ELSEVIER Computational Statistics & Data Analysis 30 (1999) 13–17

COMPUTATIONAL STATISTICS & DATA ANALYSIS

The lower bound method in probit regression

Dankmar Böhning [*]

Department of Epidemiology, Free University Berlin, Fabeckstr. 60-62, Haus 562, 14195 Berlin, Germany

Received 1 April 1996; received in revised form 1 July 1998; accepted 30 November 1998

Abstract

The lower bound principle consists of replacing the second derivative matrix of the log-likelihood by a global lower bound in the Loewner ordering. This bound is then used in the Newton–Raphson iteration instead of the Hessian matrix leading to a monotonically converging sequence of iterates. In this note we apply this principle to probit regression as an example of a generalized linear model for which observed and expected information matrix do not coincide. This note also points out that it is *not* sufficient to use a bound for the expected value of the Hessian matrix as has been suggested recently in the context of generalized linear models. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Probit link function and probit regression; Generalized linear model; Lower bound method; Quadratic approximation of log-likelihood

1. Introduction

Consider a log-likelihood function $L(\beta)$ with p-dimensional parameter $\beta = (\beta_1, \dots, \beta_p)^{\mathrm{T}}$ and associated second-order Taylor expansion around β_0:

$$L(\beta) = L(\beta_0) + \nabla L(\beta_0)^{\mathrm{T}}(\beta - \beta_0) + \tfrac{1}{2}(\beta - \beta_0)^{\mathrm{T}}\nabla^2 L(\beta^*)(\beta - \beta_0), \qquad (1)$$

where $\beta^* = (1-\alpha)\beta + \alpha\beta_0$ for some $\alpha \in (0,1)$.

[*] Corresponding author.

PII: S 0167-9473(98)00094-2

The lower bound method[1] (Böhning and Lindsay, 1988; Böhning, 1989, 1992, 1993) requires the existence of a lower bound matrix B for the second derivative matrix $\nabla^2 L(\beta)$ with elements $\nabla^2 L(\beta)_{jk} = (\partial^2 L/\partial\beta_j\partial\beta_k)(\beta)$:

$$\nabla^2 L(\beta) \geq B, \tag{2}$$

where "$\geq$" is Loewner ordering, e.g. for non-negative definite, symmetric matrices A, B : $A \geq B$ if their difference $A - B$ is non-negative definite. Then $\beta_{\mathrm{LB}} = \beta_0 - B^{-1}\nabla L(\beta_0)$ maximizes the quadratic form $Q_B(\beta) = L(\beta_0) + \nabla L(\beta_0)^{\mathrm{T}}(\beta - \beta_0) + \frac{1}{2}(\beta - \beta_0)^{\mathrm{T}}B(\beta - \beta_0)$ and, since $Q_B(\beta)$ is a global lower bound for the log-likelihood $L(\beta)$, β_{LB} leads to a monotonic increase of the log-likelihood:

$$\begin{aligned} L(\beta_{\mathrm{LB}}) - L(\beta_0) &\geq Q_B(\beta_{\mathrm{LB}}) - L(\beta_0) \\ &= -\nabla L(\beta_0)^{\mathrm{T}} B^{-1}\nabla L(\beta_0) + \tfrac{1}{2}(B^{-1}\nabla L(\beta_0))^{\mathrm{T}} B (B^{-1}\nabla L(\beta_0)) \\ &= -\tfrac{1}{2}\nabla L(\beta_0)^{\mathrm{T}} B^{-1}\nabla L(\beta_0) \geq 0, \end{aligned}$$

since B as a lower bound of $\nabla^2 L$ is also negative definite. These results have been developed previously in Böhning and Lindsay (1988), Böhning (1989, 1992, 1993) and applied favourably in the context of logistic regression and Cox regression. For example, in logistic regression it was shown that $B = -\frac{1}{4}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}$, where $\boldsymbol{x}_i$ is the ith vector of covariates. For this case in addition, a comparison of the lower bound method with the conventional Newton–Raphson method has been undertaken (Böhning, 1993) showing some benefit in overall computational efficiency for the lower bound method. The similarity to the EM algorithm (Dempster et al., 1977) is rather striking where also an approximating form, the expected complete-data log-likelihood, serves as a lower bound for the incomplete log-likelihood.

