Vol. 35 (2015)
No. 6

数 学 杂 志
J. of Math. (PRC)

Logistic 回归模型中参数极大似然估计的二次下界算法及其应用

王 佳, 丁洁丽
(武汉大学数学与统计学院, 湖北 武汉 430072)

摘 要: 本文研究了 Newton-Raphson 等算法无法进行时探寻更加稳定的数值解法的问题. 利用 Böhning & Linday (1988) 提出的二次下界算法 (Quadratic lower-bound), 文中在 Logistic 回归模型下构造了极大似然函数的代理函数并进行数值模拟, 获得了二次下界算法是 Newton-Raphson 算法的合理替代的结果, 推广了数值方法在 Logistic 回归模型中的应用.

关键词: minorization-maximization 算法; Logistic 回归模型; quadratic lower-bound 算法; 极大似然估计; Newton-Raphson 算法

MR(2010) **主题分类号:** 62J12; 65C20 **中图分类号:** O212.1

文献标识码: A **文章编号:** 0255-7797(2015)06-1521-12

1 引言

统计推断常常通过建立各种模型来探讨因变量与相关协变量之间的关系. 关于模型中参数的估计, 例如, 极大似然估计, 最小二乘估计等, 往往涉及到求解一些目标函数的极值问题. 以下我们将着重关注极大似然估计的求解问题. 很多模型下, 似然方程没有解析解, 这时需要采用数值方法求解参数的估计值, 例如, Newton-Raphson 算法, 随机梯度下降法 [1] 等. 其中, Newton-Raphson 算法是最为常用的数值方法之一, 然而, Newton-Raphson 算法在实际应用中可能遇到下述问题: 每次迭代均需要计算复杂的二次导数矩阵; 当参数维数较大时, 算法涉及到大维矩阵求逆的问题; 由于不同初值的选取, 算法可能收敛到似然函数的极小值点或者鞍点; 多次迭代导致矩阵奇异, 迭代无法继续进行; 由于算法没有上升性, 可能会导致不收敛的情况发生 [2]. 因此, 发展和研究更为稳定的数值算法是统计研究中的热点问题之一.

近年来, 对 Minorization-Maximization 算法 (以下简称 "MM 算法") 的研究与应用越来越广泛. MM 算法的中心思想是 "优化转移", 关键之处在于构造一个代理函数 (surrogate function), 从而将求解复杂的目标函数的极值问题转移为求解具有优良性质的代理函数的极值问题, 进而有效地避免上述提到的 Newton-Raphson 算法可能遇到的问题. MM 算法近来被广泛地应用到统计学各个研究领域, 例如, 稳健回归 [3], 关联度分析 [4], 最小二乘估计 [5–7], 医学影像分析 [8, 9], 分位数回归 [10], 生存分析 [11], DNA 排列分析 [12], 多重比较 [13], 变量选择 [14], 判别分析 [15, 16] 等等. MM 算法是一种算法的思想, 而非局限于一种具体的计算方法. 算法原理的关键处是构造一个合理的代理函数, 常用方法有: 线性化; 参数分离; 规避求解大维矩阵; 利用光滑代理函数替代非光滑目标函数等等 [17–22].

众所周知, 线性回归模型是定量分析研究中最为流行的统计分析方法. 然而, 当因变量为分类变量而不是连续变量时, 线性回归模型就不再适用了. 对于二元因变量的研究, Logistic

*收稿日期: 2014-11-15 接收日期: 2015-01-08
基金项目: 国家自然科学基金 (11101314).
作者简介: 王佳 (1989–), 女, 陕西咸阳, 硕士, 主要研究方向: 数理统计, 应用统计.

回归模型是使用最为普遍和广泛的分析方法. 在 Logistic 回归模型下, Böhning & Linday (1988) 为其极大似然估计的求解提出了一种二次下界 (Quadratic Lower-Bound) 算法 (以下简称“QLB 算法”) 来构造 MM 算法中的代理函数 [2]. QLB 算法用一个与参数无关的二次下界矩阵替代了负的观测信息矩阵, 从而, 在求解极大似然估计的整个迭代过程中仅需要计算一次该二次下界矩阵的逆, 而非如 Newton-Raphson 算法在每次迭代中均需要计算一次观测信息矩阵的逆.

本文以探讨 Logistic 回归模型中参数极大似然估计问题的 QLB 算法理论为支撑, 主要研究其在实际中的应用. 首先, 我们利用数值模拟方法比较了 Newton-Raphson 算法和 QLB 算法, 并评估了极大似然估计的大样本性质在有限样本下的表现. 进一步, 利用 QLB 算法分析了三个实际数据: 急性淋巴细胞性白血病病人数据, 教职工升职情况数据, 农村居民健康行为数据. 本文结构如下: 第 2 节中, 我们首先介绍了一般情形下的 QLB 算法的原理和方法, 接着, 探讨在 Logistic 回归模型下, 如何利用 QLB 算法求解参数极大似然估计. 第 3 节中, 我们先利用数值模拟比较 Newton-Raphson 和 QLB 两种算法, 最后应用 QLB 算法分析了三个实际数据.

