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.

Recall the score function

$$S(\boldsymbol{Y}, \boldsymbol{\theta}) =$$

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.

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

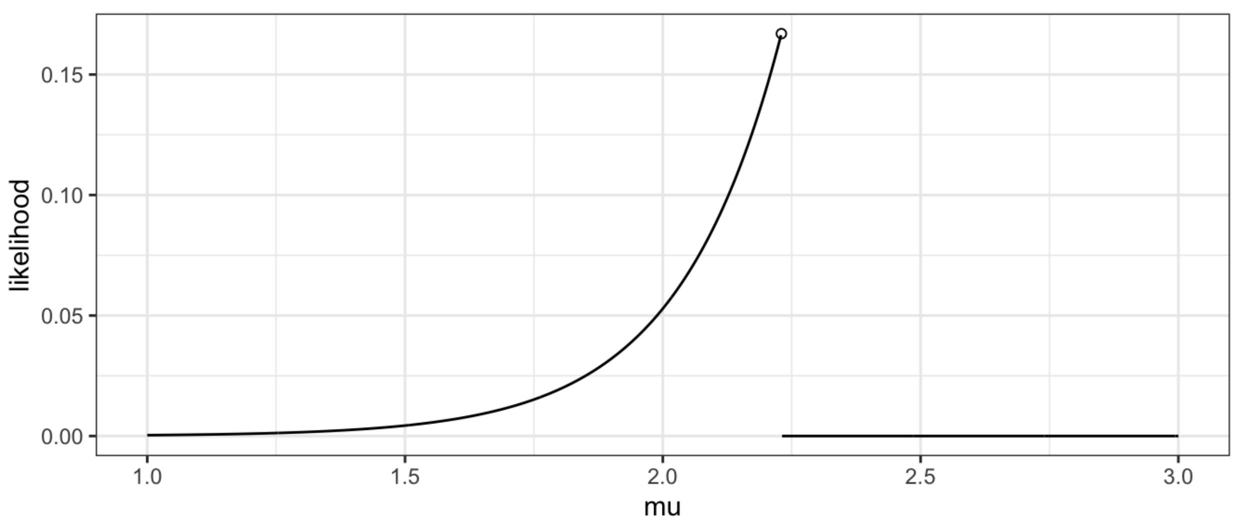
But not always (?!).

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

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

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

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



1.5.1 The Fisher Information Matrix

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

$$I_{ij}(\boldsymbol{\theta}) =$$

In matrix form,

$$I(\boldsymbol{\theta}) =$$

Fisher information facts:

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

2. If regularity conditions are met,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\text{MLE}} - \theta) \xrightarrow{d} \text{N}_b(0, I(\boldsymbol{\theta})^{-1}).$$

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

4. The information matrix is related to the curvature of the log-likelihood contribution.

1.5.2 Observed Information

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

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

taking two derivatives and dividing by n results in

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

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 $I(\boldsymbol{\theta}) = \mathrm{E}\left[\left\{\frac{\partial}{\partial\boldsymbol{\theta}^\top}\log f(Y_1;\boldsymbol{\theta})\right\}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\log f(Y_1;\boldsymbol{\theta})\right\}\right]$?

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