Haberman (1974, p. 37) proves a very general theorem on necessary and sufficient conditions for the maximum likelihood estimate to exist. In his terminology, existence means finiteness of the solution. He also demonstrates that for most models, if the maximum likelihood solution exists, it is unique, as a result of the concavity of the likelihood function. Wedderburn (1976) presents a series of sufficient, but not necessary, conditions for the existence, uniqueness and location, on the boundary or not, of maximum likelihood estimates for the parameters of the generalized linear model. Unfortunately, for multinomial logistic regression these conditions are not satisfied in many practically important cases, such as completely and quasicompletely separated data configurations.

Though powerful, both Haberman's and Wedderburn's results fall short of providing conditions for nonexistence of maximum likelihood estimates and general practical procedures for identifying infinite parameter values, except in problems with a special structure.

Recently, Silvapulle (1981) made a first step in that direction by proving that a certain degree of overlap is a necessary and sufficient condition for the existence of the maximum likelihood estimates for the binomial response model.

In the present paper we prove existence theorems that go beyond previously reported results, by classifying the different types of data sets. We also suggest ways of recognizing those data configurations that lead to infinite estimates, thus avoiding unnecessary iterations and giving insight into why infinite estimates occur.

2. THE LIKELIHOOD FUNCTION

It is convenient at this stage to present the notation in terms of the logistic model, but most of the results apply to the probit and many other similar models for binary or multinomial regression.

It is assumed that a sample of n independent points is available. At each point, observations are made on the variables (x, H) where $x^{\mathrm{T}} = (x_0, \ldots, x_p)$ is a vector of variables which may be discrete or continuous, and H is a variable which takes values $H_1, \ldots, H_g$, the indicator of group membership. For convenience $x_0 \equiv 1$. The logistic approach to regression and discrimination is to assume that the conditional probabilities have the extended logistic form (Anderson, 1972):

$$\begin{aligned} \operatorname{pr}(H_s \mid x) &= \exp(\alpha_s^{\mathrm{T}} x) \operatorname{pr}(H_g \mid x) \quad (s = 1, \ldots, g-1), \\ \operatorname{pr}(H_g \mid x) &= 1 \Big/ \left\{ \sum_{s=1}^{g} \exp(\alpha_s^{\mathrm{T}} x) \right\}, \end{aligned} \tag{1}$$

where

$$\alpha_s^{\mathrm{T}} = (\alpha_{s0}, \ldots, \alpha_{sp}) \quad (s = 1, \ldots, g-1), \quad \alpha_g^{\mathrm{T}} = 0.$$

For discrimination, the simplest classification rule is to allocate the observation x to the group H_s $(s = 1, \ldots, g)$ if and only if

$$(\alpha_s - \alpha_t)^{\mathrm{T}} x \geqslant 0 \quad (t = 1, \ldots, g). \tag{2}$$

Let x_i be the column vector of observations for the ith point $(i = 1, \ldots, n)$ and denote X the $n \times (p+1)$ matrix with the x_i^{T} as rows. We assume that X is of full rank, $p+1$. Denote by E_s the set of row identifiers of X for observations from H_s, and by n_s the number of observations from H_s. Then $E = \cup E_s$, the set of integers $\{1, \ldots, n\}$. Let $\alpha^{\mathrm{T}} = (\alpha_1^{\mathrm{T}}, \ldots, \alpha_{g-1}^{\mathrm{T}})$

represent the $v = (p+1)(g-1)$ vector of unknown parameters. It is also assumed that $n > v$. The subsequent results hold whether sampling is from the mixture distribution (x, H) or from either the conditional distributions $H \mid x$ or $x \mid H$. In all cases, the function to be maximized is

$$\log L(X, \alpha) = \sum_{j=1}^{g} \sum_{i \in E_j} \log \left[1 \Big/ \sum_{t=1}^{g} \exp \{(\alpha_t - \alpha_j)^{\mathrm{T}} x_i\} \right]. \tag{3}$$

For $H \mid x$ sampling this is the log likelihood; for (x, H) sampling it the conditional log likelihood; for $x \mid H$ sampling it is a quasi log likelihood function with asymptotic properties similar to maximum likelihood. See Anderson & Blair (1982) and Prentice & Pyke (1979) for a discussion of these topics.

3. Maximum likelihood solution: existence and nonexistence

3·1. *General comments*

The problem of maximizing the log likelihood function (3) will be examined by considering the possible configurations of the n sample points in the observation space R^p. The possible configurations fall essentially into three mutually exclusive and exhaustive categories, complete separation, quasicomplete separation and overlap. In each case, we concentrate on the finiteness of the maximum likelihood solution, and on the actual value of the maximum of the likelihood function (3). