Newton–Raphson Iteration and the Method of Scoring

Steffen Lauritzen, University of Oxford

BS2 Statistical Inference, Lecture 4, Hilary Term 2009

February 4, 2009

Under suitable regularity conditions, the maximum likelihood estimator is a solution to the score equation

$$s(\theta) = s(x;\theta) = \frac{\partial}{\partial\theta} l(\theta) = \frac{\partial}{\partial\theta} \log L(\theta;x) = 0,$$

where $S(\theta) = s(X;\theta)$ is the *score statistic.*

Generally the solution to this equation must be calculated by iterative methods.

One of the most common methods is the *Newton–Raphson method* and this is based on successive approximations to the solution, using Taylor's theorem to approximate the equation.

Thus, we take an initial value θ_0 and write

$$0 = S(\theta_0) - J(\theta_0)(\theta - \theta_0),$$

ignoring the remainder term. Here

$$J(\theta) = J(\theta; X) = -\frac{\partial}{\partial\theta} S(\theta) = -\frac{\partial^2}{\partial\theta^2} l(\theta).$$

Solving this equation for θ then yields a new value θ_1

$$\theta_1 = \theta_0 + J(\theta_0)^{-1} S(\theta_0)$$

and we keep repeating this procedure as long as $|S(\theta_j)| > \epsilon$, i.e.

$$\theta_{k+1} = \theta_k + J(\theta_k)^{-1} S(\theta_0).$$

Clearly, $\hat{\theta}$ is a fixed point of this iteration as $S(\hat{\theta}) = 0$ and, conversely, any fixpoint is a solution to the likelihood equation.

If $\hat{\theta}$ is a local maximum for the likelihood function, we must have

$$J(\hat{\theta}) = -\frac{\partial^2}{\partial \theta^2} l(\hat{\theta}) > 0.$$

The quantity $J(\hat{\theta})$ determines the sharpness of the peak in the likelihood function around its maximum. It is also known as the *observed information*.

Occasionally we also use this term for $J(\theta)$ where θ is arbitrary but strictly speaking this can be quite inadequate as $J(\theta)$ may well be negative (although positive in expectation).

Recall that the (expected) Fisher information is

$$I(\theta) = \mathbf{E}\{J(\theta)\}$$

and that for large i.i.d. samples it holds approximately that $\hat{\theta} \sim \mathcal{N}(\theta, I(\theta)^{-1})$. In contrast to the observed information, $I(\theta)$ is non-negative everywhere, and in regular cases even strictly positive.

But it is also approximately true, to be elaborated later, under the same assumptions that

$$\sqrt{J(\hat{\theta})}(\hat{\theta} - \theta) \sim \mathcal{N}(0, 1),$$

so we could write $\hat{\theta} \sim \mathcal{N}(\theta, J(\hat{\theta})^{-1})$.

In fact, the observed information is in many ways preferable to the expected information. Indeed, *$\hat{\theta}$ is approximately sufficient* and *$J(\hat{\theta})$ is approximately ancillary.*

Formally the iteration becomes

- Choose an initial value θ; calculate $S(\theta)$ and $J(\theta)$;

Formally the iteration becomes

- Choose an initial value θ; calculate $S(\theta)$ and $J(\theta)$;
- `while` $|S(\theta)| > \epsilon$ `repeat`
  1. $\theta \leftarrow \theta + J(\theta)^{-1} S(\theta)$
  2. Calculate $S(\theta)$ and $J(\theta)$ `go to 1`

Formally the iteration becomes

- Choose an initial value θ; calculate $S(\theta)$ and $J(\theta)$;
- `while` $|S(\theta)| > \epsilon$ `repeat`
  1. $\theta \leftarrow \theta + J(\theta)^{-1} S(\theta)$
  2. Calculate $S(\theta)$ and $J(\theta)$ `go to 1`
- `return` θ;

Formally the iteration becomes

- Choose an initial value θ; calculate $S(\theta)$ and $J(\theta)$;
- `while` $|S(\theta)| > \epsilon$ `repeat`
  1. $\theta \leftarrow \theta + J(\theta)^{-1} S(\theta)$
  2. Calculate $S(\theta)$ and $J(\theta)$ `go to 1`
- `return` θ;

Other criteria for terminating the iteration can be used. To get a criterion which is insensitive to scaling of the variables, one can instead use the criterion $J(\theta)^{-1} S(\theta)^2 > \epsilon$.

