Modèles Linéaires Appliqués

Arthur Charpentier

Automne 2020

Rappels #3.1 (statistique & maximum de vraisemblance)

Likelihood / Vraisemblance

Assume that $\{x_1, x_2, \cdots, x_n\}$ are obtained from i.i.d. random variables $X_1, X_2, \cdots, X_n$, with identical distribution F_θ, and density f_θ.

$$\mathcal{L}(\theta) = f_\theta(\mathbf{x}) = f_\theta(x_1, \cdots, x_n) = \prod_{i=1}^{n} f_\theta(x_i)$$

$$\log \mathcal{L}(\theta) = \sum_{i=1}^{n} \log f_\theta(x_i)$$

$\widehat{\theta}$ is a maximum likelihood estimator of parameter θ if

$$\widehat{\theta} \in \text{argmax}\Big\{\mathcal{L}(\theta)\Big\} = \text{argmax}\Big\{\log \mathcal{L}(\theta)\Big\}$$

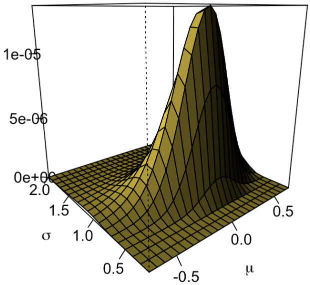
Likelihood / Vraisemblance

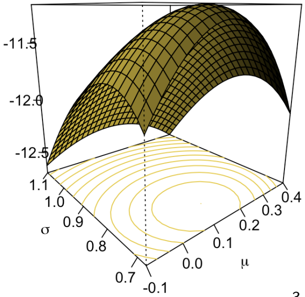
Given some sample $\{x_1, \cdots, x_n\}$ from a $\mathcal{N}(\mu, \sigma^2)$ distribution,

$$\mathcal{L}(\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2}\right)$$

$$\log \mathcal{L}(\mu, \sigma^2) = -\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^2} - \frac{n}{2} \log(2\pi\sigma^2)$$

Here $\boldsymbol{\theta} = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$.





Likelihood / Vraisemblance

The first order condition (also called likelihood equations) is

$$\left.\frac{\partial \log\left(\mathcal{L}(\theta; x_1, \cdots, x_n)\right)}{\partial \theta}\right|_{\theta=\widehat{\theta}} = 0$$

Second order condition is

$$\left.\frac{\partial^2 \log\left(\mathcal{L}(\theta; x_1, \cdots, x_n)\right)}{\partial \theta}\right|_{\theta=\widehat{\theta}} < 0$$

Example: if $X \sim \mathcal{P}(\lambda)$,

$$\log \mathcal{L}(\lambda; x_1, \cdots, x_n) = \sum_{i=1}^{n} \log\left(e^{-\lambda}\frac{\lambda^{x_i}}{x_i!}\right) = -n\lambda + n\overline{x}\log(\lambda) - \log\left(\prod_{i=1}^{n} x_i!\right)$$

$$\frac{\partial \log\left(\mathcal{L}(\lambda; x_1, \cdots, x_n)\right)}{\partial \lambda} = -n + \frac{n\overline{x}}{\lambda},\ \text{so } \widehat{\lambda} = \overline{x}.$$

Likelihood / Vraisemblance

$x \mapsto \dfrac{d}{d\theta}\log f_\theta(x)$ is called score. If $X \sim f_\theta$, $\mathbb{E}\left(\dfrac{d}{d\theta}\log f_\theta(X)\right) = 0$

Example: if $X \sim \mathcal{P}(\lambda)$,

$$\frac{d\log f_\lambda(X)}{d\lambda} = -1 + \frac{X}{\lambda},\ \text{so}\ \mathbb{E}\left(\frac{d}{d\lambda}\log f_\lambda(X)\right) = -1 + \frac{\mathbb{E}(X)}{\lambda} = 0$$

Fisher information associated with a density f_θ, with $\theta \in \mathbb{R}$ is

$$I(\theta) = \mathbb{E}\left(\frac{d}{d\theta}\log f_\theta(X)\right)^2 \text{ where } X \text{ has distribution } f_\theta,$$

$$I(\theta) = \text{Var}\left(\frac{d}{d\theta}\log f_\theta(X)\right) = -\mathbb{E}\left(\frac{d^2}{d\theta^2}\log f_\theta(X)\right).$$

For a sample of size n,

$$I_n(\theta) = \mathbb{E}\left(\frac{\partial}{\partial\theta}\log \mathcal{L}(\theta, X_1, \cdots, X_n)\right)^2 = nI(\theta)$$

