TCP Transl Clin Pharmacol
2015;23(1):08-14
http://dx.doi.org/10.12793/tcp.2015.23.1.8

TUTORIAL

Statistical basis for pharmacometrics: maximum likelihood estimator and its asymptotics

Kyungmee Choi[1,+] and Dong-Seok Yim[2,*]

[1]Division of Mathematics, College of Science and Technology, Hongik University at Sejong, Jochiwon, Sejong 339-701, South Korea, [2]Department of Clinical Pharmacology & Therapeutics, Seoul St. Mary's Hospital, PIPET (Pharmacometrics Institute for Practical Education & Training), College of Medicine, The Catholic University of Korea, Seoul 137-701, South Korea

*Correspondence: D. S. Yim; Tel: +82-2-2258-7327, Fax: +82-2-536-2485, E-mail: yimds@catholic.ac.kr

[+]This work was supported by the 2015 Hongik University Academic Research Support Fund.

Received 10 May 2015
Revised 5 Jun 2015
Accepted 10 Jun 2015

Keywords
Taylor series, Newton-Raphson method, Maximum Likelihood Estimator

pISSN: 2289-0882
eISSN: 2383-5427

The maximum likelihood estimator is the point estimator of the top priority in statistical data analysis because of its optimum properties for large sample size. While the maximum likelihood estimator is widely used, it has been an abstruse subject for pharmacometricians without statitics bagkround because of high dimensional calculus and asymptotic theories. This tutorial provides a general and brief introduction to the maximum likelihood estimator and its related caluculus for non-statisticians.

Introduction

The maximum likelihood estimator (MLE) is a "strongly consistent, asymptotically normal, and asymptotically efficient" point estimator of the parameter because of its optimum properties.[1] In other words, for large sample size, it is unbiased, normally distributed, and its variance is the Cramer-Rao lower bound for the variance of unbiased estimators of the parameter.[2] Because a linear combination of normal random variables is again normally distributed,[3] the asymptotic normality and its optimum properties are great advantage of the MLE. The model of our interest is, however, often nonlinear function of multiple PK and/or PD parameters which requires linear approximation of the MLE.[4] The high dimensional calculus used in the approximation steps has been an obstacle for non-statisticians to overcome. This tutorial aims to provide a general introduction to MLE along with a review of its associated caluculus. See p 144 of reference,[1] p 318 of reference,[2] and p 157 of reference.[3]

Calculus plays a vital role in statistics. Among massive amount of caluculus, we review the Taylor series to get approximate polynomials of nonlinear functions, the Newton-Raphson method to get approximate solutions of equations and the eigenvalue and eigenvector problem to understand correlation of random variables. We will also review the likelihood function and the MLE. The Newton-Raphson method will be applied to get the approximate MLE from the nonlinear likelihood function. Further asymptotic properties of the MLE will be reviewed. We assume that readers are familiar with differentiation and introductory probability.

Review of calculus

Taylor series

Polynomials appear in wide application areas because of their simple additive form. Functions of our interest, however, is not often polynomial and scientists and engineers want to transform their nonlinear functions into polynomials. For this purpose, the Taylor series provides easy approximate polynomials around a non-singular point.

Denition 1 *If f(x) is defined and infinitely differentiable at x = c, then f(x) can be expressed as a power series of the form*

$$f(x) = f(c) + f'(x)(x-c) + \frac{f''(c)}{2!}(x-c)^2 + \cdots + \frac{f^{(n)}(c)}{n!}(x-c)^n + \cdots,$$

which is called a Taylor series.

In order to prove it, let us first express $f(x)$ as a power series as follows.

$$f(x) = a_0 + a_1(x-c) + a_2(x-c)^2 + a_3(x-c)^3 + a_4(x-c)^4 + \cdots$$

Then,

$$\begin{aligned} f'(x) &= a_1 + 2a_2(x-c) + 3a_3(x-c)^2 + 4a_4(x-c)^3 + \cdots \\ f''(x) &= 2a_2 + 2\cdot 3a_3(x-c) + 3\cdot 4a_4(x-c)^2 + \cdots \\ f'''(x) &= 2\cdot 3a_3 + 2\cdot 3\cdot 4a_4(x-c) + \cdots \end{aligned}$$

$$f(c) = a_0, \quad f'(c) = a_1, \quad f''(c) = 2a_2, \quad f'''(c) = 3!a_3,$$

Therefore,

$$f^{(n)}(c) = n!a_n, \quad a_n = \frac{1}{n!}f^{(n)}(c)$$

The partial sums of the Taylor series, the Taylor polynomials, provide easy linear approximations to the whole function $f(x)$.

Example 1 *Let us find a Taylor series of* $f(x) = e^x$ *at* $x = 0$.

Since $f'(x) = f''(x) = \cdots = f^{(n)}(x) = e^x$, $f(0) = f'(0) = f''(0) = f^{(n)}(0) = 1$, and thus the coefficient $a_n = \frac{1}{n!}$. We consequently have the famous Taylor series of the exponential function $f(x) = e^x$ as follows:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots + \frac{x^n}{n!} + \cdots$$

The number e can be approximated by the partial sum of the first four terms of the Taylor series. That is,

$$e \approx 1 + 1.0 + \frac{1}{2} + \frac{1}{3!} + \frac{1}{4!} = 2.708333$$

with an error of 0.009948.

The Taylor polynomials can be easily extended to a bivariate function $f(x, y)$ and its second order Taylor polynomial is as follows:

$$\begin{aligned} f(x,y) &= f(x_0,y_0) + f_x(x_0,y_0)(x-x_0) + f_y(x_0,y_0)(y-y_0) \\ &\quad + \frac{1}{2}f_{xx}(x_0,y_0)(x-x_0)^2 + f_{xy}(x_0,y_0)(x-x_0)(y-y_0) \\ &\quad + \frac{1}{2}f_{yy}(x_0,y_0)(y-y_0)^2 + error \end{aligned}$$

where f_x is partial derivative of f with respect to x, f_y is partial derivative of f with respect to y, f_{xx} is partial derivative of f_x with respect to x, and f_{yy} is partial derivative of f_y with respect to y, f_{xy} is partial derivative of f_x with respect to y, and f_{yx} is partial derivative of f_y with respect to x.

Let us now present the first order Taylor polynomial for two bivariate functions $f(x, y)$ and $g(x, y)$.

$$f(x,y) \approx f(x_0,y_0) + f_x(x_0,y_0)(x-x_0) + f_y(x_0,y_0)(y-y_0)$$

$$g(x,y) \approx g(x_0,y_0) + g_x(x_0,y_0)(x-x_0) + g_y(x_0,y_0)(y-y_0)$$

By rewriting them in a vector form, we have the the first order Taylor polynomial for a vector function as follows:

$$\begin{pmatrix} f(x,y) \\ g(x,y) \end{pmatrix} \approx \begin{pmatrix} f(x_0,y_0) \\ g(x_0,y_0) \end{pmatrix} + \begin{pmatrix} f_x(x_0,y_0) & f_y(x_0,y_0) \\ g_x(x_0,y_0) & g_y(x_0,y_0) \end{pmatrix} \begin{pmatrix} x-x_0 \\ y-y_0 \end{pmatrix}$$

Introducing the gradient of $f(x, y)$

$$\nabla f = grad\, f = (f_x, f_y)$$

and Jacobian J

$$J(f,g) = \begin{pmatrix} f_x & f_y \\ g_x & g_y \end{pmatrix}$$

let us define the operater H as follows

$$\mathrm{H}f = J(\nabla f) = J(f_x, f_y) = \begin{pmatrix} f_{xx} & f_{yx} \\ f_{xy} & f_{yy} \end{pmatrix}$$

The bivariate function $f(x, y)$ represents a surface in R^3 like a mound in space and its gradient ∇f points to the direction of steepest change. This notation will be further used to obtain an approximate solution of equations expressed in a vector function.

Newton-Raphson method

The Newton-Raphson method is used to look for an approximate root of a real-valued function $f(x)$ by iteratively solving the equation $f(x) = 0$.

