SCHOOL OF OPERATIONS RESEARCH
AND INDUSTRIAL ENGINEERING
COLLEGE OF ENGINEERING
CORNELL UNIVERSITY
ITHACA, NEW YORK 14853

TECHNICAL REPORT NO. 662

June 1985
(Revised March 1986)

A NOTE ON ALBERT AND ANDERSON'S CONDITIONS FOR THE EXISTENCE OF MAXIMUM LIKELIHOOD ESTIMATES IN LOGISTIC REGRESSION MODELS*

By

Thomas J. Santner & Diane E. Duffy

*This research was partially supported by the Hospital for Special Surgery Bioengineering Program at Cornell University.

A NOTE ON ALBERT AND ANDERSON'S CONDITIONS FOR THE EXISTENCE OF MAXIMUM LIKELIHOOD ESTIMATES IN LOGISTIC REGRESSION MODELS

by Thomas J. Santner and Diane E. Duffy
School of Operations Research and Industrial Engineering
Cornell University, Ithaca, New York 14853, U.S.A.

SUMMARY

This note expands the paper by Albert and Anderson (1984) on the existence and uniqueness of maximum likelihood estimates in logistic regression models. Their three possible mutually exclusive data patterns: (i) overlap, (ii) complete separation, and (iii) quasi-separation are considered. The maximum likelihood estimate exists only in (i). Modifications of the statement and proofs of Albert and Anderson's results are given for (ii) and (iii). The identifiability for a more general model arising in the study of (iii) is discussed together with the maximization of the corresponding likelihood. A linear program is presented which determines whether data is of type (i), (ii), or (iii) and in the case of (ii) identifies Albert and Anderson's minimal set Q^m.

Key words: Complete separation; Log linear model; Logistic discrimination; Mixed linear program; Overlap; Quasiseparation.

1. INTRODUCTION

We assume the logistic regression model of Albert and Anderson (1984) and adopt their notation in toto; references to equations (1)-(18) and to Lemma 1 and Theorem 2 are to expressions in their paper. Equations in this note are numbered beginning with (19). A few key equations and notations are repeated here to facilitate reading. In particular the observed data are (x_i, Y_i), $i = 1, \ldots, n$ where

$x_i' = (1, x_{i1}, \ldots, x_{ip})$ and the $n \times (p+1)$ matrix X whose ith row is x_i' has full column rank $(p+1)$. The $\{Y_i : i = 1, \ldots, n\}$ are mutually independent random variables with Y_i having probability mass function:

$$P[Y_i = j | x_i] = \frac{\exp\{\alpha_j' x_i\}}{\sum_{t=0}^{g} \exp\{\alpha_t' x_i\}} \quad \text{for} \quad j = 0, 1, \ldots, g$$

for some $g \geq 1$. Here $\alpha_j' = (\alpha_{j0}, \ldots, \alpha_{jp}) \in \mathbb{R}^{p+1}$ for $1 \leq j \leq g$ and $\alpha_0 \equiv 0$. Let $\alpha' = (\alpha_1', \ldots, \alpha_g') \in \mathbb{R}^{\upsilon}$ where $\upsilon \equiv (p+1)g$. For each i, $1 \leq i \leq n$, let $j(i)$ denote the outcome j, $0 \leq j(i) \leq g$ such that $Y_i = j(i)$.