2. The result

Let Y be a binary variable with $\Pr\{Y=1\} = p$ and $\Pr\{Y=0\} = 1 - p$, $p \in (0,1)$. Also, let $\boldsymbol{x}$ be a p-dimensional vector of covariates, which is connected to $E(Y)$ by a generalized linear model with link function g : $p = E(Y) = g(\eta)$, η being the linear predictor $\eta = \beta^{\mathrm{T}}\boldsymbol{x}$ (McCullagh and Nelder, 1989). For convenience, only one observation is considered. The likelihood is $p^y(1-p)^{1-y}$ and the log-likelihood is $L(\beta) = y\log(p) + (1-y)\log(1-p)$. Straightforward computation yields

$$\frac{\partial L}{\partial\beta_j} = \mathrm{d}L/\mathrm{d}\eta \times \frac{\partial\eta}{\partial\beta_j} = \mathrm{d}L/\mathrm{d}\eta \times x_j$$

and

$$\frac{\partial^2 L}{\partial\beta_j\partial\beta_k} = \mathrm{d}^2 L/(d\eta)^2 \times \frac{\partial\eta}{\partial\beta_j}\frac{\partial\eta}{\partial\beta_k} = \mathrm{d}^2 L/(d\eta)^2 \times x_j x_k,$$

[1] The lower bound method was originally introduced in Böhning and Lindsay (1988), where it was abbreviated as LB-method. In the sequel other authors understood this abbreviation as a short form of Lindsay–Böhning method, which, of course, was not intended.

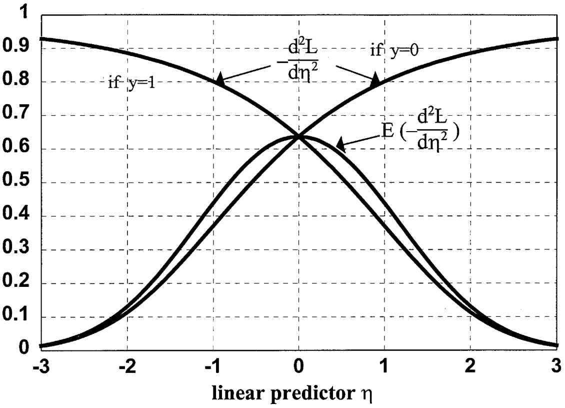

Fig. 1. Observed and expected information for the probit link function.

or, in vector notation

$$\nabla L(\beta) = \mathrm{d}L/\mathrm{d}\eta(\beta) \times \boldsymbol{x}, \qquad \nabla^2 L(\beta) = \mathrm{d}^2L/(\mathrm{d}\eta)^2(\beta) \times \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}}.$$

The existence of a lower bound depends solely on the existence of a lower bound for

$$\begin{aligned}\mathrm{d}^2L/(\mathrm{d}\eta)^2(\beta) &= y[g''(\eta)g(\eta) - g'(\eta)^2]/g(\eta)^2 \\ &\quad -(1-y)[g''(\eta)(1-g(\eta)) + g'(\eta)^2]/[1-g(\eta)]^2,\end{aligned}$$

where $g'(\eta)=\mathrm{d}g/\mathrm{d}\eta$, $g''(\eta)=\mathrm{d}^2g/(\mathrm{d}\eta)^2$. It is tempting to take expected values because a simple algebraic form can be reached for the *expected* information:

$$\begin{aligned}E\{\mathrm{d}^2L/(\mathrm{d}\eta)^2\} &= [g''(\eta)g(\eta) - g'(\eta)^2]/g(\eta) \\ &\quad -[g''(\eta)(1-g(\eta)) + g'(\eta)^2]/[1-g(\eta)] \\ &= -g'(\eta)^2/\{g(\eta)(1-g(\eta))\}.\end{aligned}$$

This is a simple function, for which bounds can be easily found depending on the form of the link function. Note that these results are independent on the form of the link function. Let us consider the probit link function, that is $g(\eta) = \Phi(\eta)$, Φ being the cumulative distribution function of the standard normal distribution and ϕ it's density: $\phi(\eta) = \exp(-\frac{1}{2}\eta^2)/\sqrt{2\pi}$. Fig. 1 shows $-E\{\mathrm{d}^2L/(\mathrm{d}\eta)^2\}$ for the probit link function, and it is easily seen that the maximum is attained for $\eta = 0$ for which $-E\{\mathrm{d}^2L/(\mathrm{d}\eta)^2(0)\} = 4/(2\pi) = 2/\pi \approx 0.6369$. This approach has been taken by Devidas and George (1995), and bounds could be provided for a whole class of link functions. However, it must be pointed out that this approach– though intuitive and appealing–suffers under the weakness of no longer guaranteeing the before mentioned monotonicity property, since it works with the *approximating* log-likelihood $L(\beta_0) + \nabla L(\beta_0)^{\mathrm{T}}(\beta - \beta_0) + \frac{1}{2}(\beta - \beta_0)^{\mathrm{T}}E[\nabla^2 L(\beta^*)](\beta - \beta_0)$, for which a global lower bound is found. This lower bound, however, might not be a global lower bound for $L(\beta)$ and the mathematical argument of replacing the log-likelihood by a lower bound collapses. It is necessary to find a lower bound for the *observed* information which is provided in the following theorem.