Let us start with an initial point x_0, and the tangent line L_1 to $f(x)$ at x_0. We want to find the equation of the line L_1. L_1 passes through $(x_0, f(x_0))$ and its slope is $f'(x_0)$. Solving the equation for y, we have a line equation

$$L_1 : y = f(x_0) + f'(x_0)(x - x_0)$$

Note that it is the first order Taylor polynomial of $f(x)$. The x-intercept of L_1 is obtained by solving the equation for x

$$0 = f(x_0) + f'(x_0)(x - x_0)$$

and more explicitly it is

$$x = x_0 - \frac{f(x_0)}{f'(x_0)}$$

We name the x-intercept as x_1, a new approximate root. Then,

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

For further iterations, let us define a sequence of approximate roots

TCP
Transl Clin Pharmacol

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

Example 2 *Find approximate roots of* $x^2 = 2$ *using the Newton-Raphson method.*
Let us first define the equation $f(x)$ such as

$$f(x) = x^2 - 2$$

Then the derivative of $f(x)$ is given by

$$f'(x) = 2x$$
$$x_{n+1} = x_n - \frac{{x_n}^2 - 2}{2x_n} = \frac{1}{2}x_n + \frac{1}{x_n}$$

As an exercise, let us start from $x_0 = 2$ and have the first two approximate roots x_1 and x_2. Then,

$$\begin{aligned} x_1 &= \frac{1}{2}x_0 + \frac{1}{x_0} = \frac{1}{2} \times 2 + \frac{1}{2} = \frac{3}{2} = 1.5 \\ x_2 &= \frac{1}{2}x_1 + \frac{1}{x_1} = \frac{1}{2} \times \frac{3}{2} + \frac{2}{3} = 1.4167 \end{aligned}$$

The sequence converges very quickly to $\sqrt{2}$, which we know as 1.414214.

Eigenvalue and eigenvector

For notational convenience, we use boldface for vectors from now on.

Definition 2 *For a square matrix A, a scalar* λ *is called an eigenvalue of A if the relationship*

$$A\,\mathbf{x} = \lambda\mathbf{x}$$

holds for a nonzero vector $\mathbf{x}$*, which is called an eigenvector of A.*

In order to look for λ and nonzero $\mathbf{x}$, we solve the linear equations for $\mathbf{x}$. Since $I\mathbf{x} = \mathbf{x}$ for an identity matrix I, we have the same linear equations such as

$$A\mathbf{x} - \lambda I\mathbf{x} = 0$$

and

$$(A - \lambda I)\mathbf{x} = 0$$

If $|A - \lambda I| \neq 0$, then $(A - \lambda I)^{-1}$ exists and $\mathbf{x} = 0$ which is a trivial solution and of no use. Equivalently, for $\mathbf{x} \neq 0$, we need $|A - \lambda I| = 0$. Two steps are suggested to calculate λ and nonzero $\mathbf{x}$ for a given matrix A.

*STEP*1 : Solve $|A - \lambda I| = 0$ for λ.
*STEP*2 : Get nonzero $\mathbf{x}$ satisfying $A\mathbf{x} = \lambda\mathbf{x}$.

Example 3 *Find eigenvalues and eigenvectors of* $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$

Since A is 2×2, there can be two eigenvalues.

$$|A - \lambda I| = \begin{vmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = \lambda^2 - 4\lambda + 3 = 0$$

By solving a quadric equation, we get two eigenvectors $\lambda_1 = 3$ or $\lambda_2 = 1$.

Now for each eigenvalue, let us get a corresponding eigenvector. For $\lambda_1 = 3$ and $\mathbf{x} = (x_1, x_2)^T \in R^2$, solve

$$(A - 3I)\mathbf{x} = 0$$

Equivalently,

$$\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

and from the following two equations

$$\begin{aligned} -x_1 + x_2 &= 0 \\ x_1 - x_2 &= 0 \end{aligned}$$

we have

$$x_1 = x_2$$

Therefore, the eigenvector we are looking for can be one of the following vectors

$$x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \cdots$$

Note that there are infinitely many eigenvectors corresponding to an eigenvalue and all of them are on the same line $x_2 = x_1$. The eigenvector is not unique.

Similarly, for $\lambda_2 = 1$, solve $(A - I)\mathbf{x} = 0$. Then we have

$$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

and

$$\begin{aligned} x_1 + x_2 &= 0 \\ x_1 + x_2 &= 0 \end{aligned}$$

Therefore,

$$x_1 = -x_2$$

and

$$x = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \cdots$$

In summary, the eigenvalues of A are 3 and 1, and its corresponding eigenvectors of unit length are $\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$ and $\left(-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$.

