Maximum Likelihood Estimation

Let $Y_1, \ldots, Y_n$ be independent and identically distributed random variables.

Assume: Data are sampled from a distribution with density $f(y|\theta_0)$ for some (unknown but fixed) parameter θ_0 in a parameter space Θ.

Definition Given the data Y, the *likelihood function* $L_n(\theta|Y)$ is

$$L_n(\theta|Y) = f_Y(Y|\theta) = \prod_{i=1}^{n} f_{Y_i}(Y_i|\theta)$$

More generally, we may define $L_n(\theta|Y)$ as any function of $\theta \in \Theta$ proportional to $f_Y(Y|\theta)$.

Definition The *log-likelihood function* $l_n(\theta|Y)$ is the (natural) logarithm of the likelihood function $L_n(\theta|Y)$,

$$l_n(\theta|Y) = \log L_n(\theta|Y) = \sum_{i=1}^{n} \log f_{Y_i}(Y_i|\theta).$$

Example: For $Y_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu_0, \sigma_0^2)$, the likelihood function is

$$L_n(\mu, \sigma^2|Y) = f_Y(Y|\mu, \sigma^2) = \big(2\pi\sigma^2\big)^{-\frac{n}{2}} \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\big(Y_i - \mu\big)^2\Big)$$

and the log-likelihood function is (ignoring the additive constant)

$$l_n(\mu, \sigma^2|Y) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\big(Y_i - \mu\big)^2.$$

The parameter is $\theta = (\mu, \sigma^2)$ and the parameter space is $\Theta = \mathbb{R} \times \mathbb{R}^+$.

Maximum Likelihood Estimation

Definition A *maximum likelihood estimator (MLE)* $\hat{\theta}_{\mathrm{ML}}$ of θ maximizes the likelihood $L_n(\theta|Y)$, or equivalently, the log-likelihood $l_n(\theta|Y)$:

$$\hat{\theta}_{\mathrm{ML}} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, l_n(\theta|Y).$$

Assume: $L_n(\theta|Y)$ differentiable and bounded above (in θ)

$\rightsquigarrow$ solve the likelihood equation

$$S(\theta|Y) = \frac{\partial l_n(\theta|Y)}{\partial \theta} = 0.$$

($S(\theta|Y)$ is called *score function*)

Example: $Y_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu_0, \sigma_0^2)$

The log-likelihood function is:

$$l_n(\mu, \sigma^2|Y) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\big(Y_i - \mu\big)^2$$

Differentiation with respect to μ:

$$\frac{\partial l_n(\mu, \sigma^2|Y)}{\partial \mu} = 0 \;\Leftrightarrow\; \frac{n}{\sigma^2}\big(\bar{Y} - \mu\big) = 0 \;\Rightarrow\; \hat{\mu}_{\mathrm{ML}} = \bar{Y}$$

Differentiation with respect to σ^2:

$$\frac{\partial l_n(\mu, \sigma^2|Y)}{\partial \sigma^2} = 0 \;\Leftrightarrow\; -\frac{1}{\sigma^2} + \frac{1}{\sigma^4}\sum_{i=1}^{n}\big(Y_i - \mu\big)^2 = 0$$

$$\Rightarrow\; \hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \bar{Y}\big)^2$$

Maximum Likelihood Estimation

Large-sample Properties

For large n (and under certain regularity conditions), the MLE is approximately normally distributed:

$$\hat{\theta}_{\mathrm{ML}} - \theta_0 \approx \mathcal{N}(0, C)$$

Assume: Model is correctly specified (Y is sampled from density $f(\cdot|\theta_0)$).

Then the covariance matrix C is given by

$$C = I(\theta_0)^{-1}$$

where $I(\theta_0)$ is the *expected (Fisher) information (matrix)*

$$I(\theta) = \mathbb{E}\big(I(\theta|Y)|\theta\big) = \int I(\theta|y) f_Y(y|\theta)\, dy$$

and

$$I(\theta|Y) = -\frac{\partial^2 l_n(\theta|Y)}{\partial\theta^2}$$

is the *observed information (matrix).*

Example: $Y_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu_0, \sigma_0^2)$

$$I(\mu, \sigma^2|Y) = \begin{pmatrix} \frac{n}{\sigma^2} & \frac{n}{\sigma^4}(\bar{Y} - \mu) \\ \frac{n}{\sigma^4}(\bar{Y} - \mu) & -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6}\sum_{i=1}^n (Y_i - \mu)^2 \end{pmatrix}$$

Note that at $(\hat{\mu}, \hat{\sigma}^2)$, the observed Fisher information becomes

$$I(\hat{\mu}, \hat{\sigma}^2|Y) = \begin{pmatrix} \frac{n}{\hat{\sigma}^2} & 0 \\ 0 & \frac{n}{2\hat{\sigma}^4} \end{pmatrix}.$$