Formally the iteration becomes

- Choose an initial value θ; calculate $S(\theta)$ and $J(\theta)$;
- `while` $|S(\theta)| > \epsilon$ `repeat`
  1. $\theta \leftarrow \theta + J(\theta)^{-1}S(\theta)$
  2. Calculate $S(\theta)$ and $J(\theta)$ `go to 1`
- `return` θ;

Other criteria for terminating the iteration can be used. To get a criterion which is insensitive to scaling of the variables, one can instead use the criterion $J(\theta)^{-1}S(\theta)^2 > \epsilon$.

Note that, as a by-product of this algorithm, the final value of $J(\theta)$ is the observed information which can be used to assess the uncertainty of $\hat{\theta}$.

If θ_0 is chosen sufficiently near $\hat{\theta}$ convergence is very fast.

It can be computationally expensive to evaluate $J(\theta)$ a large number of times. This is sometimes remedied by only changing J every 10 iterations or similar.

Another problem with the Newton–Raphson method is its lack of stability. When the initial value θ_0 is far from θ it might wildly oscillate and not converge at all. This is sometimes remedied by making smaller steps as

$$\theta \leftarrow \theta + \gamma J(\theta)^{-1} S(\theta)$$

where $0 < \gamma < 1$ is a constant. An alternative (or additional) method of stabilization is to let

$$\theta \leftarrow \theta + \gamma \{J(\theta) + S(\theta)^2\}^{-1} S(\theta)$$

as this avoids taking large steps when $S(\theta)$ is large.

The iteration has a tendency to be unstable for many reasons, one of them being that $J(\theta)$ may be negative unless θ already is very close to the MLE $\hat{\theta}$. In addition, $J(\theta)$ might sometimes be hard to calculate.

R. A. Fisher introduced the *method of scoring* which simply replaces the observed second derivative with its expectation to yield the iteration

$$\theta \leftarrow \theta + I(\theta)^{-1} S(\theta).$$

In many cases, $I(\theta)$ is easier to calculate and $I(\theta)$ is always positive. This generally stabilizes the algorithm, but here it can also be necessary to iterate as

$$\theta \leftarrow \theta + \gamma \{I(\theta) + S(\theta)^2\}^{-1} S(\theta).$$

In the case of n independent and identically distributed observations we have $I(\theta) = nI_1(\theta)$ so

$$\theta \leftarrow \theta + I_1(\theta)^{-1} S(\theta)/n$$

where $I_1(\theta)$ is the Fisher information in a single observation.

In a linear canonical one-parameter exponential family

$$f(x;\theta) = b(x)e^{\theta t(x) - c(\theta)}$$

we get

$$J(\theta) = \frac{\partial^2}{\partial \theta^2}\{c(\theta) - \theta t(X)\} = c''(\theta) = I(\theta).$$

so *for canonical exponential families the method of scoring and the method of Newton–Raphson coincide.*

If we let $v(\theta) = c''(\theta) = I(\theta) = \mathbf{V}(t(X))$ the iteration becomes

$$\theta \leftarrow \theta + v(\theta)^{-1} S(\theta)/n.$$

The identity of Newton–Raphson and the method of scoring *only holds for the canonical parameter.* If $\theta = g(\mu)$

$$\begin{aligned} J(\mu) &= \frac{\partial^2}{\partial \mu^2}[c\{g(\mu)\} - g(\mu)t(X)] \\ &= \frac{\partial}{\partial \mu}\left[g'(\mu)\tau\{g(\mu)\} - g'(\mu)t(X)\right] \\ &= v\{g(\mu)\}\{g'(\mu)\}^2 + g''(\mu)\left[\tau\{g(\mu)\} - t(X)\right] \end{aligned}$$

where we have let $\tau(\theta) = c'(\theta) = \mathbf{E}_\theta\{t(X)\}$ and $v(\theta) = c''(\theta) = \mathbf{V}_\theta\{t(X)\}$.