Eigenvalues and eigenvectors have very useful properties which determine the characteristics of A. Suppose that $A\mathbf{x}_i = \lambda_i\mathbf{x}_i$ for $i = 1, \cdots, n$.

Theorem 1 *The determinant of A is the product of eigenvalues. In other words,*

$$|A| = \prod_{i=1}^{n} \lambda_i$$

If $\lambda_i = 0$ for any i, then $|A| = 0$ and A^{-1} does not exist. This property is important because we often need to know whether or not there exists the inverse matrix and zero eigenvalue is an indicator for singularity of the given matrix.

Theorem 2 (Spectral Decomposition Theorem) *If A is symmetric, then its eigenvalues are real and A can be decomposed as follows:*

$$A = \lambda_1 \mathbf{x}_1 \mathbf{x}_1^T + \lambda_2 \mathbf{x}_2 \mathbf{x}_2^T + \cdots + \lambda_n \mathbf{x}_n \mathbf{x}_n^T$$

Example 4 *(Continued) Let us find the spectral decomposition of* $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$.

$$\begin{aligned} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} &= 3 \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} + 1 \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \\ &= 3 \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} + \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \end{aligned}$$

Definition 3 $q(\mathrm{x}) = \mathbf{x}^T A \mathbf{x}$ *is called a quadratic form.*

In R^2,

$$\begin{aligned} q(\mathbf{x}) &= \mathbf{x}^T A \mathbf{x} = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= a_{11} x_1^2 + (a_{12} + a_{21}) x + a_{22} x_2^2 \end{aligned}$$

Definition 4 *A is positive definite (pd) if $\mathbf{x}^T A \mathbf{x} > 0$ for all $\mathbf{x} \in R^n$. A is positive semidefinite (psd) if $\mathbf{x}^T A \mathbf{x} \geq 0$ for all $\mathbf{x} \in R^n$. A is negative definite (nd) $\mathbf{x}^T A \mathbf{x} < 0$ for all $\mathbf{x} \in R^n$. A is negative semidefinite (nsd) if $\mathbf{x}^T A \mathbf{x} \leq 0$.*

Theorem 3 *A is pd if all $\lambda > 0$ and A is psd if all λ's ≥ 0. A is nd if all $\lambda < 0$ and A is nsd if all λ's ≤ 0.*

If A is either pd or nd, then its inverse exists. Eigenvalues of the inverse matrix are the reciprocals of the eigenvalues of A. If A is pd, then A^{-1} is also pd.

Let us now think about the meaning of eigenvalues and eigenvectors in statistics. See p 153 of reference.[3] Let $\mathbf{x} = (x_1, x_2, \cdots, x_p)$ be a vector of random variables which has a multivariate normal distribution with the mean μ and the covariance Σ. The contour of a constant density is then an oblique ellipsoid

$$(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c^2$$

which is centered at μ and has axes $\pm c\sqrt{\lambda_i}\mathbf{e}_i$, where $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$.

Imagine a contour of an ellipsoidal bell which is rotated on the xy-plane. For $\mathbf{x} \in R^2$, the oblique ellipsoid has two axes e_1 and e_2 which are not parallel to the xy-axes and whose magnitudes are proportional to $\sqrt{\lambda_1}$ and $\sqrt{\lambda_2}$, respectively. First of all, each λ is nonnegative since it is length. In other words, each λ is nonnegative variation in the data to the direction of the corresponding eigenvector. Thus, Σ is psd. Secondly, if $\lambda_2 = 0$, then the ellipsoid becomes a line and a perfect collinearity between x_1 and x_2 emerges to the direction of e_1. In summary, if there exists a zero eigenvalue, then Σ is singular, that is, not invertible, an ellipsoid becomes a line, and therefore there exists high correlation among variables.

For the sample covariance matrix S, the followings are equivalent.

1. S is psd, but not pd.
2. There exist zero eigenvalues.
3. $|S| = 0$.
4. S is singular.
5. $X_1, X_2, \cdots, X_p$ are highly correlated.

Maximum Likelihood Estimator and its optimum properties

The maximum likelihood estimator is a good estimator of the parameter θ because it is asymptotically normally distributed with mean θ and variance which is the Cramer-Rao lower bound for the variance of unbiased estimators of θ. We use the word of "asymptotic" for large sample size. Let us start with the definitions of the likelihood function and the maximum likelihood estimator. Read p 318 of reference[2] for the Cramer-Rao lower bound and p 359 of reference[2] for the optimum properties of the maximum likelihood of estimator.