Likelihood / Vraisemblance

Example: if X has a Poisson distribution $\mathcal{P}(\theta)$,

$$\log f_\theta(x) = -\theta + x \log\theta - \log(x!) \text{ and } \frac{d^2}{d\theta^2}\log f_\theta(x) = -\frac{x}{\theta^2}$$

$$I(\theta) = -\mathbb{E}\left(\frac{d^2}{d\theta^2}\log f_\theta(X)\right) = -\mathbb{E}\left(-\frac{X}{\theta^2}\right) = \frac{1}{\theta}$$

Example: if X has a binomial distribution $\mathcal{B}(n,\theta)$, $I(\theta) = \dfrac{n}{\theta(1-\theta)}$

Cramér-Rap bound: If $\widehat{\theta}$ is an unbiased estimator of θ, then

$$\text{Var}(\widehat{\theta}) \geq \frac{1}{nI(\theta)}$$

If that bound is attained, the estimator is said to be efficient.
An unbiased estimator $\widehat{\theta}$ is said to be optimal if it has the lowest variance among all unbiased estimators, see bias, minimum variance unbiased estimator

Likelihood / Vraisemblance

$$\text{if } \boldsymbol{\theta} \in \mathbb{R}^k,\ \left.\frac{\partial \log\left(\mathcal{L}(\boldsymbol{\theta}; x_1, \cdots, x_n)\right)}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = \mathbf{0}$$

Second order condition is

$$\text{if } \boldsymbol{\theta} \in \mathbb{R}^k,\ \left.\frac{\partial^2 \log\left(\mathcal{L}(\boldsymbol{\theta}; x_1, \cdots, x_n)\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} \text{ is definite negative}$$

If $\boldsymbol{\theta} \in \mathbb{R}^k$, then Fisher information is the $k \times k$ matrix $I = [I_{i,j}]$ with

$$I_{i,j} = \mathbb{E}\left(\frac{\partial}{\partial \theta_i} \log f_{\boldsymbol{\theta}}(X) \frac{\partial}{\partial \theta_j} \log f_{\boldsymbol{\theta}}(X)\right).$$

i.e.

$$I(\theta) = \mathbb{E}\left[\left(\frac{d}{d\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(X)\right)\left(\frac{d}{d\boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(X)\right)^{\top}\right]$$

$$I(\theta) = -\mathbb{E}\left(\frac{d^2}{d\boldsymbol{\theta} d\boldsymbol{\theta}^{\top}} \log f_{\boldsymbol{\theta}}(X)\right)$$

Likelihood / Vraisemblance

For a Gaussian distribution $\mathcal{N}(\theta,\sigma^2)$, $I(\theta)=\dfrac{1}{\sigma^2}$

For a Gaussian distribution $\mathcal{N}(\mu,\theta)$, $I(\theta)=\dfrac{1}{2\theta^2}$

For a Gaussian distribution $\mathcal{N}(\boldsymbol{\theta})$, $I(\boldsymbol{\theta})=\begin{pmatrix}1/\sigma^2 & 0\\ 0 & 2/\sigma^2\end{pmatrix}$

Cramér-Rao bound is $\dfrac{1}{n}I^{-1}=\dfrac{1}{n}\begin{pmatrix}\sigma^2 & 0\\ 0 & \sigma^2/2\end{pmatrix}$

Likelihood / Vraisemblance

Let $\{x_1, \cdots, x_n\}$ be a sample with distribution $f_{\boldsymbol{\theta}}$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. The maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_n$ of $\boldsymbol{\theta}$ is

$$\widehat{\boldsymbol{\theta}}_n \in \text{argmax}\Big\{\mathcal{L}(\boldsymbol{\theta}; x_1, \cdots, x_n), \boldsymbol{\theta} \in \boldsymbol{\Theta}\Big\}.$$

Under some technical assumptions $\widehat{\boldsymbol{\theta}}_n$ converges almost surely towards $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}_n \overset{a.s.}{\to} \boldsymbol{\theta}$, as $n \to \infty$.

