta}}_{MLE} = \underset{\underline{\theta}}{\operatorname{argmax}}\, L(\underline{\theta}|\underline{Y}) \iff L(\hat{\underline{\theta}}_{MLE}|\underline{Y}) \geq L(\theta|\underline{Y}) \;\; \forall\, \theta \in \Theta$$

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

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

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

support depends on parameter.

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

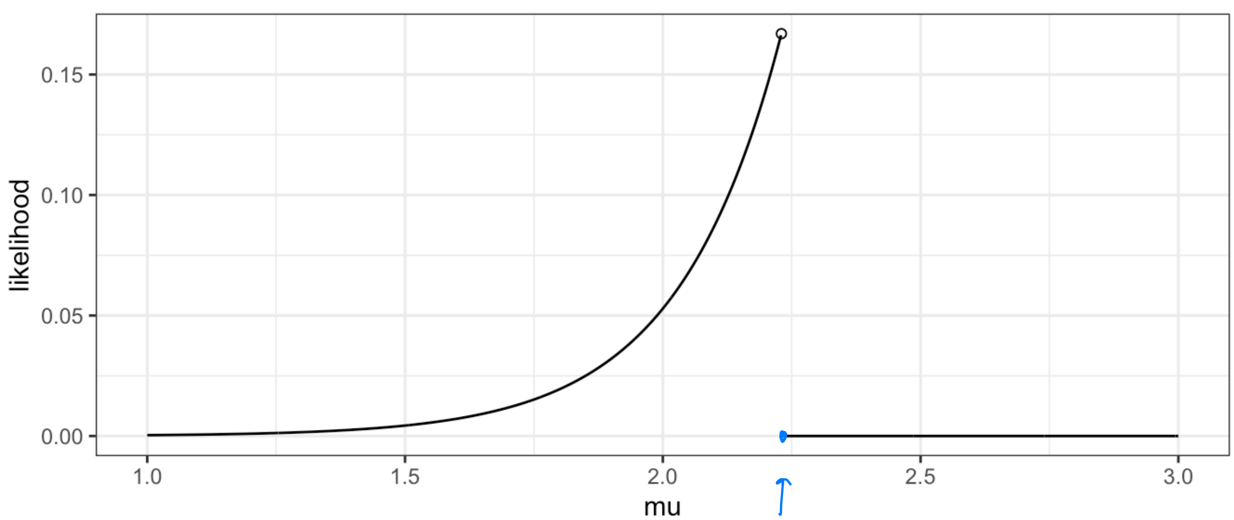
for $\infty < \mu < \infty$.

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

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

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

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



$Y_{(1)} = 2.23$

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

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

If replace $\mu < y$ with $\mu \leq y$ in $f(y;\mu)$ then $(*)$ will hold, but not the 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 $\lim_{h \to 0^+} L_h(\mu|\underline{Y})$.

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

1.5.1 The Fisher Information Matrix

$\underline{\theta} \in \mathbb{R}^b$

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

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

Note:

This is the "information" in one observation.

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

In matrix form,

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

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

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

again just depends on 1 observation (not n)

Fisher information facts:

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

Why? $E[s(y_i;\underline{\theta})] = 0$

fact (1) from GLM section / Homework

Big Result

2. If regularity conditions are met,

defined wrt a single observation.

$$\underbrace{\sqrt{n}(\hat{\boldsymbol{\theta}}_{\text{MLE}} - \underline{\theta})}_{\text{based on } n \text{ observations}} \xrightarrow{d} \text{N}_b(0, \underbrace{\boldsymbol{I}(\boldsymbol{\theta})^{-1}}_{\text{inverse Fisher information}}).$$

unbiased.

"approximately" distributed as

in practice

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

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

$$\hat{\theta}_{MLE} \overset{\cdot}{\sim} N(\underline{\theta}, \underbrace{\tfrac{1}{n} I(\underline{\theta})^{-1}}_{= \{n I(\theta)\}^{-1}})$$

We will prove this result for $b=1$ next week.

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

Cramer-Rao lower bound.

If $b>1$: If Σ is the asymptotic covariance 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.

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

$$= E\left[\left(\frac{\partial}{\partial \underline{\theta}} \log f(Y_i;\underline{\theta})\right)^2\right]$$

$$= E\left[-\frac{\partial^2}{\partial \underline{\theta}\, \partial \underline{\theta}^T} \log f(Y_i;\underline{\theta})\right]$$

assuming ℓ is twice differentiable and uses fact (2) from GLM section/homework.

$$= E\left[-\frac{\partial}{\partial \underline{\theta}} s(Y_i;\underline{\theta})\right]$$ (written a different way)

Information is related to Hessian.

Comments:

(2) + (4) often combined to produce asymptotic confidence intervals for MLE's.

Equivalence of $E\left[\left\{\frac{\partial}{\partial \underline{\theta}^T} \log f(Y_i;\underline{\theta})\right\}\left\{\frac{\partial}{\partial \underline{\theta}} \log f(Y_i;\underline{\theta})\right\}\right]$ and $E\left[-\frac{\partial^2}{\partial \theta\, \partial \underline{\theta}^T} \log f(Y_i;\underline{\theta})\right]$

relies on $Y_1, \ldots, Y_n \overset{iid}{\sim} f(y;\underline{\theta})$

When this is not true, equality will not hold and (4) doesn't hold. (more later in the course)