2 QLB 算法

本节先介绍一般情形下的 QLB 算法的原理和方法, 然后, 在 Logistic 回归模型下, 利用 QLB 算法求解回归参数的极大似然估计.

2.1 QLB 算法的原理与方法

设 $l(\theta)$ 为感兴趣的对数似然函数, θ 为待估的 q 维参数, 则 θ 的极大似然估计为

$$\hat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, l(\theta). \tag{2.1}$$

上述极大值问题的求解常常转化为求解得分方程:

$$\frac{\partial l(\theta)}{\partial \theta} = 0.$$

然而, 实际中, 上述方程的解通常没有解析表达式, 故而需要采用数值方法求解方程, 其中 Newton-Raphson 算法是广泛使用的方法之一. 设 $\theta^{(n)}$ 表示第 n 步迭代所求得的 θ 的解, Newton-Raphson 迭代公式为

$$\theta^{(n+1)} = \theta^{(n)} - \left(\frac{\partial^2 l(\theta^{(n)})}{\partial\theta\partial\theta'}\right)^{-1} \frac{\partial l(\theta^{(n)})}{\partial\theta}. \tag{2.2}$$

前面提到, 在实际应用中, Newton-Raphson 算法常会遭遇初值选取, 观测信息阵奇异以及算法可能不收敛等问题. 而此时, MM 算法可以有效地避免上述问题的发生, 故而成为一种运用越来越广泛的方法.

MM 算法的主要思想是“优化转移”, 其关键之处在于构造一个代理函数 $Q(\theta|\theta^{(n)})$, 使之满足

$$l(\theta) - Q(\theta|\theta^{(n)}) \geq 0, \qquad \forall\theta\in\Theta,$$

且等号成立当且仅当 $\theta=\theta^{(n)}$. 若定义

$$\theta^{(n+1)}=\underset{\theta\in\Theta}{\operatorname{argmax}}\,Q(\theta|\theta^{(n)}),$$

则可得如下事实:

$$\begin{aligned} l(\theta^{(n+1)}) &= l(\theta^{(n+1)})-Q(\theta^{(n+1)}|\theta^{(n)})+Q(\theta^{(n+1)}|\theta^{(n)}) \\ &\geq l(\theta^{(n)})-Q(\theta^{(n)}|\theta^{(n)})+Q(\theta^{(n)}|\theta^{(n)}) \\ &= l(\theta^{(n)}). \end{aligned}$$

此性质使得 MM 算法与 EM 算法一样具有上升性, 这一优良性质. 正是由于这种上升性, 使得 (2.1) 中的极值问题可以转化为

$$\hat{\theta}=\underset{\theta\in\Theta}{\operatorname{argmax}}\,Q(\theta|\theta^{(n)}). \tag{2.3}$$

即将求解似然函数 $l(\theta)$ 极大值点的问题转移为求解代理函数 $Q(\theta|\theta^{(n)})$ 极大值点的问题.

构造代理函数的方法很多, 文献 [17–22] 以及文献 [11] 等分别对不同的统计模型构造了与之相应的代理函数. Böhning & Linday (1988) 提出了一种 QLB 算法来构造代理函数 [2]. 假设 $l(\theta)$ 满足二阶连续可微, 将 $l(\theta)$ 在 $\theta^{(n)}$ 处进行 Taylor 展开, 可得

$$l(\theta)=l(\theta^{(n)})+\left(\frac{\partial l(\theta^{(n)})}{\partial\theta}\right)'(\theta-\theta^{(n)})+\frac{1}{2}(\theta-\theta^{(n)})'\left(\frac{\partial^2 l(\tilde{\theta})}{\partial\theta\partial\theta'}\right)(\theta-\theta^{(n)}),$$

其中 $\tilde{\theta}$ 为 θ 和 $\theta^{(n)}$ 连线上的一点. QLB 算法关键之处在于找到一个与 θ 无关的负定矩阵 B, 使得

$$\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta'}\geq B, \qquad \forall\theta\in\Theta.$$

此时, 可构造代理函数形式如下:

$$Q(\theta|\theta^{(n)})=l(\theta^{(n)})+\left(\frac{\partial l(\theta^{(n)})}{\partial\theta}\right)'(\theta-\theta^{(n)})+\frac{1}{2}(\theta-\theta^{(n)})'B(\theta-\theta^{(n)}).$$

注意到, 上述代理函数 $Q(\theta|\theta^{(n)})$ 用 B 替代了似然函数 $l(\theta)$ 二阶展开式中的 $\left(\frac{\partial^2 l(\tilde{\theta})}{\partial\theta\partial\theta'}\right)$, 且满足

$$l(\theta)-Q(\theta|\theta^{(n)})\geq 0, \qquad \forall\theta\in\Theta,$$

且 $l(\theta^{(n)})=Q(\theta^{(n)}|\theta^{(n)})$. 故而可构造求解极值问题 (2.3) 的迭代公式如下:

$$\theta^{(n+1)}=\theta^{(n)}-B^{-1}\left(\frac{\partial Q(\theta^{(n)}|\theta^{(n)})}{\partial\theta}\right)=\theta^{(n)}-B^{-1}\left(\frac{\partial l(\theta^{(n)})}{\partial\theta}\right). \tag{2.4}$$