Definition 5 *The likelihood function of n random variables $X_1, X_2, \cdots, X_n$ is defined to be their joint probability density function,*

$$L(\theta; x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} p(x_i; \theta)$$

where $p(x_i;\theta)$ is the pdf of X_i and $\theta = (\theta_1, \theta_2, \cdots, \theta_p)$ is the unknown parameter vector.

The likelihood function $\mathrm{L}(\theta)$ is considered as the function of the parameter θ for the fixed data $x_1, x_2, \cdots, x_n$.

Definition 6 *Maximum likelihood estimator (MLE) of θ is defined to be*

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\theta)$$

We want to know which particular value of the random variables is most likely to occur or equivalently want to know where the likelihood becomes the maximum. In other words, we are looking for the value of θ which makes the likelihood function the greatest.

In order to look for $\hat{\theta}_{MLE}$, we solve $\partial l / \partial \theta = 0$.

Example 5 *Let $X_1, X_2, \cdots, X_n$ be a random sample from $N(\theta_1, \theta_2)$,*

where $-\infty < \theta_1 < \infty$, $\theta_2 > 0$. *Note that* $\theta_1 = \mu$, $\theta_2 = \sigma^2$. *Find the MLEs of* μ *and* σ^2.

Let us start from the likelihood function by producting the pdf's such as

$$\begin{aligned} L(\theta_1, \theta_2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta_2}} \exp\left(-\frac{(x_i-\theta_1)^2}{2\theta_2}\right) \\ &= (2\pi)^{-\frac{n}{2}} \theta_2^{-\frac{n}{2}} \exp\left(-\frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i-\theta_1)^2\right) \end{aligned}$$

Since $\arg\max_{\theta_1,\theta_2} L(\theta_1,\theta_2) = \arg\max_{\theta_1,\theta_2} \ln L(\theta_1,\theta_2)$, we define

$$l(\theta_1,\theta_2) = \ln L(\theta_1,\theta_2) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\theta_2 - \frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i-\theta_1)^2$$

and maximize $l(\theta_1, \theta_2)$ instead of $L(\theta_1, \theta_2)$. This is easier because terms in $l(\theta_1, \theta_2)$ are additive.

$$\begin{aligned} \frac{\partial l}{\partial \theta_1} &= -\frac{2}{2\theta_2}\sum_{i=1}^{n}(x_i-\theta_1)(-1) = 0 \\ \frac{\partial l}{\partial \theta_2} &= -\frac{n}{2}\frac{1}{\theta_2} + \frac{1}{2\theta_2^2}\sum_{i=1}^{n}(x_i-\theta_1)^2 = 0 \end{aligned}$$

Solving the two equations simultaneously, we have

$$\hat{\theta}_{1,MLE} = \bar{X}$$

$$\hat{\theta}_{2,MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\theta}_{1,MLE})^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

Notice that $\hat{\theta}_{MLE}$'s are obtained in closed forms. Let us calculate the expectation of $\hat{\theta}_{MLE}$ to check their unbiasedness.

$$\begin{aligned} E[\hat{\theta}_{1,MLE}] &= E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n}X_i\right] = \frac{1}{n}\sum_{i=1}^{n}E[X_i] = \mu \\ E[\hat{\theta}_{2,MLE}] &= E\left[\frac{1}{n}\sum_{i=1}^{n}(X-\bar{X})^2\right] = \frac{n-1}{n}E[S^2] = \frac{n-1}{n}\sigma^2 \end{aligned}$$

where

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

Note that $\hat{\theta}_{1,MLE}$ is an unbiased estimator of μ. The $\hat{\theta}_{2,MLE}$ is not an unbiased estimator of σ^2 and it underestimates σ^2.

For large sample size, the variance of MLE is the Cramer-Rao lower bound which is the minimum among the unbiased estimators. It is closely associated with the Fisher's Information that is described as the amount of information X carries about. Let us first define the Fisher Information. For notational convenience, we use $E_\theta[\cdot] = E[\cdot|\theta]$.

Definition 7 *The Fisher Information is*

$$I(\theta) = E_\theta\left[\left(\frac{\partial}{\partial\theta}\log p(X;\theta)\right)^2\right]$$

Note that $\log p(X;\theta)$ is a special form of log-likelihood function $l(\theta)$ when $n = 1$. $l(\theta)$ is the function of θ which is not a random variable, where X is fixed. The $\log p(X;\theta)$ is the function of X which is a random variable, where θ is given. It is useful to know that

$$E_\theta\left[\frac{\partial}{\partial\theta}\log p(X;\theta)\right] = 0$$

Here is the proof.

$$\begin{aligned} E_\theta\left[\frac{\partial}{\partial\theta}\log p(X;\theta)\right] &= E_\theta\left[\frac{\frac{\partial p(X;\theta)}{\partial\theta}}{p(X;\theta)}\right] = \int \frac{\frac{\partial p(X;\theta)}{\partial\theta}}{p(X;\theta)} p(X;\theta)dX \\ &= \frac{\partial}{\partial\theta}\int p(X;\theta)dX = \frac{\partial}{\partial\theta}(1) = 0 \end{aligned}$$

where the integral of pdf is 1. From this property, we can conjecture that

$$E[\partial l(\theta)/\partial\theta] = 0$$

and derive another efficient form of $I(\theta)$ such as

$$I(\theta) = -E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log p(X;\theta)\right]$$

See p 320 of reference.[2]

Theorem 4 *The maximum likelihood estimator of* θ, $\hat{\theta}_{MLE}$, *is asymptotically normally distributed with mean and variance*

$$\mathrm{Var}(\hat{\theta}_{MLE}) = \frac{1}{n}I^{-1}(\theta)$$

Note that $\mathrm{Var}(\hat{\theta}_{MLE})$ is the Cramer-Rao lower bound for the variance of the unbiased estimators of θ. Note that for $I(\theta)$ to be invertible, it should be pd. In other words, if its eigenvalues are all positive, $I(\theta)$ is invertible. Especially for two parameters θ_1 and θ_2, $I(\theta)$ is given by

$$I(\theta) = -E\begin{bmatrix} \frac{\partial^2 p(X;\theta)}{\partial\theta_1^2} & \frac{\partial^2 p(X;\theta)}{\partial\theta_2\partial\theta_1} \\ \frac{\partial^2 p(X;\theta)}{\partial\theta_2\partial\theta_1} & \frac{\partial^2 p(X;\theta)}{\partial\theta_2^2} \end{bmatrix}$$

where $I(\theta) = -E_\theta[Hp]$.

Besides, asymptotic properties of the MLE holds for the case of multidimensional $\theta = (\theta_1, \cdots, \theta_p)$. The joint distribution of the maximum likelihood estimators is asymptotically multivariate normal. This is very powerful property because any combination of normal random variables is again normally distributed and it can be used to test the parameters.

Note that $\hat{\theta}_{2,MLE}$ in Example 5 is biased and underestimates σ^2. In order to get its unbiased estimator, the Residual (Restricted) Maximum Likelihood Estimation (RMLE) are used. See pp 62-76 of reference.[4]

Approximate MLE in a vector form

Suppose that the model of our interest is in general

$$Y_i = f(\mathbf{X}_i;\theta) + \epsilon_i$$

where Y_i is a response variable, $\mathbf{X}_i$ is a vector of explanatory variables in R^p, $\theta = (\theta_1, \theta_2, \cdots, \theta_p)$, and ϵ_i is multivariate normal with mean 0 and variance $\sigma^2 I$.
Note that

$$E[Y_i] = f(\mathbf{X}_i;\theta)$$

Then, the log-likelihood function is defined as follows:

$$l(\theta,\sigma^2) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - f(\mathbf{X}_i;\theta))^2$$

Unlike Example 5, it is hard to solve $\partial l/\partial\theta = 0$ to get the MLE in a closed form if $E[Y_i] = f(\mathbf{X}_i;\theta)$ is nonlinear. Then, using Newton-Raphson method, we get an approximate MLE of θ and its corrected variance as follows:
1. Get $\theta^{(i)}$ by solving $\nabla l = 0$ at the i^{th} step.
2. Get MLE as the limit of $\theta^{(i)}$.
3. Get $E[\theta^{(i)}]$, Var($\theta^{(i)}$) and their limits.