Theorem. *Let* $g(\eta) = \Phi(\eta)$ *the probit link function. Then the parameter-dependent part of the observed information matrix is bounded above by* 1:

$$-\nabla^2 L(\beta) = -\mathrm{d}^2 L/(d\eta)^2(\beta) \times \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \leq \boldsymbol{x}\boldsymbol{x}^{\mathrm{T}} \quad (\textit{for one observation})$$

$$-\nabla^2 L(\beta) = -\sum_{i=1}^{n} \mathrm{d}^2 L_i/(\mathrm{d}\eta_i)^2(\beta)\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} \leq \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} \quad (\textit{for a sample of size } n\textit{; the}$$

index i is referring to the ith contribution to the likelihood).

Proof. It is sufficient to consider $\mathrm{d}^2 L/(\mathrm{d}\eta)^2(\beta) = y[-\eta\phi(\eta)\Phi(\eta) - \phi(\eta)^2]/\Phi(\eta)^2 - (1-y)[-\eta\phi(\eta)(1-\Phi(\eta)) + \phi(\eta)^2]/[1-\Phi(\eta)^2]$. Note that one of the two terms must be 0 always.

Case 1: Let $y = 1$. Then $\mathrm{d}^2 L/(\mathrm{d}\eta)^2(\eta) = -\eta\phi(\eta)/\Phi(\eta) - [\phi(\eta)/\Phi(\eta)]^2$ (see Fig. 1). If η approaches $+\infty$, $\mathrm{d}^2 L/(\mathrm{d}\eta)^2(\eta)$ becomes 0. If η approaches $-\infty$, the rule of l'Hospital provides clarity: $\lim_{\eta\to\infty}\{[-\eta\phi(\eta)]'/\Phi'(\eta) - [\phi'(\eta)/\Phi'(\eta)]^2\} = \lim_{\eta\to\infty}\{[-\phi(\eta) + \eta^2\phi(\eta)]/\phi(\eta) - [-\eta\phi(\eta)/\phi(\eta)]^2\} = \lim_{\eta\to\infty}[-1 + \eta^2 - \eta^2] = -1$.

Case 2: The case $y = 0$ can be treated analogously and ends the proof. □

3. Discussion

The result leads to a particular simple form of iteration: $\beta_{\mathrm{LB}} = \beta_0 - B^{-1}\nabla L(\beta_0) = \beta_0 + [X^{\mathrm{T}}X]^{-1}\nabla L(\beta_0)$, where $X^{\mathrm{T}}X = \sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}$. Because of the fact that $X^{\mathrm{T}}X$ needs to be inverted only once, the lower bound method should also compare favourably in this case to the Newton–Raphson method as well as to Fisher–Scoring which also requires the inversion of $E(\nabla^2 L(\beta_0))$ at each iteration: $\beta_{\mathrm{FS}} = \beta_0 - E(\nabla^2 L(\beta_0))^{-1}\nabla L(\beta_0)$. This fact, in connection with the guaranteed monotonicity, makes the lower bound method attractive. The convergence rate is only linear, but the overall numerical complexity is rather low in comparison to Newton–Raphson or Fisher–Scoring as it has been demonstrated for logistic regression in Böhning (1993).

Acknowledgements

This research is under current support of the German Research Foundation.

References

Böhning, D., 1989. Likelihood inference for mixtures: geometrical and other constructions of monotone step-length algorithms. Biometrika 76, 375–383.

Böhning, D., 1992. Multinomial logistic regression algorithm. Ann. Inst. Statist. Math. 44, 197–200.

Böhning, D., 1993. Construction of reliable maximum likelihood algorithms with application to logistic and Cox regression. In: Rao, C.R. (Ed.), Handbook of Statistics, vol. 9. North-Holland, Amsterdam, pp. 409–422.

Böhning, D., Lindsay, B.G., 1988. Monotonicity of quadratic approximation algorithms. Ann. Inst. Statist. Math. 40, 641–663.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). J. Roy. Statist. Soc. B 39, 1–38.

Devidas, M., George, E.O., 1995. Monotonic algorithms for computing maximum likelihood estimates in generalized linear models. Preprint, Division of Biostatistics, University of Mississipi Medical Center.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, 2nd ed. Chapman & Hall, London.