因为 $\theta^{(n+1)}=\underset{\theta\in\Theta}{\operatorname{argmax}}\,Q(\theta|\theta^{(n)})$, 即每一步的迭代都使得 $Q(\theta|\theta^{(n)})$ 达到极大值, 故而所得到的估计值序列 $\{\theta^{(n)}:\ n\geq 1\}$ 会逐渐逼近 $l(\theta)$ 的极大值点. Böhning & Linday (1988) 给出了上述 QLB 算法上升性并证明了其算法的收敛性 [2]. 注意到, 与 Newton-Raphson 算法相

比, QLB 算法的最大优点在于: 由于二次下界矩阵 B 与 θ 无关, 因此, QLB 算法不需要在每次迭代中计算观测信息矩阵的逆, 这个优势在计算大维矩阵的逆时尤为突出. 当然, 二次下界矩阵 B 的选取不是唯一的, 不同的 B 会产生不同的迭代步长, 从而产生不同的收敛速度.

2.2 Logistic 回归模型下的 QLB 算法

Logistic 回归模型广泛地应用于各个领域, 例如, 流行病学, 生物医学, 经济学等等. 假设有 N 个独立样本, 记 Y_i 为第 i 个样本的二元响应变量, X_i 为第 i 个样本的 q 维协变量, $i = 1, \cdots, N$. 我们考虑下述 Logistic 回归模型:

$$P(Y_i = 1) = \frac{e^{X_i'\theta}}{1 + e^{X_i'\theta}}, \quad i = 1, \cdots, N,$$

其中 θ 为待估参数. 记 $p_i = P(Y_i = 1)$, 对数似然函数有如下形式:

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{N} (Y_i \log p_i + (1 - Y_i) \log(1 - p_i)) \\ &= \sum_{i=1}^{N} \left(Y_i (X_i'\theta) - \log(1 + e^{X_i'\theta}) \right). \end{aligned}$$

从而得分方程和观测信息矩阵分别为

$$\begin{aligned} \frac{\partial l(\theta)}{\partial \theta} &= \sum_{i=1}^{N} (Y_i - p_i) X_i, \\ -\frac{\partial^2 l(\theta)}{\partial \theta \partial \theta'} &= \sum_{i=1}^{N} p_i (1 - p_i) X_i X_i', \end{aligned} \tag{2.5}$$

从而可得求解极大似然估计 $\hat{\theta}$ 的 Newton-Raphson 迭代公式为

$$\theta^{(n+1)} = \theta^{(n)} + \left(\sum_{i=1}^{N} p_i^{(n)} (1 - p_i^{(n)}) X_i X_i' \right)^{-1} \left(\sum_{i=1}^{N} (Y_i - p_i^{(n)}) X_i \right), \tag{2.6}$$

其中 $p_i^{(n)} = \frac{e^{X_i'\theta^{(n)}}}{1 + e^{X_i'\theta^{(n)}}}$ 为 p_i 的第 n 次迭代值.

下面我们来应用 QLB 算法求解 Logistic 模型中回归参数的极大似然估计. 我们知道, 算法关键是寻找二次下界矩阵 B, 使得 B 是二阶导数矩阵 $\left(\frac{\partial^2 l(\theta)}{\partial \theta \partial \theta'} \right)$ 的一个下界. Böhning & Linday (1988)[2] 通过对 (2.5) 式的观察得到: (1) $X_i X_i'$ 是非负定矩阵; (2) $p_i(1 - p_i) \leq \frac{1}{4}$, $i = 1, \cdots, N$. 因此有

$$\frac{\partial^2 l(\theta)}{\partial \theta \partial \theta'} = -\sum_{i=1}^{N} p_i (1 - p_i) X_i X_i' \geq -\frac{1}{4} \sum_{i=1}^{N} X_i X_i',$$

对任意 $\theta \in \Theta$ 均成立. 从而找到了 QLB 算法中的一个二次下界矩阵 $B = -\frac{1}{4} \sum_{i=1}^{N} X_i X_i'$. 根据均值不等式可知, 这里的矩阵 B 是二阶导数矩阵的最大下界, 也就是说此处的矩阵 B 是二阶

导数矩阵所有下界的上确界. 因此, 我们可得求解 $\hat{\theta}$ 的 QLB 算法迭代公式如下:

$$\theta^{(n+1)} = \theta^{(n)} + \left(\frac{1}{4}\sum_{i=1}^{N} X_i X_i'\right)^{-1} \left(\sum_{i=1}^{N}(Y_i - p_i^{(n)})X_i\right). \tag{2.7}$$

Böhning & Linday (1988) 证明了此迭代算法是单调线性收敛的 [2]. 由于 QLB 算法中的矩阵 B 只与样本有关, 若矩阵 $\left(\sum\limits_{i=1}^{N} X_i X_i'\right)$ 不奇异或不接近奇异, 那么仅需要在整个迭代过程计算一次矩阵的逆即可.

进一步, 对于极大似然估计 $\hat{\theta}$ 方差的估计问题, 在 Newton-Raphson 算法下, 由于极大似然估计渐近正态性成立, 我们可以利用渐近方差的样本经验估计作为 $\hat{\theta}$ 方差的估计, 即 $\hat{\theta}$ 的协方差估计为 $\widehat{\mathrm{Cov}}(\hat{\theta}) = (-\frac{\partial^2 l(\hat{\theta})}{\partial\theta\partial\theta'})^{-1}$, 从而 $\hat{\theta}$ 的标准差 (Standard Errors) 为 $\widehat{\mathrm{Cov}}(\hat{\theta})$ 对角元素的平方根. 而在 QLB 算法下, 我们则可以应用广泛使用的非参数 Bootstrap 重抽样方法 [23, 24] 估计 $\hat{\theta}$ 的方差.

3 QLB 算法的应用

本节中, 我们将 QLB 算法应用于 Logistic 回归模型中参数的估计问题. 首先, 我们进行数值模拟, 分别应用 QLB 算法与 Newton-Raphson 算法计算参数的极大似然估计. 然后, 应用 QLB 算法分析了三个实际数据. 所有分析均应用 R 软件.

3.1 数值模拟

我们考虑下述 Logistic 模型:

$$P(Y=1) = \frac{e^{\theta_0+\theta_1 X_1+\theta_2 X_2}}{1+e^{\theta_0+\theta_1 X_1+\theta_2 X_2}}.$$

先从成功概率为 0.2 的 Bernoulli 分布生成协变量 X_1, 从标准正态分布生成 X_2, 从而二元响应变量 Y 可由逆变换法产生. 为了更好地研究模拟结果, 我们考虑不同的参数取值组合. 分别取 $\theta_0 = 1$, $\theta_1 = -0.5,\ 0,\ 0.5$, $\theta_2 = -0.5,\ 0,\ 0.5$. 分别取样本量 $N = 300,\ 400$, 并设定 QLB 算法下 Bootstrap 次数为 500 次.

对于每种参数组合, 比较两种算法的参数估计结果:

(1) $\hat{\theta}_N$: 应用 Newton-Raphson 算法, 基于迭代公式 (2.6) 计算所得的极大似然估计;

(2) $\hat{\theta}_Q$: 应用 QLB 算法, 基于迭代公式 (2.7) 计算所得的极大似然估计. 基于 1000 次模拟下, 我们得到估计的均值 (Mean), 估计的样本标准差 (SD), 标准差估计的均值 (SE), 以及 95% 正态区间估计覆盖率 (CP). 模拟结果请见表 1. 表 1 中结果表明, 在每种模拟设定下, θ_1 和 θ_2 的极大似然估计均为无偏估计. SE 与 SD 充分接近, 表明提出的标准差的估计很好地估计了极大似然估计的样本标准差, 说明提出的标准差估计在实际应用中的合理性. 区间估计覆盖率均接近 95%, 表明在考虑的有限样本下, 极大似然估计的渐近正态性表现优良. 进一步, 当样本量增大时, 估计的效率提高. 数值模拟结果也表明, QLB 算法与 Newton-Raphson 算法所得的结果一致, 这表明, QLB 算法作为 Newton-Raphson 算法的合理替代, 能很好地应用于 Logistic 回归参数的估计问题.