Under some technical assumptions $\widehat{\boldsymbol{\theta}}_n$ is asymptotically efficient,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \overset{\mathcal{L}}{\to} \mathcal{N}(0, I^{-1}(\boldsymbol{\theta})).$$

See maximum likelihood estimation

Consider some Poisson model,

```
> set.seed(1)
> (y=rpois(10,3))
[1] 2 2 3 5 2 5 6 4 3 1
> mean(y)
[1] 3.3
> NLogL = function(lambda) -sum(log(dpois(y,lambda)))
> optim(fn = NLogL,par = 1)
$par
[1] 3.3

$value
[1] 18.59581
```

Likelihood / Vraisemblance

Consider a sample $\mathbf{X} = (X_1, \cdots, X_n)$ i.id. from $F_{\boldsymbol{\theta}}$. Let

$$S_{n,\boldsymbol{\theta}}(\mathbf{x}) = \frac{\partial \log \mathcal{L}(\boldsymbol{\theta}; \mathbf{x})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{n} S_{1,\boldsymbol{\theta}}(x_i)$$

denote the score function. Then $S_{n,\boldsymbol{\theta}}(\mathbf{X})$ is a random vector. Then

$$\mathbb{E}[S_{n,\boldsymbol{\theta}}(\mathbf{X})] = \mathbf{0}$$

while

$$\text{Var}[S_{n,\boldsymbol{\theta}}(\mathbf{X})] = I_n(\boldsymbol{\theta}) = \mathbb{E}\left(\frac{\partial}{\partial \boldsymbol{\theta}} S_{n,\boldsymbol{\theta}}(\mathbf{X})\right).$$

$$\frac{S_{n,\theta}(\mathbf{X})}{n} \overset{a.s.}{\to} 0 \quad \text{and} \quad \frac{S_{n,\boldsymbol{\theta}}(\mathbf{X})}{\sqrt{n}} \overset{\mathcal{L}}{\to} \mathcal{N}(0, I(\boldsymbol{\theta})).$$

Likelihood / Vraisemblance

If θ is univariate, use Taylor approximation of S_n in the neighbourhood of θ_0 (the true value)

$$S_n(x) = S_n(\theta_0) + (x - \theta_0)S_n'(y) \text{ for some } y \in [x, \theta_0]$$

Set $x = \widehat{\theta}_n$, then

$$S_n(\widehat{\theta}_n) = 0 = S_n(\theta_0) + (\widehat{\theta}_n - \theta_0)S_n'(y) \text{ for some } y \in [\theta_0, \widehat{\theta}_n]$$

Hence, $\widehat{\theta}_n = \theta_0 - \dfrac{S_n(\theta_0)}{S_n'(y)}$ for $y \in [\theta_0, \widehat{\theta}_n]$. Hence,

$$\widehat{\theta}_n^{(i+1)} = \widehat{\theta}_n^{(i)} - \frac{S_n(\widehat{\theta}_n^{(i)})}{S_n'(\widehat{\theta}_n^{(i)})},$$

from some starting value $\widehat{\theta}_n^{(0)}$ (hopefully well chosen).

Newton-Raphson:

$$\widehat{\theta}_n^{(i+1)} = \widehat{\theta}_n^{(i)} - \frac{S_n(\widehat{\theta}_n^{(i)})}{S_n'(\widehat{\theta}_n^{(i)})},$$

where

$$S_n'(\widehat{\theta}_n^{(i)}) \sim \frac{S_n(\widehat{\theta}_n^{(i)} + h) - S_n(\widehat{\theta}_n^{(i)} - h)}{2h}$$

from some starting value $\widehat{\theta}_n^{(0)}$ (hopefully well chosen), and some small $h > 0$.

Fisher-Scoring technique:

$$\widehat{\theta}_n^{(i+1)} = \widehat{\theta}_n^{(i)} + \frac{S_n(\widehat{\theta}_n^{(i)})}{nI(\widehat{\theta}_n^{(i)})},$$

again from some starting value.

Newton-Raphson

Consider some Poisson model, $S_1(\theta) = -1 + \dfrac{x}{\theta}$

```
> Sn = function(lambda) sum(-1+y/lambda)
> h = 1e-7
> dSn = function(lambda) (Sn(lambda+h)-Sn(lambda-h))
    /(2*h)
> L = rep(NA,10)
> L[1] = 1
> for(i in 1:9){
+   L[i+1] = L[i] - Sn(L[i])/dSn(L[i])
+ }
> L
[1] 1.000 1.697 2.521 3.116 3.290 3.300 3.300 3.300
```

Scoring-Fisher

Consider some Poisson model, with Fisher information $I(\theta) = \dfrac{1}{\theta}$

```
> I = function(lambda) 1/lambda
> L = rep(NA,10)
> L[1] = 1
> for(i in 1:9){
+   L[i+1] = L[i] - Sn(L[i])/(length(y)*I(L[i]))
+ }
> L
[1] 1.0 3.3 3.3 3.3 3.3 3.3 3.3 3.3 3.3 3.3
```