We review pp 117-130 of Lecture B10 NONMEM Estimation of Noh[5] using simpler notaions. Let us start with an example of $\theta = (\theta_1, \theta_2)$ and solve the following equation based on Newton-Raphson method:

$$\nabla l = \begin{pmatrix} \frac{\partial l}{\partial\theta_1} \\ \frac{\partial l}{\partial\theta_2} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

We are looking for the point θ, where $l(\theta)$ does not change in any direction. Then, the first order Taylor polynomial is given by

$$\begin{pmatrix} \frac{\partial l}{\partial\theta_1} \\ \frac{\partial l}{\partial\theta_2} \end{pmatrix} \approx \begin{pmatrix} \frac{\partial l}{\partial\theta_1} \\ \frac{\partial l}{\partial\theta_2} \end{pmatrix}\Bigg|_{\substack{\theta_1=\theta_1^{(i)} \\ \theta_2=\theta_2^{(i)}}} + \begin{pmatrix} \frac{\partial^2 l}{\partial\theta_1^2} & \frac{\partial^2 l}{\partial\theta_2\partial\theta_1} \\ \frac{\partial^2 l}{\partial\theta_1\partial\theta_2} & \frac{\partial^2 l}{\partial\theta_2^2} \end{pmatrix}\Bigg|_{\substack{\theta_1=\theta_1^{(i)} \\ \theta_2=\theta_2^{(i)}}} \begin{pmatrix} \theta_1 - \theta_1^{(i)} \\ \theta_2 - \theta_2^{(i)} \end{pmatrix} = \mathbf{0}$$

Adopting the gradient ∇l and the matrix $H = \mathrm{H}l$ defined in Taylor series section, we can easily extend the previous equations to multidimensional parameters, where $\theta = (\theta_1, \cdots, \theta_p)$. Then, we want to solve the equation of more general form as follows:

$$\nabla l(\theta) \approx \nabla l(\theta^{(i)}) + H|_{\theta=\theta^{(i)}}(\theta - \theta^{(i)}) = \mathbf{0}$$

Solving it for θ, we have

$$H|_{\theta=\theta^{(i)}}(\theta - \theta^{(i)}) = -\nabla l(\theta^{(i)})$$

$$\theta - \theta^{(i)} = -H^{-1}\big|_{\theta=\theta^{(i)}}\nabla l(\theta^{(i)})$$

Let us define the solution θ at the i^{th} step as $\theta^{(i+1)}$. Then,

$$\theta^{(i+1)} = \theta^{(i)} - H^{-1}\big|_{\theta=\theta^{(i)}}\nabla l(\theta^{(i)})$$

Therefore, the limit of the sequence $\{\theta^{(i)}\}$ is called the MLE.

$$\hat{\theta}_{MLE} = \lim_{i\to\infty}\theta^{(i+1)} = \theta^{(\infty)}$$

Let us prove that, for large sample size, the MLE is an unbiased estimator of θ.

$$\begin{aligned} E[\theta^{(i+1)}] &= E[\theta^{(i)}] - H^{-1}\big|_{\theta=\theta^{(i)}}E\left[\nabla l(\theta^{(i)})\right] \\ &= \theta - H^{-1}\big|_{\theta=\theta^{(i)}}E\left[\nabla l(\theta^{(i)})\right] \\ &\to \theta \text{ as } n\to\infty \end{aligned}$$

since $E[\nabla l(\theta^{(i)})] = 0$.

Let us obtain the variance of approximate MLE.

$$\begin{aligned} & Var\left(\theta^{(i+1)}\right) \\ = \; & Var\left(\theta^{(i)} - H^{-1}\big|_{\theta=\theta^{(i)}}\nabla l(\theta^{(i)})\right) \\ = \; & H^{-1}\big|_{\theta=\theta^{(i)}}Var\left(\nabla l(\theta^{(i)})\right)H^{-1}\big|_{\theta=\theta^{(i)}} \\ & \text{considering } \theta^{(i)} \text{ and } H^{-1}\big|_{\theta=\theta^{(i)}} \text{ as fixed at step } i \\ \to \; & R^{-1}S_oR^{-1} \end{aligned}$$

where

$$R = \lim_{i\to\infty}H|_{\theta=\theta^{(i)}}$$

$$S_o = \lim_{i\to\infty}Var\left(\nabla l(\theta^{(i)})\right) = \lim_{i\to\infty}E[\nabla l(\theta^{(i)})(\nabla l(\theta^{(i)}))^T]$$