表 1: Logistic 回归模型下参数估计的模拟结果

			$\hat{\theta}_1$				$\hat{\theta}_2$			
N	$(\theta_1,\ \theta_2)$		$Mean$	SD	SE	CP	$Mean$	SD	SE	CP
300	$(-0.5,\ -0.5)$	$\hat{\theta}_N$	−0.4616	0.2919	0.3169	0.970	−0.5017	0.1409	0.1391	0.950
		$\hat{\theta}_Q$	−0.4966	0.3215	0.3260	0.960	−0.5101	0.1369	0.1422	0.962
	$(-0.5,\ 0)$	$\hat{\theta}_N$	−0.4939	0.3114	0.3089	0.955	−0.0031	0.1356	0.1299	0.945
		$\hat{\theta}_Q$	−0.4932	0.3281	0.3190	0.956	−0.0025	0.1285	0.1330	0.951
	$(-0.5,\ 0.5)$	$\hat{\theta}_N$	−0.4960	0.3123	0.3157	0.947	0.5066	0.1444	0.1381	0.942
		$\hat{\theta}_Q$	−0.4994	0.3113	0.3250	0.958	0.5082	0.1457	0.1421	0.953
	$(0,\ -0.5)$	$\hat{\theta}_N$	0.0303	0.3444	0.3376	0.955	−0.5147	0.1399	0.1406	0.956
		$\hat{\theta}_Q$	0.0071	0.3490	0.3519	0.965	−0.5032	0.1380	0.1439	0.948
	$(0,\ 0)$	$\hat{\theta}_N$	−0.0015	0.3318	0.3319	0.954	−0.0029	0.1367	0.1321	0.950
		$\hat{\theta}_Q$	0.0005	0.3405	0.3476	0.957	0.0018	0.1330	0.1349	0.960
	$(0,\ 0.5)$	$\hat{\theta}_N$	0.0286	0.3547	0.3379	0.947	0.5157	0.1422	0.1406	0.955
		$\hat{\theta}_Q$	0.0139	0.3397	0.3520	0.971	0.5115	0.1386	0.1447	0.954
	$(0.5,\ -0.5)$	$\hat{\theta}_N$	0.5512	0.3832	0.3773	0.957	−0.5007	0.1436	0.1434	0.955
		$\hat{\theta}_Q$	0.5436	0.3798	0.4030	0.971	−0.5178	0.1406	0.1470	0.955
	$(0.5,\ 0)$	$\hat{\theta}_N$	0.5397	0.3840	0.3762	0.954	−0.0037	0.1331	0.1352	0.960
		$\hat{\theta}_Q$	0.5380	0.4157	0.4040	0.952	−0.0008	0.1334	0.1386	0.965
	$(0.5,\ 0.5)$	$\hat{\theta}_N$	0.5337	0.3927	0.3767	0.948	0.5059	0.1459	0.1436	0.948
		$\hat{\theta}_Q$	0.5384	0.3845	0.4020	0.972	0.5079	0.1425	0.1470	0.964
400	$(-0.5,\ -0.5)$	$\hat{\theta}_N$	−0.4623	0.2670	0.2734	0.959	−0.5016	0.1165	0.1194	0.959
		$\hat{\theta}_Q$	−0.5004	0.2873	0.2789	0.949	−0.5133	0.1173	0.1217	0.965
	$(-0.5,\ 0)$	$\hat{\theta}_N$	−0.5099	0.2636	0.2667	0.967	−0.0020	0.1112	0.1116	0.950
		$\hat{\theta}_Q$	−0.5013	0.2662	0.2723	0.958	−0.0029	0.1127	0.1137	0.955
	$(-0.5,\ 0.5)$	$\hat{\theta}_N$	−0.4967	0.2646	0.2723	0.952	0.5069	0.1190	0.1190	0.956
		$\hat{\theta}_Q$	−0.4980	0.2672	0.2782	0.960	0.5152	0.1213	0.1217	0.952
	$(0,\ -0.5)$	$\hat{\theta}_N$	0.0017	0.2910	0.2891	0.956	−0.5077	0.1220	0.1214	0.954
		$\hat{\theta}_Q$	0.0085	0.3044	0.2992	0.950	−0.5072	0.1253	0.1234	0.943
	$(0,\ 0)$	$\hat{\theta}_N$	0.0121	0.2957	0.2869	0.940	0.0033	0.1175	0.1142	0.955
		$\hat{\theta}_Q$	0.0125	0.2995	0.2953	0.953	0.0044	0.1161	0.1163	0.953
	$(0,\ 0.5)$	$\hat{\theta}_N$	0.0296	0.2891	0.2902	0.961	0.5098	0.1214	0.1212	0.958
		$\hat{\theta}_Q$	0.0228	0.2964	0.2995	0.962	0.5046	0.1295	0.1236	0.936
	$(0.5,\ -0.5)$	$\hat{\theta}_N$	0.5076	0.3362	0.3225	0.945	−0.5027	0.1292	0.1237	0.948
		$\hat{\theta}_Q$	0.5307	0.3400	0.3410	0.962	−0.5080	0.1219	0.1259	0.956
	$(0.5,\ 0)$	$\hat{\theta}_N$	0.5292	0.3400	0.3244	0.950	−0.0027	0.1235	0.1173	0.937
		$\hat{\theta}_Q$	0.5429	0.3289	0.3415	0.971	−0.0074	0.1178	0.1186	0.948
	$(0.5,\ 0.5)$	$\hat{\theta}_N$	0.5239	0.3185	0.3232	0.958	0.5064	0.1242	0.1241	0.955
		$\hat{\theta}_Q$	0.5404	0.3329	0.3396	0.964	0.5161	0.1285	0.1265	0.946

说明: $\hat{\theta}_N$ 为 Newton-Raphson 算法; $\hat{\theta}_Q$ 为 QLB 算法.

3.2 实例分析

实例 1 (急性淋巴细胞性白血病病人数据)

我们首先分析了一个 50 位急性淋巴细胞性白血病病人的数据, 数据来源于薛毅等 [25]. 在入院治疗时, 研究测得了病人外辕血中的细胞数 (Cell), 单位为千个/mm^3; 淋巴结浸润等级 (Lymph), 分为 0, 1, 2, 3 级; 出院后有无巩固治疗 (Consolidation), 0 表示无巩固治疗, 1 表示有巩固治疗. 因变量 Y 为病人的生存时间, $Y=0$ 表示生存时间在 1 年以内, $Y=1$ 表示生存时间在 1 年或 1 年以上.

首先, 对淋巴结浸润等级 (Lymph) 以及出院后有无巩固治疗 (Consolidation) 变量的频数表 (表 2) 和条形图 (图 1) 分析发现, 出院后有无巩固治疗对病人生存时间影响显著, 即有巩固治疗的病人生存时间比没有巩固治疗的病人生存时间更长. 而淋巴结浸润等级变量对病人的生存时间的影响没有明显的规律可循.

