STAT 604: David B. Dahl
Intro. to Statistical Computing Fall 2006

NEWTON-RAPHSON ALGORITHM

Introduction

A maximum likelihood estimator $\hat{\theta}_{\text{MLE}}$ is an estimator that maximizes the likelihood function of a parameter θ given data $x_1, \ldots, x_n$. Since the logarithm is a strictly-monotonic function, the maximizer of the likelihood function is also the maximizer of the log of the likelihood function. Often it is easier to maximize the log-likelihood than to maximize the likelihood itself.

Recall from calculus that the first derivative (or gradient) of a function evaluated at its maximizer must be zero. The Newton-Raphson method is an efficient root-finding method (i.e., a method for finding zeros) of real-valued functions. The method is based on a second-order Taylor series approximation, which is easy to maximize. For maximum likelihood estimation in statistics, the real-valued function of interest is the first derivative (or gradient) of the log-likelihood function.

Univariate Version

Let $f(\theta) = l(\theta|x_1, \ldots, x_n)$ denote the log-likelihood function evaluated at θ. This section assumes that θ is real-valued. (The multivariate case is considered later.) The goals is to find $\hat{\theta}$ such that $f(\theta)$ is maximized. Equivalently, for some θ_i, the goal is to find $\hat{h}$ such that $f(\theta_i + h)$ is maximized. Although $f(\theta_i + h)$ may be hard to maximize as a function of h, polynomials are trivial to maximize. For h sufficient close to zero (in absolute value), the Taylor's series approximation of $f(\theta_i + h)$ is:

$$f(\theta_i + h) \approx f(\theta_i) + f'(\theta_i)h + \frac{1}{2}f''(\theta_i)h^2, \tag{1}$$

where $f'(\theta_i)$ and $f''(\theta_i)$ denote the first and second derivatives of f evaluated at θ_i. This is a second-order polynomial in h. The derivative (with respect to h) of the right hand side of the previous equation is:

$$f'(\theta_i) + f''(\theta_i)h.$$

Setting the derivative above to zero and solving for h yields:

$$\hat{h} = -\frac{f'(\theta_i)}{f''(\theta_i)}.$$

If $f''(\theta_i) < 0$, $\theta_i + \hat{h}$ is a maximizer of the right-hand side of 1 and is an approximate maximizer of $f(\theta_i + h)$, the left-hand side of 1. This approximation can be improved by setting $\theta_{i+1} = \theta_i + \hat{h}$ and repeating the procedure above until $|f'(\theta_i)| \leq \epsilon$.

Algorithmically, the univariate Newton-Raphson method is:

1. Let θ_0 denote an initial guess as to the maximizer of f.
2. Let $i = 0$.
3. While $|f'(\theta_i)| > \epsilon$:
   (a) Let $i = i + 1$.
   (b) Let $\theta_i = \theta_{i-1} - f'(\theta_{i-1})/f''(\theta_{i-1})$
4. θ_i corresponds either to an approximate maximum, an approximate minimum, or an approxmate point of inflection of f. To verify that it is a maximizer, check that $f''(\theta_i) < 0$. If so, set $\hat{\theta}_{\text{MLE}} = \theta_i$, otherwise try another starting point for θ_0 or use a different algorithm.

Multivariate Version

The multivariate version of the Newton-Raphson is conceptually the same as the univariate version and notationally very similar. The log-likelihood $f(\boldsymbol{\theta})$ is now a function of a vector $\boldsymbol{\theta}$, and $\boldsymbol{h}$ is a vector of the same dimension as $\boldsymbol{\theta}$. For $\boldsymbol{h}$ sufficient close to zero (in terms of its norm), the Taylor's series approximation of $f(\boldsymbol{\theta}_i + \boldsymbol{h})$ is:

$$f(\boldsymbol{\theta}_i + \boldsymbol{h}) \approx f(\boldsymbol{\theta}_i) + [\nabla f(\boldsymbol{\theta}_i)]'\boldsymbol{h} + \frac{1}{2}\boldsymbol{h}'[D^2 f(\boldsymbol{\theta}_i)]\boldsymbol{h}, \tag{2}$$

where $\nabla f(\boldsymbol{\theta}_i)$ is the gradient of f evaluated at $\boldsymbol{\theta}_i$ and $D^2 f(\boldsymbol{\theta}_i)$ is Hessian matrix evaluated at $\boldsymbol{\theta}_i$. This is a second-order polynomial in the vector $\boldsymbol{h}$.

The gradient (with respect to $\boldsymbol{h}$) of the right hand side of the previous equation is:

$$\nabla f(\boldsymbol{\theta}_i) + [D^2 f(\boldsymbol{\theta}_i)]\boldsymbol{h}.$$

Setting the gradient above to zero gives a linear system of equations:

$$D^2 f(\boldsymbol{\theta}_i)\hat{\boldsymbol{h}} = -\nabla f(\boldsymbol{\theta}_i)$$

Solving this linear system of equations yields:

$$\hat{\boldsymbol{h}} = -[D^2 f(\boldsymbol{\theta}_i)]^{-1}[\nabla f(\boldsymbol{\theta}_i)].$$

For numerical stability on the computer, this system of linear equations should not be solved with an inverse. Instead, software should use one of the many matrix decompositions in linear alegra (such as the QR or LU decomposition).

If $D^2 f(\boldsymbol{\theta_i})$ is a negative definite matrix (i.e., all of its eigenvalues are negative), $\boldsymbol{\theta}_i + \hat{\boldsymbol{h}}$ is a maximizer of the right-hand side of 2 and is an approximate maximizer of $f(\boldsymbol{\theta}_i + \boldsymbol{h})$, the left-hand side of 2. This approximation can be improved by setting $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + \hat{\boldsymbol{h}}$ and repeating the procedure above until $|\nabla f(\boldsymbol{\theta}_i)| \leq \epsilon$.

Algorithmically, the multivariate Newton-Raphson method is:

1. Let $\boldsymbol{\theta}_0$ denote an initial guess as to the maximizer of f.
2. Let $i = 0$.
3. While $|\nabla f(\boldsymbol{\theta}_i)| > \epsilon$:
   (a) Let $i = i + 1$.
   (b) Let $\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i-1} - [D^2 f(\boldsymbol{\theta}_{i-1})]^{-1}[\nabla f(\boldsymbol{\theta}_{i-1})]$
4. $\boldsymbol{\theta}_i$ corresponds either to an approximate maximum, an approximate minimum, or an approxmate point of inflection of f. To verify that it is a maximizer, check that $D^2 f(\boldsymbol{\theta_i})$ is negative definate. If so, set $\hat{\boldsymbol{\theta}}_{\text{MLE}} = \boldsymbol{\theta}_i$, otherwise try another starting point for $\boldsymbol{\theta}_0$ or use a different algorithm.

Examples

See the class website for examples.