$$\hat{\theta}_N = \arg\max_{\theta \in \Theta \subseteq \mathbb{R}^L} E[\ell_i(y_i | x_i; \theta)]$$

Average loglikelihood

$$E_N[\cdot] = \sum_{i=1}^{N} [\cdot]$$

in this case loglikelihood function

estimator is implicit function of the sample size

- θ is an $L \times 1$ vector of unknown parameters

returns a scalar

$E_{\theta_0}[\ell_i(\theta_0; y_i)]$

↳ is the true loglikelihood function

$$= \int_S \ell_i(y_i | x_i; \theta) f(y_i | x_i; \theta_0)\, dy$$

Example

$\{x_i\}_{i=1}^{N}$ iid $\sim \exp(\lambda)$

scalar, i can draw on a cartesian axis

$$f_X(x_i, \lambda) = \lambda \exp(-\lambda x_i)$$

$$\downarrow$$

$$\ell_X(x_i, \lambda) = \log \lambda - \lambda x_i$$

loglikelihood for each individual observation

$$E_N[\ell_X(x_i, \lambda)] = \frac{1}{N} \sum [\log \lambda - \lambda x_i] = \log \lambda - \lambda \boxed{\frac{1}{N} \sum x_i}$$

$$E_{\theta_0}[x_i] = \frac{1}{\lambda_0}$$

$$= \log \lambda - \lambda \bar{x}$$

$$\hat{\theta} = \arg\max_{\lambda > 0} [\log \lambda - \lambda \bar{x}]$$

$$\frac{\partial E_N[x, \lambda]}{\partial \lambda} = \frac{1}{\lambda} - \bar{x} = 0$$

$$\hat{\lambda} = \frac{1}{\bar{x}} = \frac{N}{\sum x_i}$$

$$\mathbb{E}_{\theta_0}\left[\ell(x_i,\lambda)\right] = \log\lambda - \frac{\lambda}{\lambda_0}$$

does not depend on x anymore because we took expectation
(in reality we do not know this quantity)

$\downarrow$

but at least i can study if theoretically the model is sound

$$\underset{\lambda>0}{\operatorname{argmax}}\left[\log\lambda - \frac{\lambda}{\lambda_0}\right]$$

true pop. exp.
$\mathbb{E}_{\theta_0}[\theta;\theta_0]$
no more from the data

$$\frac{\partial\, \mathbb{E}_{\theta_0}\left[\ell(\lambda,\lambda_0)\right]}{\partial\lambda} = \frac{1}{\lambda} - \frac{1}{\lambda_0} = 0 \Rightarrow \boxed{\lambda = \lambda_0}$$

Exactly what we wanted

$\downarrow$

the solution is unique!

$$\mathbb{E}_N\left[\ell(y_i|x_i,\theta)\right] \xrightarrow{P} \mathbb{E}_{\theta_0}\left[\ell(y_i|x_i;\theta)\right]$$

(sample average) convergence pointwise, for each value of θ (population moment)

$\Downarrow$

$$\hat{\theta}_N \to \theta_0$$

$$\mathbb{E}_N\left[\ell_i(y_i;\theta)\right] \xrightarrow[\text{pointwise}]{?} \mathbb{E}_{\theta_0}\left[\ell_i(y_i;\theta)\right]$$

ASYMPTOTIC NORMALITY

$$\sqrt{N}\left(\hat{\theta}-\theta_0\right) \xrightarrow{d} \mathcal{N}_L\left(0;\ \underset{(L\times L)}{\Sigma_{\theta_0}}\right)$$

multidimensional

$$\Sigma_{\theta_0} = \left(-E_{\theta_0}\left[\frac{\partial^2 (\ell(\theta))}{\partial\theta\,\partial\theta'}\Big|_{\theta=\theta_0}\right]\right)^{-1}$$

inverse of the variance is sort of a measure of the precision

the more curve is the function in θ_0 the more we are precise with the estimation, the smaller is the curvature, the less the data are informative

less curvature, less precision, more variance and viceversa

$$E_N\left[\frac{\partial \ell(\theta)}{\partial\theta}\Big|_{\theta=\hat{\theta}}\right] = 0$$

by construction the empirical score evaluated in $\hat{\theta}$ is zero

EXPANSION OF THE SCORE AROUND θ_0

$$\underbrace{E_N\left[\frac{\partial \ell(\theta)}{\partial\theta}\Big|_{\theta=\theta_0}\right] + E_N\left[\frac{\partial^2 \ell(\theta)}{\partial\theta\,\partial\theta}\Big|_{\theta=\bar{\theta}}\right](\hat{\theta}-\theta_0)}_{\text{TAYLOR}} = 0$$