表 2: 急性淋巴细胞性白血病病人数据的描述性统计表

		Y=0	Y=1			Y=0	Y=1
Lymph	0	18	17	Consolidation	0	22	3
	2	12	3		1	8	17

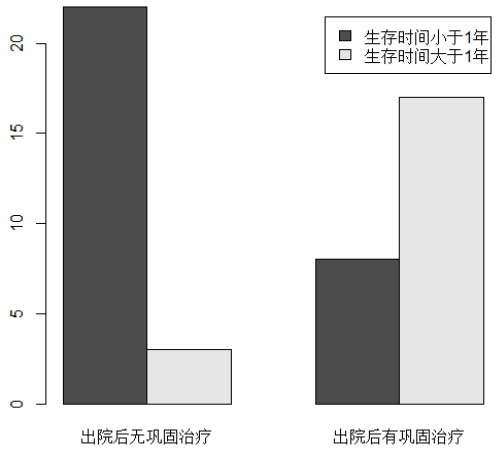

图 1: 出院后生存时间的条形图

基于上述影响因素, 在如下 Logistic 模型框架下:

$$\log \frac{P(Y=1)}{1+P(Y=1)} = \theta_0 + \theta_1 \,\mathrm{Cell} + \theta_2 \,\mathrm{Lymph} + \theta_3 \,\mathrm{Consolidation},$$

对数据进行了进一步的分析, 应用 QLB 算法计算回归参数的极大似然估计 (Est), 估计的标

准差 (SE) 仍由 Bootstrap 方法求得, Bootstrap 次数设为 1500 次, 进而得到参数的 95% 置信区间 (CI), 以及回归参数显著性检验的 p 值 (p-value). 数据分析结果请见表 3. 由表中结果可得, 出院后有无巩固治疗对病人生存时间有显著性影响 (Est$=2.8304$, p-value<0.001), 表明出院后接受巩固治疗能显著延长病人的生存时间. 而外辕血中的细胞数与淋巴结浸润等级对病人生存时间无显著性影响.

表 3: 急性淋巴细胞性白血病病人数据的分析结果

	Est	SE	95% CI	p-value
Intercept	-1.6965	0.7212	$(-2.1830,\ -1.2101)$	0.019^*
Cell	0.0023	0.0392	$(-0.0241,\ 0.0288)$	0.953
Lymph	-0.7922	0.7529	$(-1.3000,\ -0.2840)$	0.293
Consolidation	2.8304	0.8374	$(2.2656,\ 3.3952)$	$<0.001^*$

实例 2 (教职工升职情况数据)

这里, 我们分析一个教职工升职情况的调查数据, 数据来源于王静龙等 [26]. 数据给出了某校 5221 位教职工升职情况以及性别 (Sex)、工龄 (Age)、文化程度 (Education) 三个相关影响因素. 工龄取值为 1, 2, 3, 4, 分别表示工龄小于 5 年, $6-15$ 年, $16-29$ 年以及大于等于 30 年; 文化程度取值为 1, 2, 3, 4, 分别表示中学、大学、硕士、博士; 性别变量取值 0, 1, 其中 0 表示女性, 1 表示男性. 因变量 Y 取值 0, 1, 其中 0 表示无升职, 1 表示升职. 通过频率表和条形图可以看出: 随着工龄的增加, 或者教育程度的提高, 升职的人数逐渐增加 (表 4, 图 2–3). 因此可初步表明升职与否与工龄和教育程度有着明显相关性.

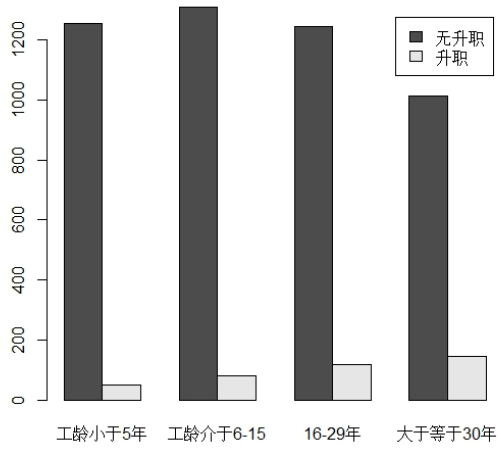

图 2: 不同工龄的升职统计.

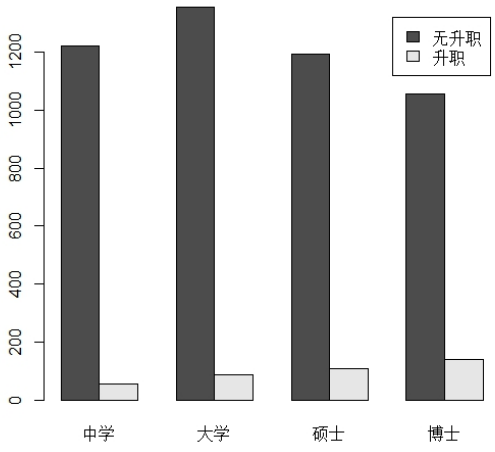

图 3: 不同文化程度的升职统计.