Albert and Anderson classify logistic regression data into three mutually exclusive and exhaustive categories: complete separation, quasiseparation and overlap. The data are completely separated if there is an $\alpha \in \mathbb{R}^{\upsilon}$ such that for every $i = 1, \ldots, n$ and $t \in \{0, \ldots, g\} \backslash \{j(i)\}$

$$(\alpha_{j(i)} - \alpha_t)' x_i > 0. \tag{19}$$

Let A^c denote the set of $\alpha \in \mathbb{R}^{\upsilon}$ satisfying (19). To define quasiseparated data, let A^q denote the set of $\alpha \in \mathbb{R}^{\upsilon}$ such that for all $i = 1, \ldots, n$ and $t \in \{0, \ldots, g\} \backslash \{j(i)\}$

$$(\alpha_{j(i)} - \alpha_t)' x_i \geq 0 \tag{20}$$

with equality for <u>at least</u> one (i,t) pair. The data are quasiseparated if there is a <u>nonzero</u> $\alpha \in A^q$. Our definition of quasiseparation differs slightly from that of Albert and Anderson who consider data quasiseparated when A^q is nonempty. Since $\alpha \equiv 0 \in A^q$ for all data sets by equation (20), all data sets are trivially

quasiseparated by their definition. Data which is neither completely nor quasicompletely separated is called overlapped. The maximum likelihood estimate of α exists only for overlapped data.

For quasiseparated data, §2 contains revised versions of Lemma 1 and Theorem 2. Comments on the identifiability of the reduced problem in (14) are made in §3. For overlapped data, we note that $\alpha = 0$ must be excluded from equation (15) otherwise it is trivially false. A linear program is formulated in §4 which classifies the data into one of the three categories and, for quasiseparated data, determines Q^m.

2. QUASICOMPLETE SEPARATION

We begin by noting that in Albert and Anderson's discussion of the special case $g = 2$, if $Q(\alpha)$ has $r \geq 1$ elements, then the r points corresponding to the rows of X^q generate an affine space of dimension $\text{rank}(X^q)-1$ rather than "a linear subspace of dimension $\leq p-1$." Lemma 1' and Theorem 2' are modified versions of Lemma 1 and Theorem 2. The symbol ϕ denotes the empty set throughout.

LEMMA 1'. If $A^c = \phi$ and $A^q \neq \{0\}$, then (i) and (ii) of Lemma 1 hold.

Lemma 1' differs from Lemma 1 in requiring the existence of a nonzero α in A^q. Part (i) of Lemma 1' is proved as in Albert and Anderson but their proof of (ii) requires modification.

Let $\alpha,\beta \in A^q$, $\lambda,\mu > 0$ and $\gamma = \lambda\alpha + \mu\beta$; Albert and Anderson show $\gamma \in A^q$ and $Q(\gamma) \subseteq Q(\alpha) \cap Q(\beta)$. They claim equality holds, but for $g \geq 3$, examples of strict containment can be constructed. Instead, induction shows that the finite intersection property holds

for $\{Q(\alpha)\colon \alpha \in A^q\}$, i.e. for integer $k \geq 1$ and $\alpha^1,\ldots,\alpha^k$ in A^q, $\cap\{Q(\alpha^j)\colon j = 1,\ldots,k\} \neq \phi$. The compactness of $\{1,\ldots,n\}$ under the power set topology gives $Q^m \equiv \cap\{Q(\alpha)\colon \alpha \in A^q\} \neq \phi$. There exists $\alpha^m \in A^q$ with $Q(\alpha^m) = Q^m$ for if not, an analogous argument to that above shows $\cap\{Q(\alpha)\backslash Q^m\colon \alpha \in A^q\} \neq \phi$ contradicting the definition of Q^m.

THEOREM 2'. If $A^c = \phi$ and $A^q \neq \{0\}$, then (i) and (ii) of Theorem 2 hold.

The hypothesis of Theorem 2' is a modification of that of Theorem 2. Albert and Anderson's proof of (i) and (ii) is based on the decomposition of $\alpha \in \mathbb{R}^v$ as $\alpha = \gamma+\beta$ with $\gamma \notin A^q$ and β the unique element of A^q closest to α. This decomposition requires A^q be closed and convex, thus 0 must be considered an element of A^q. Their claim that $\gamma \notin A^q$ holds if and only if $\alpha \notin A^q$, else $\gamma = 0 \in A^q$. Given $\alpha \in \mathbb{R}^v$, they set $\alpha(k) = \gamma+k\beta$ with γ, β as above, $k > 0$, and consider the limiting behavior of $\log L(X,\alpha(k))$ as $k \to \infty$. This argument fails to consider those $\alpha \notin A^q$ for which $\beta = 0$; that is, vectors not in the cone for which the nearest element of the cone is the zero vector.