The expected information matrix is

$$I(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}.$$

Maximum Likelihood Estimation

Confidence interval for θ:

An approximate $(1-\alpha)$ confidence interval for θ_j is

$$\hat{\theta}_j \pm z_{\alpha/2}\sqrt{I(\hat{\theta}|Y)^{-1}_{jj}}$$

or

$$\hat{\theta}_j \pm z_{\alpha/2}\sqrt{I(\hat{\theta})^{-1}_{jj}}$$

Incorrect specified model

If the model is incorrectly specified and the data Y are sampled from a true density f^* then the ML estimate converges to the value θ^* which minimizes the *Kullback-Leibler information*

$$\mathbb{E}\left[\log\left(\frac{f(Y|\theta)}{f^*(Y)}\right)\right].$$

In this case, we have

$$\hat{\theta}_{\text{ML}} - \theta^* \approx \mathcal{N}(0, C^*)$$

where

$$C^* = I(\theta^*)^{-1}K(\theta^*)I(\theta^*)^{-1}$$

and

$$K(\theta) = \mathbb{E}\big(S(\theta|Y)S(\theta|Y)^{\mathsf{T}}\big).$$

In that case, the covariance matrix can be estimated by the estimator

$$\hat{C}^* = I(\hat{\theta}|Y)^{-1}\hat{K}(\hat{\theta})I(\hat{\theta}|Y)^{-1}.$$

where

$$\hat{K}(\theta) = S(\theta|Y)S(\theta|Y)^{\mathsf{T}}.$$

Newton-Raphson Method

Aim: Find $\hat{\theta}$ such that

$$S(\hat{\theta}|Y) = \frac{\partial l_n(\theta|Y)}{\partial \theta}\Big|_{\theta=\hat{\theta}} = 0.$$

Problem: Analytic solution of likelihood equations not always available.

Example: Randomly censored normal data

$$L_n(\theta|Y_{\text{obs}}, R) = \prod_{i=1}^{m} \frac{1}{\sigma}\varphi\left(\frac{Y_i - \mu}{\sigma}\right) \prod_{i=m+1}^{n} \left[1 - \Phi\left(\frac{c-\mu}{\sigma}\right)\right]$$

Example: Bivariate normal data, both variables subject to nonresponse

$$L_n(\theta|Y_{\text{obs}}) = \prod_{i=1}^{l} f_{Y_i}(Y_i|\mu, \Sigma) \prod_{i=l+1}^{m} f_{Y_{i1}}(Y_{i1}|\mu_1, \sigma_1^2) \prod_{i=m+1}^{n} f_{Y_{i2}}(Y_{i2}|\mu_2, \sigma_2^2)$$

Computational approach: Solve likelihood equation iteratively

Let $\theta^{(k)}$ be the current estimate. Taylor expansion of the score function about $\theta^{(k)}$ yields

$$S(\hat{\theta}|Y) \approx S(\theta^{(k)}|Y) - I(\theta^{(k)}|Y)(\hat{\theta} - \theta^{(k)})$$

Since $S(\hat{\theta}|Y) = 0$ ($\hat{\theta}$ maximizes $l_n(\theta|Y)$) we obtain

$$\hat{\theta} \approx \theta^{(k)} + I(\theta^{(k)}|Y)^{-1} S(\theta^{(k)}|Y).$$

This suggests the following iteration:

Newton-Raphson method:

$$\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} + I(\hat{\theta}^{(k)}|Y)^{-1} S(\hat{\theta}^{(k)}|Y)$$

Newton-Raphson Method

Example: Censored exponentially distributed observations

Suppose that $T_i \overset{\text{iid}}{\sim} \text{Exp}(\theta)$ and that the censored times

$$Y_i = \begin{cases} T_i & \text{if } T_i \leq C \\ C & \text{otherwise} \end{cases}$$

are observed. Let m be the number of uncensored observations. Then

$$l_n(\theta|Y) = m\log(\theta) - \theta \sum_{i=1}^{n} Y_i$$

with first and second derivative

$$\frac{\partial l_n(\theta|Y)}{\partial\theta} = \frac{m}{\theta} - \sum_{i=1}^{n} Y_i \qquad \text{and} \qquad \frac{\partial^2 l_n(\theta|Y)}{\partial\theta^2} = -\frac{m}{\theta^2}$$