表 4: 教职工升职情况数据的描述性统计表

		$Y=0$	$Y=1$			$Y=0$	$Y=1$
Age	1	1256	49	Education	1	1221	57
	2	1309	82		2	1355	87
	3	1246	118		3	1193	110
	4	1015	146		4	1057	141

进一步, 建立升职与否与上述协变量之间的 Logistic 回归模型如下:

$$\log \frac{P(Y=1)}{1+P(Y=1)} = \theta_0 + \theta_1 \,\mathrm{Sex} + \theta_2 \,\mathrm{Age} + \theta_3 \,\mathrm{Education}.$$

应用 QLB 算法, 设定 Bootstrap 次数为 500 次, 所得数据分析结果请见表 5, 结果表明: 工龄 (Est= 0.5053, p-value< 0.001) 与文化程度 (Est= 0.4385, p-value< 0.001) 对升职有显著性影响, 特别地, 工龄越高, 文化程度越高, 越有利于升职. 结果同时也表明了性别对升职没有显著性影响.

表 5: 教职工升职情况数据的分析结果

	Est	SE	95%CI	p-value
Intercept	−5.0674	0.2442	(−5.2321, −4.9027)	$< 0.001^*$
Sex	0.1104	0.1105	(0.0359, 0.1850)	0.318
Age	0.5053	0.0508	(0.4711, 0.5395)	$< 0.001^*$
Education	0.4385	0.0538	(0.4023, 0.4748)	$< 0.001^*$

实例 3 (农村居民健康行为数据)

最后, 我们讨论一个农村居民健康行为及其影响因素的研究数据, 数据来源于方积乾等 [27]. 在这项调查研究中, 随机调查了 9760 名 15 − 55 岁的居民, 其中吸烟男性有 2229 人, 过去 1 年内有过戒烟行为的有 393 人. 此项研究的主要目的是分析和寻找对 “戒烟” 这一行为有影响的相关因素. 研究中对多个指标进行合并得到了主要影响因素如下: 年龄 X_1(岁), 家庭人均年实际收入 X_2(元), 婚姻状况 X_3(未婚 = 0, 已婚 = 1, 离婚丧偶 = 2), 职业 X_4(学生、家务及无业者 = 0, 农业劳动 = 1, 城乡农民工 = 2, 干部职工等其他 = 3), 受教育程度 X_5(文盲或半文盲 = 0, 小学 = 1, 初中 = 2, 高中及以上 = 3), 饮酒 X_6(不饮酒 = 0, 饮酒 = 1), 主动获取保健知识 X_7(不经常获取 = 0, 经常获取 = 1), 开始吸烟年龄 X_8(岁), 常去公共场所有无禁烟规定 X_9(无 = 0, 有 = 1), 家庭住室有无吸烟限制 X_{10}(无 = 0, 有 = 1), 有无经医生诊断的慢性病 X_{11}(无 = 0, 有 = 1), 自我健康总体评价 X_{12}(一般或较差 = 0, 较好 = 1, 很好 = 2). 响应变量 Y 表示过去 1 年内是否有过戒烟行为 (未曾戒烟 = 0, 曾戒烟 = 1). 进一步, 对原始数据中年龄 X_1 以及开始吸烟年龄 X_8 进行处理: 对年龄 X_1, 年龄低于 40 岁的赋值为 0, 大

于 40 岁的赋值为 1; 对开始吸烟年龄 X_8, 小于 25 岁赋值为 0, 大于 25 岁赋值为 1. 对收入 X_2, 做中心标准化处理.

表 6: 农村居民健康行为数据的分析结果

	Est	SE	95% CI	p-value
Intercept	−1.9801	0.2881	(−2.1744, −1.7858)	$< 0.001^*$
X_1	−0.3427	0.1260	(−0.4277, −0.2578)	0.007^*
X_2	0.0108	0.0551	(−0.0264, 0.0479)	0.845
X_3	0.3281	0.2011	(0.1925, 0.4637)	0.103
X_4	0.1019	0.0839	(0.0453, 0.1584)	0.225
X_5	0.1739	0.0787	(0.1208, 0.2269)	0.027^*
X_6	0.0671	0.1132	(−0.0092, 0.1435)	0.553
X_7	0.1156	0.1198	(0.0348, 0.1965)	0.335
X_8	0.2957	0.1466	(0.1968, 0.3946)	0.044^*
X_9	−0.0965	0.1265	(−0.1818, −0.0112)	0.446
X_{10}	0.7519	0.2731	(0.5676, 0.9361)	0.006^*
X_{11}	0.4368	0.1908	(0.3081, 0.5655)	0.022^*
X_{12}	−0.3027	0.0848	(−0.3599, −0.2455)	$< 0.001^*$

研究上述所有潜在影响因素与戒烟与否之间的关系, 建立相应 Logistic 模型, 应用 QLB 算法并设定 Bootstrap 次数为 1500 次, 所得数据分析结果请见表 6. 结果表明, 年龄超过 40 岁的男性更难戒烟, 而开始吸烟年龄大于 25 岁的男性更倾向于戒烟. 受教育程度越高, 家庭住室有吸烟限制、经医生诊断患有慢性病, 以及自我健康总体评价较低的男性更倾向于戒烟.