Instead part (i) can be proved as below. Fix $\beta \in A^q\backslash\{0\}$. For any $\alpha \in \mathbb{R}^v$ let $\alpha*(k) = \alpha+k\beta$. Consideration of the limit as $k \to \infty$ of $\log L(X, \alpha*(k))$ shows that the maximum likelihood estimate of α does not exist. To prove (ii) define $T(\beta)$ to be the Cartesian product $X\{T_i(\beta)\colon i \in Q^m\}$ for $\beta \in A^q$. Then $T(\beta) \neq \phi$ for all $\beta \in A^q$ since $Q^m \subset Q(\beta)$. Let $T^m = \cap\{T(\beta)\colon \beta \in A^q\}$. Arguments similar to those in Lemma 1', give $T^m \neq \phi$ and there exists $\beta^m \in A^q$ with $T(\beta^m) = T^m$. For any $\alpha \in \mathbb{R}^v$ we have

$$\log L(X,\alpha) \le \sup_{0\le k<\infty} \log L(X,\alpha+k\beta^m)$$

$$= \sum_{i\in Q^m} \log[1/(1 + \sum_{t\in T_i^m} \exp\{(\alpha_t-\alpha_{j(i)})'x_i\}] \qquad (21)$$

Expression (21) includes the summand 1 in the denominator of each logarithm which was incorrectly omitted in (13). Taking the supremum

$$\sup_{\alpha\in\mathbb{R}^\nu} \log L(X,\alpha) = \sup_{\alpha\in\mathbb{R}^\nu} \sum_{i\in Q^m} \log[1/(1 + \sum_{t\in T_i^m} \exp\{(\alpha_t-\alpha_{j(i)})'x_i\})]. \qquad (22)$$

The right hand side of (22) is constant for all $\alpha \in A^q$, and, as is correct, includes this value in the maximization. Further, $\{\alpha: (\alpha_t-\alpha_{j(i)})'x_i < 0$ for all $i \in Q^m$ and $t \in T_i^m\} = \phi$ so that (22) is less than 0, completing the proof.

3. REDUCED PROBLEMS

The right hand side of (22) appears to be a logistic log likelihood for data Y_i, $i \in Q^m$ with values in $S_i \equiv T_i^m \cup \{j(i)\}$. This reduced problem involves α_j only for $j \in G \equiv \cup\{S_i: i \in Q^m\}$. Abusing notation, let $\alpha \equiv (\alpha_j: j \in G) \in \mathbb{R}^r$, $r \le \nu$ be its parameter vector. This problem is identifiable i.e. there is a 1-1 correspondence between α and the joint distributions of Y_i, $i \in Q^m$, if and only if

$$\{\alpha \in \mathbb{R}^r: x_i'(\alpha_{j(i)}-\alpha_t) = 0 \text{ for all } i \in Q^m \text{ and all } t \in T_i^m\} = \{0\}. \qquad (23)$$

In the absence of (23), if a maximum likelihood estimator of α exists, it is nonunique.

Regardless of the validity of (23), Theorem 2' shows that (22) is less than zero. Hence the problem cannot be completely separated in the sense that for some α, (19) holds for all $i \in Q^m$ and $t \in T_i^m$.

It can also be shown that it is not quasiseparated in the sense there is a nonzero α for which (20) holds for all $i \in Q^m$ and $t \in T_i^m$.