Thus we obtain for the observed and expected information

$$I(\theta|Y) = I(\theta) = \frac{m}{\theta^2}.$$

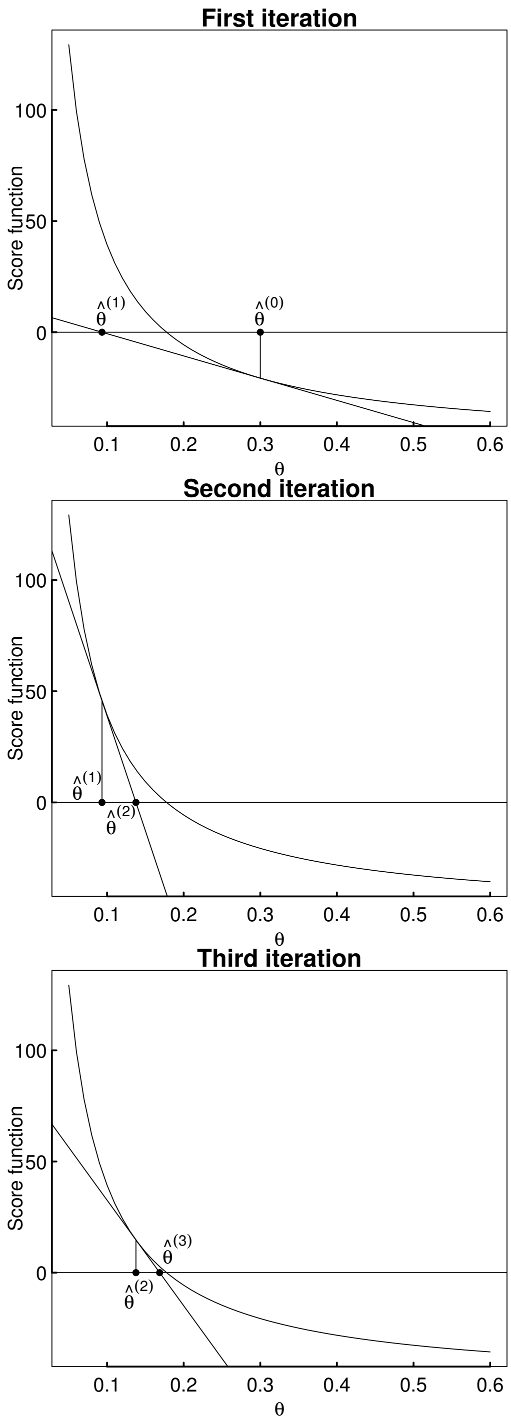
Thus the MLE can be obtained be the Newton-Raphson iteration

$$\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} + \frac{(\hat{\theta}^{(k)})^2}{m} \cdot \Big(\frac{m}{\hat{\theta}^{(k)}} - \sum_{i=1}^{n} Y_i\Big)$$

Numerical example: Choose starting value in $(0, 1)$

	Starting value		
Iteration k	**0.01**	**0.4**	**0.6**
1	0.0196	0.0764	-0.1307
2	0.0374	0.1264	-0.3386
3	0.0684	0.1805	-1.1947
4	0.1157	0.2137	-8.8546
5	0.1708	0.2209	-372.3034
6	0.2097	0.2211	-627630.4136
7	0.2205	0.2211	*
8	0.2211	0.2211	*
9	0.2211	0.2211	*
10	0.2211	0.2211	*

Newton-Raphson Method



Implementation in R:

```
#Log-likelihood, 1st & 2nd derivative
ln<-function(p,Y,R) {
  m<-sum(R==1)
  ln<-m*log(p)-p*sum(Y)
  attr(ln,"gradient")<-m/p-sum(Y)
  attr(ln,"hessian")<--m/p^2
  ln
}
#Newton-Raphson method
newmle<-function(p,ln,...) {
  l<-ln(p,...)
  pnew<-p-attr(l,"gradient")/attr(l,"hessian")
  pnew
}
#Simulate censored data~Exp(1/5)
Y<-rexp(10,1/5)
R<-ifelse(Y>10,0,1)
Y[R==0]=10
#Plot first derivative of the log-likelihood
x<-seq(0.05,0.6,0.01)
plot(x,attr(ln(x,Y,R),"gradient"),type="l",
  xlab=expression(theta),ylab="Score function")
abline(0,0)
#Apply Newton-Raphson iteration 3 times
#Starting value p=0.3
p<-0.3
p<-newmle(p,ln,Y=Y,R=R)
p
p<-newmle(p,ln,Y=Y,R=R)
p
p<-newmle(p,ln,Y=Y,R=R)
p
```