$\hat{\theta}_N < \bar{\theta} < \theta_0$

$f(x) = f(x_0) + f'(x_0)(x-x_0)$

this is to have a precise estimation and to not care about the other terms of the taylor polynomial

MEAN VALUE THEOREM

Now apply the CLT

$$\sqrt{N}\left(\hat{\theta}-\theta_0\right) = -E_N\left[\frac{\partial^2 \ell(\theta)}{\partial\theta\,\partial\theta}\Big|_{\theta=\bar{\theta}}\right]^{-1}\sqrt{N}\ E_N\left[\frac{\partial \ell(\theta)}{\partial\theta}\Big|_{\theta=\theta_0}\right]$$

$$\sqrt{N}\ \underset{L\times 1}{E_N\left[\frac{\partial \ell(\theta)}{\partial\theta}\Big|_{\theta=\theta_0}\right]} \xrightarrow{d} N_L\left(0;\ -E_{\theta_0}\left[\frac{\partial^2 \ell(\theta)}{\partial\theta\,\partial\theta'}\Big|_{\theta=\theta_0}\right]\right)$$

WHY THE MEAN OF THIS IS ZERO?

$$E_{\theta_0}\left(\frac{\partial \ell(\theta)}{\partial\theta}\Big|_{\theta=\theta_0}\right)$$

$$\text{Var}\left(\frac{\partial \ell(\theta)}{\partial\theta}\Big|_{\theta=\theta_0}\right) = E_{\theta_0}\left[\underbrace{\frac{\partial \ell(\theta)}{\partial\theta}\cdot\frac{\partial \ell(\theta)}{\partial\theta'}}_{\text{outer product}}\Big|_{\theta=\theta_0}\right]$$

$$= -E_{\theta_0}\Big[\underbrace{\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta'}}_{L\times L}\Big|_{\theta=\theta_0}\Big]$$ Fisher's information

• $$\operatorname*{plim}_{N\to\infty} E_N\left[\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta'}\Big|_{\theta=\bar\theta}\right] = E_{\theta_0}\left[\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta'}\Big|_{\theta=\bar\theta}\right]$$

true expected Hessian evaluated in $\bar\theta$

$$E_N[\cdot] \xrightarrow[\substack{\text{WLLN}\\ \text{or}\\ \text{ULLN}}]{p} E_{\theta_0}[\cdot]$$

if $\hat\theta \xrightarrow{p} \theta_0$, $\bar\theta = w\hat\theta + (1-w)\theta_0$

$w\ (0,1)$

$$\Downarrow$$

$$\bar\theta \longrightarrow \theta_0$$

we call the expected Hessian H and the expected gradient g

$$\sqrt{N}(\hat\theta-\theta_0) \xrightarrow{d} -H^{-1}\sqrt{N}\,g \longrightarrow \mathcal{N}(0; -H)$$

$$\sim \mathcal{N}(0; -H^{-1}\cancel{H}\cancel{H^{-1}})$$

Outer product of the gradient estimator to estimate the inverse of the Hessian

$$H = E_{\theta_0}\left[\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta'}\Big|_{\theta=\theta_0}\right]$$

$$g = E_{\theta_0}\left[\frac{\partial \ell(\theta)}{\partial\theta}\Big|_{\theta=\theta_0}\right]$$

$I \to$ Fisher information

measure of the information we have about a parameter

it is the variance of the score

$$I = Var\left(\frac{\partial l_i (y_i; \theta)}{\partial \theta} \Big|_{\theta = \theta_0} \right)$$

$$I = -E\left[\frac{\partial l_i^2 (y_i; \theta)}{\partial \theta \partial \theta'} \Big|_{\theta = \theta_0} \right]$$

measure of curvature of the function of which

the more the function is curve, less is the uncertainty about the parameter

if we have a sample of 100 obs. the loglikelihood for that sample will be a function of θ

we want to maximise it, we do the first derivative and we pose it $= 0$

if $k > 1$ this derivative is a vector (gradient) and the variance is a matrix (Hessian)