Applied Multivariate and Longitudinal Data Analysis

Maximum Likelihood Estimation

Ana-Maria Staicu
SAS Hall 5220; 919-515-0644; astaicu@ncsu.edu

Parameter estimation: Maximum likelihood estimation (MLE)

Previously: we reviewed various distributions and noted that most of them depend on parameters.

In practice we collect DATA and make assumptions about the underlying distributions where they come from. Often these underlying distributions ar known up to a parameter, which could be scalar or vector. For example the data is assumed to come from $N(\mu, \sigma^2)$; in this case the unknown parameter is a vector $\theta = (\mu, \sigma^2)$. A major objective is then to use the data to estimate the parameter that fully determines the underlying distribution. In this section we review maximum likelihood estimation, in short MLE, as one approach for parameter estimation.

The maximum likelihood estimator is defined to be the maximizing value of a certain function called *likelihood function*; hence the name of the procedure: MLE. The likelihood function has a very wide use in statistical theory. An important principle in statistics essentially states that the *likelihood function contains all the information about an unknown parameter in the data.*

Recall the *likelihood function*: If we observe a random sample $y_1, y_2, \ldots, y_n$ from a distribution denoted by $f(y; \theta)$, where θ is the unknown parameter that is of interest. Then the likelihood function is

$$L(\theta) = f(y_1; \theta) \times f(y_2, ; \theta) \times \ldots \times f(y_n; \theta). \tag{1}$$

As the data are observed, the only unknown terms in the right hand side are those based on θ. The function $L(\theta)$ gives the probability of observing the (observed) data $y_1, \ldots, y_n$. The intuition behind MLE is to estimate the unknown parameter θ by the maximizer of this function; this value maximizes the probability of observing the sample we already observed. By an abuse of notation, the maximizer value of $L(\theta)$, the maximum likelihood estimator is also abbreviated in the literature by MLE. The context distinguishes if we talk about a procedure or an estimator or an estimate. Formally we write

$$\widehat{\theta}_{MLE} = \arg max_\theta L(\theta) \tag{2}$$

Most of the time, this maximization does not produce an analytical expression. In practice, because each of the values $f(y_i; \theta)$ are tiny, their product is even tinier and may result in numerical challenges. To avoid this numerical issue, it is more convenient to work with a different quantity, referred by the name of *log-likelihood function*, which is essentially the log on the natural base of the likelihood function, and is denoted by $\ell(\theta)$.

$$\ell(\theta) = \log L(\theta) = \log f(y_1; \theta) + \ldots + \log f(y_n; \theta) \tag{3}$$

Since *log* is a monotone increasing function, the maximizer of $L(\theta)$ also maximizes $\ell(\theta)$; and the latter one is much easier to determine and more stable numerically.

1 MLE for univariate Normal distribution

Suppose that $y_1, y_2, \ldots, y_n$ is an observed sample from $N(\mu, \sigma^2)$. Use the following steps to find the MLE of both μ and σ^2.

Recall that the density of $N(\mu, \sigma^2)$ is

$$f(y; \mu, \sigma^2) = 1/\sqrt{2\pi\sigma^2}\exp\{-(y-\mu)^2/2\sigma^2\}$$

- Write the likelihood function for the observed sample for μ, σ^2

- Write the log-likelihood function $\ell(\mu, \sigma^2)$

- Calculate the maximizers of the log-likelihood funtion among the solutions of

$$\frac{\partial \ell(\mu, \sigma^2)}{\partial \mu} = 0 \tag{4}$$

$$\frac{\partial \ell(\mu, \sigma^2)}{\partial \sigma^2} = 0 \tag{5}$$

Denote the solutions by $\widehat{\mu}_{MLE}$ and $\widehat{\sigma^2}_{MLE}$.

- Verify that $\widehat{\mu}_{MLE}$ and $\widehat{\sigma^2}_{MLE}$ are points of maximum of $\ell(\cdot, \cdot)$ by
  - Evaluating the hessian matrix (second derivative matrix) at the optimal point and showing it is negative semi-definite (this means that $\boldsymbol{a}^T H \boldsymbol{a} \leq 0$ for all 2-dimensional vectors $\boldsymbol{a}$)

$$H = \begin{pmatrix} \frac{\partial^2 \ell(\widehat{\mu}_{MLE}, \widehat{\sigma^2}_{MLE})}{\partial \mu^2} & \frac{\partial^2 \ell(\widehat{\mu}_{MLE}, \widehat{\sigma^2}_{MLE})}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 \ell(\widehat{\mu}_{MLE}, \widehat{\sigma^2}_{MLE})}{\partial \mu \partial \sigma^2} & \frac{\partial^2 \ell(\widehat{\mu}_{MLE}, \widehat{\sigma^2}_{MLE})}{\partial (\sigma^2)^2} \end{pmatrix} \tag{6}$$

  - Showing that $\ell(\widehat{\mu}_{MLE}, \widehat{\sigma^2}_{MLE}) \geq \ell(\mu, \sigma^2)$ for any values of $\mu \in R$ and $\sigma^2 > 0$.

Remark The MLE of the model parameters can be obtained relatively simply in `R`. For example:

```
> y=rnorm(25, mean=5, sd=3) # DATA
> llik= function(parm, y=y) {  # log-likelihood fn
+         mean.parm = parm[1]
+         sd.parm = sqrt(parm[2])
+         sum( dnorm(y, mean=mean.parm,sd=sd.parm , log=TRUE )) }
>
>
> llik(c(5, 9), y=y)
[1] -71.68232
```

This function evaluates the log-likelihood at any value of the parameters. R has several functions for optimization. One widely used function is nlm, which does minimization of a function of one or multiple parameters.

```
> nllik= function(parm, y=y) { # negative log-likelihood fn
+         mean.parm = parm[1]
+         sd.parm = sqrt(parm[2])
+         -sum( dnorm(y, mean=mean.parm,sd=sd.parm , log=TRUE ))}
>
> nlm(nllik, c(10, 15), y=y)
$minimum
[1] 69.38644

$estimate
[1]  4.530586 15.074981

$gradient
[1]  5.435812e-06 -1.030347e-06

$code
[1] 1

$iterations
[1] 8
```

Summary: For an observed sample $y_1, y_2, \ldots, y_n$ from $N(\mu, \sigma^2)$ the MLEs are:

$$\widehat{\mu} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{7}$$

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{8}$$

2 MLE for multivariate Normal distribution

Suppose the data $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are a random sample from the multivariate normal distribution $N_p(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu}$ is unknown p-dimensional parameter and Σ is unknown $p \times p$ matrix.

Then, using the density of the multivariate normal, $\boldsymbol{f(y; \mu, \Sigma) = (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\{-(y-\mu)^T \Sigma^{-1}(y-\mu)/2\}}$ and following the same logic as earlier we obtain that the MLE for μ and Σ are

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i \tag{9}$$

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{y}_i - \widehat{\mu})(\boldsymbol{y}_i - \widehat{\mu})^T. \tag{10}$$

The MLE estimators enjoy very nice properties. However as we will see later on, the MLE of the (co)variance parameter is biased and that a correction is needed to obtain an unbiased estimator. The unbiased (co)variance parameter uses $(n-p)$ in place of n in the expression (10), where p is the dimension of the mean vector. In the case of univariate normal, $p = 1$.