The method of scoring is simpler because the last term has expectation equal to 0:

$$I(\mu) = \mathbf{E}\{J(\mu)\} = v\{g(\mu)\}\{g'(\mu)\}^2.$$

The considerations on the previous overheads readily generalize to the multi-parameter case. The approximation to the score equation becomes

$$0 = S(\theta_0) - J(\theta_0)(\theta - \theta_0)$$

where

$$S(\theta)_r = \frac{\partial}{\partial \theta_r} l(\theta), \quad J(\theta)_{rs} = -\frac{\partial^2}{\partial \theta_r \partial \theta_s} l(\theta),$$

i.e. $S(\theta)$ is the *gradient* and $-J(\theta)$ the *Hessian* of $l(\theta)$.

The iterative step can still be written as

$$\theta \leftarrow \theta + J(\theta)^{-1} S(\theta)$$

where we just have to remember that the score statistic S is a *vector* and the Hessian $-J$ a *matrix*.

The lack of stability of the Newton–Raphson algorithm is not getting better in the multiparameter case. On the contrary there are not only problems with negativity, but the matrix can be singular and not invertible or it can have both positive and negative eigenvalues.

Recall that a symmetric matrix A is *positive definite* if all its eigenvalues are positive or, equivalently, if $x^\top Ax > 0$ for all $x \neq 0$. Sylvester's theorem says that *A is positive definite if and only if* $\det(A_R) > 0$ *for all submatrices A_R of the form* $\{a_{rs}\}_{r,s=1\ldots,R}$.

It is therefore also here advisable to replace $J(\theta)$ with its expectation, the Fisher information matrix, i.e. iterate as

$$\theta \leftarrow \theta + I(\theta)^{-1} S(\theta)$$

where now $I(\theta)$ is the Fisher information matrix which is always positive definite if the model is not over-parametrized.

Also in the multi-parameter case it can be advisable to stabilize additionally, i.e. by iterating as

$$\theta \leftarrow \theta + \gamma \{ I(\theta) + S(\theta) S(\theta)^\top \}^{-1} S(\theta)$$

or

$$\theta \leftarrow \theta + \gamma \{ I(\theta) + S(\theta)^\top S(\theta) E \}^{-1} S(\theta),$$

where E is the identity matrix.

In a multi-parameter curved exponential family with densities

$$f(x;\beta) = b(x)e^{\theta(\beta)^\top t(x) - c\{\theta(\beta)\}}$$

where β is d-dimensional, we get

$$\begin{aligned} J(\beta) &= \frac{\partial^2}{\partial\beta\partial\beta^\top}\left[c\{\theta(\beta)\} - \theta(\beta)^\top t(X)\right] \\ &= \frac{\partial}{\partial\beta}\left[\left(\frac{\partial\theta}{\partial\beta}\right)^\top \tau\{\theta(\beta)\} - \left(\frac{\partial\theta}{\partial\beta}\right)^\top t(X)\right] \\ &= \frac{\partial^2\theta}{\partial\beta\partial\beta^\top}\left[\tau\{\theta(\beta)\} - t(X)\right] + \left(\frac{\partial\theta}{\partial\beta}\right)^\top v\{\theta(\beta)\}\left(\frac{\partial\theta}{\partial\beta}\right), \end{aligned}$$

where the first term has expectation zero so

$$I(\beta) = \mathbf{E}\{J(\theta)\} = \left(\frac{\partial\theta}{\partial\beta}\right)^\top v\{\theta(\beta)\}\left(\frac{\partial\theta}{\partial\beta}\right).$$

In the multi-parameter case it is in wide generality *approximately* true that

$$\hat{\theta} \sim \mathcal{N}_d(\theta, I(\theta)^{-1})$$

or with a slight imprecision

$$\hat{\theta} \sim \mathcal{N}_d(\theta, J(\hat{\theta})^{-1})$$

where $\mathcal{N}_d$ is the d-dimensional Gaussian distribution, to be described later.

In particular it holds *approximately* that

$$(\hat{\theta}-\theta)^\top I(\theta)(\hat{\theta}-\theta) \sim (\hat{\theta}-\theta)^\top I(\hat{\theta})(\hat{\theta}-\theta) \sim (\hat{\theta}-\theta)^\top J(\hat{\theta})(\hat{\theta}-\theta) \sim \chi^2(d).$$