Newton-Raphson Method

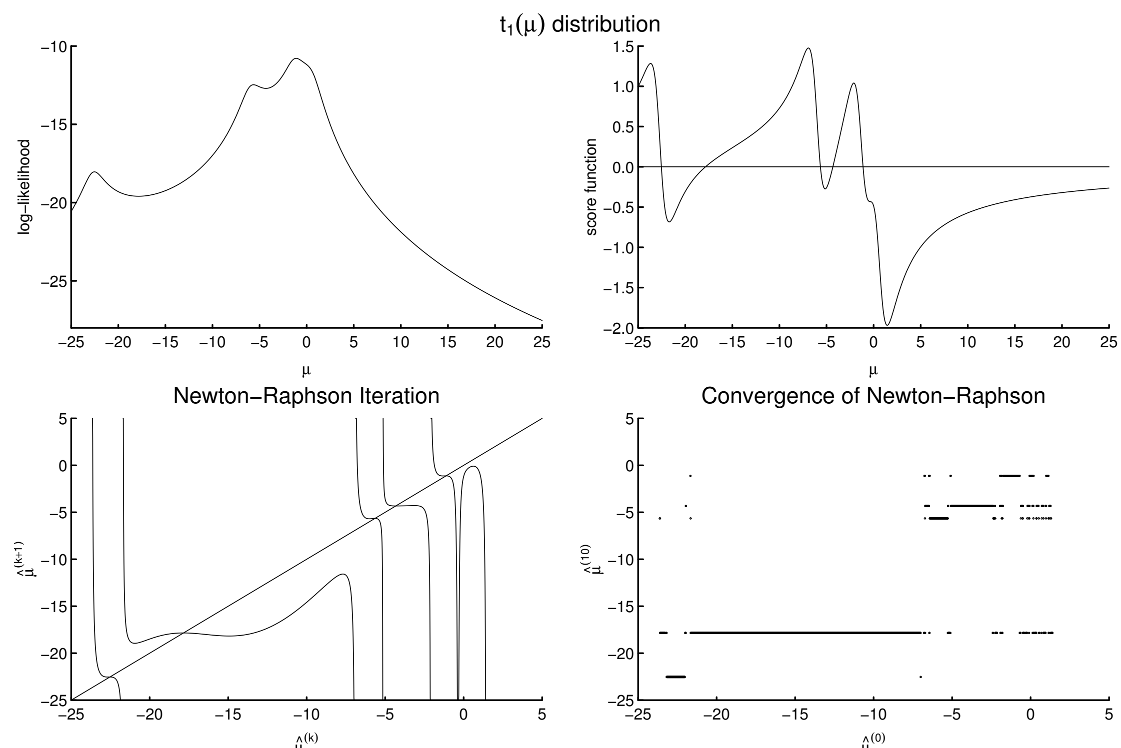
Example: t distribution
Suppose that $Y_1, \ldots, Y_n$ are independently sampled from the density

$$f_{Y_i}(y|\mu) = \frac{1}{\sqrt{\pi}\,\Gamma\big(\frac{1}{2}\big)} \big(1 + (y-\mu)^2\big)^{-1}$$

(t distribution with one degree of freedom and noncentrality parameter μ).
The log-likelihood function and its first and second derivative are given by

$$\begin{aligned} l_n(\mu|Y) &= -\sum_{i=1}^{n} \log\big(1 + (Y_i - \mu)^2\big) \\ \frac{\partial l_n(\mu|Y)}{\partial \mu} &= 2\sum_{i=1}^{n} (Y_i - \mu)\big(1 + (Y_i - \mu)^2\big)^{-1} \\ \frac{\partial^2 l_n(\mu|Y)}{\partial \mu^2} &= 2\sum_{i=1}^{n} \Big[2(Y_i - \mu)^2\big(1 + (Y_i - \mu)^2\big)^{-2} - \big(1 + (Y_i - \mu)^2\big)^{-1}\Big] \end{aligned}$$

Now suppose that $Y = (-1.318, 0.613, -6.004, -22.687)^{\mathsf{T}}$.



Alternative Methods

Quasi-Newton methods

Use iterative approximation

$$\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} - A^{-1} S(\hat{\theta}^{(k)}|Y),$$

where A is an approximation to the Hessian matrix $-I(\hat{\theta}^{(k)}|Y)$.

Modified Newton methods

- *Fisher's scoring method:*
  Replace observed information $I(\hat{\theta}^{(k)}|Y)$ by expected information

  $$I(\hat{\theta}^{(k)}) = \mathbb{E}\big(I(\hat{\theta}^{(k)}|Y)\big|\hat{\theta}^{(k)}\big)$$

- *Variant:* If the model is correctly specified

  $$I(\theta_0) = \mathrm{var}\big(S(\theta_0|Y)S(\theta_0|Y)^{\mathsf{T}}\big).$$

  For iid data, this suggests to approximate $I(\hat{\theta}^{(k)})$ by

  $$\sum_{i=1}^{n} S(\hat{\theta}^{(k)}|Y_i)S(\hat{\theta}^{(k)}|Y_i)^{\mathsf{T}} - \frac{1}{n} S(\hat{\theta}^{(k)}|Y)S(\hat{\theta}^{(k)}|Y)^{\mathsf{T}},$$

  where $S(\hat{\theta}^{(k)}|Y_i)$ is the score function based on a single observation.