Condition (23) cannot be readily established in general. Note that (23) implies full column rank p+1 for the matrices X^j, $j \in G$, where X^j has rows x_i', $j \in S_i$. This is both intuitively reasonable and implied by existence of the maximum likelihood estimator in the reduced problem. Assume rank p+1 for the X^j. If $1 \in \cap\{S_i: i \in Q^m\}$ then $\alpha_1 \equiv 0$ guarantees (23) and the α_j, $j \in G$ are easily interpretable in the original problem. In general, $\alpha_1 \equiv 0$ is not sufficient to guarantee (23) as 1 may not be in G and additional judiciously chosen α_j must be set to zero. The reduced problem may separate into subproblems with disjoint sets of outcomes requiring at least one α_j set to zero in each subproblem.

4. SEPARATION IDENTIFICATION

The linear program below classifies data as (i) completely separated, (ii) quasiseparated or (iii) overlapped and in case (ii) identifies Q^m.

Define the $n(g-1) \times v$ block matrix $\overline{X}$ to have $(g-1) \times v$ dimensional i^{th} block, $1 \le i \le n$, as below:

$$\begin{matrix} 1 \text{ row} \\ (g-2) \text{ rows} \end{matrix} \begin{cases} \left[x_i' Q_{j(i)} \right] \\ \left[\begin{matrix} \vdots \\ x_i'[Q_{j(i)} - Q_t] \end{matrix} \right] \end{cases}$$

where $t \in \{1,\ldots,g-1\}\backslash\{j(i)\}$ if $j(i) < g$, or

$$(g-1) \text{ rows } \{ [x_i'[-Q_t]]$$

where $t \in \{1,\ldots,g-1\}$ if $j(i) = g$. Here Q_t, $1 \le t < g$, is the $(p+1)\times v$ matrix of 0's and 1's such that $Q_t \alpha = \alpha_t$. Observe that $A^q = \{\alpha \in \mathbb{R}^v: \overline{X}\alpha \ge 0\}$. Consider the (mixed) linear program:

$$\max \sum_{i=1}^{n} z_i$$

$$\text{subject to } [\bar{X}-I_{n(g-1)}][\tfrac{\alpha}{s}] = 0$$

$$w_j \le \frac{s_j}{m}, \quad j = 1,\ldots,n(g-1)$$

$$z_i(g - 3/2) \le \Sigma_{j=(i-1)(g-1)+1}^{i(g-1)} w_j, \quad i = 1,\ldots,n$$

$$\alpha \text{ arbitrary}$$

$$s \ge 0$$

$$w_j,\ z_i \in \{0,1\},\ i = 1,\ldots,n \text{ and } j = 1,\ldots,n(g-1).$$

Here $I_{n(g-1)}$ is the $n(g-1)\times n(g-1)$ identity matrix and $m > 0$ is chosen so that $m \le s_j$ holds for $s_j > 0$, $j = 1,\ldots,n(g-1)$, (to the desired degree of precision).

This program is always feasible ($\alpha \equiv 0$, $s \equiv 0$, $w \equiv 0$, $z \equiv 0$). A component of s is positive if and only if the corresponding inequality of $\bar{X}\alpha \ge 0$ is satisfied <u>strictly</u>. Also, $z_i = 1$ if and only if Y_i is completely separated. The objective function maximizes the cardinality of the complement of Q^m. If $A^c \ne \phi$ the program terminates with an optimal value of n and at $\alpha \in A^c$. If $A^c = \phi$ and $A^q\backslash\{0\} \ne \phi$ the program terminates with a positive optimal value less than n, with $Q^m = \{i: z_i = 0\}$ and with α satisfying $Q(\alpha) = Q^m$. If $A^c = \phi$ and $A^q = \{0\}$, the program terminates with optimal value 0. For quasiseparated data, a (mixed) linear program similar to this can be formulated to determine T^m and $\beta \in A^q$ such that $T(\beta) = T^m$.

REFERENCES

Albert, A. and Anderson, J.A. (1984). On the existence of maximum likelihood estimates in logistic regression models. <u>Biometrika</u> 71, 1-10.