Note that $\theta^{(i)}$ does not affect the variance, $E[\nabla l(\theta^{(i)})] = 0$, and H, R, and S_o are symmetric. S is said to be an estimator of S_o and given by

$$S = \lim_{i\to\infty}\frac{1}{n}\sum_{all\ obs}\nabla l(\theta^{(i)})(\nabla l(\theta^{(i)}))^T$$

See p 172 of reference.[1] Therefore,

$$\hat{Var}\left(\hat{\theta}_{MLE}\right) = R^{-1}SR^{-1}$$

In data analyses, the followings are equivalent.
1. R is positive semidefinite, but singular.
2. H is negative semidefinite, but singular.
3. There are too many parameters and objective function could be flat.

S is singular if there are too many parameters.

More asymptotic properties of MLE

For further asymptotic properties of the MLE, we review pp151-154 of reference[1] and pp 83-84 of reference.[4] For large sample size

$$-2l(\theta) = -2\ln L \to \chi^2(p)$$

where p is the number of parameters to be estimated in the model. It can be used to test the goodness-of-fit. Moreover, it can be used to compare the nested model fits. As a general test to compare nested models, the likelihood ratio test (LRT) statistic can be defined. Suppose that L_s is for the restricted model

and L_g is for more general model. Then $L_g > L_s$, $-2 \ln L_g < -2 \ln L_s$,

$$-2 \ln L_s \rightarrow \chi^2(p_s), \text{ and } -2 \ln L_g \rightarrow \chi^2(p_g)$$

If we want to test

$$H_0 : \text{the restricted model } \theta = 0$$
$$H_1 : \text{the general model } \theta \neq 0$$

then the test statistic is given by

$$\Delta = -2 \ln L_s / L_g = -2(\ln L_s - \ln L_g) \rightarrow \chi^2(p_g - p_s)$$

Our decision is to reject H_0 at the significance level of α if $\Delta > \chi^2_{0.05}(P_g - P_s)$. Imposing penalty on many parameters, Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) are defined by

$$AIC = -2 \ln(L(\hat{\theta})) + 2n_{par}$$
$$BIC = -2 \ln(L(\hat{\theta})) + n_{par} \ln(n)$$

The models with smaller AIC and BIC are better in goodness-of-fit.

Discussion

Bonate[6] in pp 225-229 gives the nonlinear form of plasma concentration curves with both population parameters (θ, Σ) and individual parameters (η, Ω). Here θ is fixed effect and η is random effect. Since both the fixed effect and random effect appear in a model, it is a mixed-effects model whose likelihood function should include the pdf of η. Starting from the likelihood function based on the marginal pdf such as

$$L(\theta, \Sigma, \Omega | y) = \prod_{i=1}^{n} \int p(y_i | \theta, \Sigma, \eta_i) p(\eta_i | \Omega) d\eta$$

the objective function of nonlinear mixed-effects model implemented in NONMEM is rigorously derived based on the Taylor polynomial.

More details of MLE and its objective function for linear mixed-effects model can be found in pp 62-76 of reference.[4] For full algorithm of getting MLE and its objective function for nonlinear mixed-effects model, see pp 312-319 of reference.[4]

Acknowledgements

We thank Drs Gyeong Cheol Choi and Hwan Lee for their support in typing equations in Latex.

Conflict of interest

The authors have no conflict of interest.

References

1. Serfling RJ. Approximation theorems of mathematical statistics. John Wiley & Sons Inc, New York, 1980
2. Mood AM, Graybill FA, Boes DC. Introduction to the theory of statistics, 3rd ed. McGraw-Hill, New York, 1974
3. Johnson RA, Wichern DW. Applied multivariate statistical analysis, 5th ed. Englewood Cliffs, New Jersey, Prentice Hall, 2002
4. Pinheiro JC, Bates FM. Mixed-effects models in S and S-Plus. Springer, New York, 2000
5. Noh GJ. The 5th NONMEM workshop. B. NONMEM estimation (workshop manual), Seoul, 2014
6. Bonate PL. Pharmacokinetic-pharmacodynamic modeling and simulation. Springer, New York, 2006