参 考 文 献

[1] Wang Baobin, Dai Jineng. Convergence rates of stochastic gradient descent methonds[J]. J. Math., 2012, 32(1): 75–78.

[2] Böhing D, Lindsay B G. Monotonicity of quadratic approximation algorithms[J]. Annals Institute Stati. Mathe., 1988, 40(4): 641–663.

[3] Huber P J. Robust statistics[M]. New York: Wiley, 1981.

[4] Heiser W J. Correspondence analysis with least absolute residuals[J]. Comput. Stati. Data Anal., 1987, 5(4): 337–356.

[5] Bijleveld C, De Leeuw J. Fitting longitudinal reduced-rank regression models by alternating least squares[J]. Psychometrika, 1991, 56(3): 433–447.

[6] Kiers H A L, Ten Berge J. Minimization of a class of matrix trace functions by means of refined majorization[J]. Psychometrika, 1992, 57(3): 371–382.

[7] Kiers H A L. Setting up alternating least squares and iterative majorization algorithms for solving various matrix optimization problems[J]. Comput. Stati. & Data Anal., 2002, 41(1): 157–170.

[8] De Pierro A R. A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography[J]. IEEE Trans. Medical Imag., 1995, 14(1): 132–137.

[9] Lange K, Fessler J A. Globally convergent algorithms for maximum a posteriori transmission tomography[J]. IEEE Trans. Image Proc., 1994, 4(10): 1430–1438.

[10] Hunter D R, Lange K. Quantile regression via an MM algorithm[J]. J. Comput. Graph. Stati., 2000, 9(1): 60–77.

[11] Hunter D R, Lange K. Computing estimates in the proportional odds model[J]. Annals Institute Stati. Math., 2002, 54(1): 155–168.

[12] Sabatti C, Lange K. Genomewide motif identification using a dictionary model[J]. Proc. IEEE, 2002, 90(11): 1803–1810.

[13] Hunter D R. MM algorithms for generalized Bradley-Terry models[J]. Annals Stati., 2004, 32(1): 384–406.

[14] Hunter D R, Li R. Variable selection using MM algorithms[J]. Annals Stati., 2005, 33(4): 1617–1642.

[15] Groenen P J F, Nalbantov G, Bioch J C. Nonlinear support vector machines through iterative majorization and I-splines[M]. Berlin: Springer–Verlag, 2006.

[16] Lange K, Wu T T. An MM algorithm for multicategory vertex discriminant analysis[J]. J. Comput. Graph. Stati., 2008, 17(3): 527–544.

[17] De Leeuw J. Block relaxation algorithms in statistics[M]. Berlin: Springer-Verlag, 1994.

[18] Heiser W J. Convergent computing by iterative majorization: theory and applications in multidimensional data analysis[J]. Recent Advances Descriptive Multi. Anal., 1995: 157–189.

[19] Becker M P, Yang I, Lange K. EM algorithms without missing data[J]. Sati. Meth. Medical Research, 1997, 6(1): 38–54.

[20] Lange K, Hunter D R, Yang I. Optimization transfer using surrogate objective functions (with discussion)[J]. J. Comput. Graph. Stati., 2000, 9(1): 1–20.

[21] Hunter D R, Lange K. A tutorial on MM algorithms[J]. The American Statistician, 2004, 58(1): 30–37.

[22] Wu T T, Lange K. The MM alternative to EM[J]. Stati. Sci., 2010, 25(4): 492–505.

[23] Efron B. Bootstrap methods: another look at the jackknife[J]. Annals Stati., 1979, 7(1): 1–26.

[24] Efron B, Tibshirani R J. An introduction to the bootstrap[M]. Boca Raton: CRC, 1993.

[25] 薛毅, 陈立萍. 统计建模与 R 软件 (第一版)[M]. 北京: 清华大学出版社, 2007.

[26] 王静龙, 梁小筠. 定性数据分析 (第一版)[M]. 上海: 华东师范大学出版社, 2005.

[27] 方积乾. 生物医学研究的统计方法 (第一版)[M]. 北京: 高等教育出版社, 2007.

QUADRATIC LOWER-BOUND ALGORITHM FOR MAXIMUM LIKELIHOOD ESTIMATOR OF LOGISTIC REGRESSION ON PARAMETER AND ITS APPLICATION

WANG Jia, DING Jie-li

(*School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China*)

Abstract: In this paper, we study how to explore more stable numerical solution when parameters cannot be solved by using Newton-Raphson algorithm. By using the quadratic lower bound algorithm that Böhning & Linday has proposed in 1988, we construct a surrogate function for maximum likelihood function under Logistic regression model and the simulation results verify that quadratic lower bound algorithm is a reasonable algorithm of Newton-Raphson algorithm, which extend numerical method's application under Logistic regression model.

Keywords: minorization-maximization algorithm; Logistic regression model; quadratic lower-bound algorithm; maximum likelihood estimator; Newton-Raphson algorithm

2010 MR Subject Classification: 62J12; 65C20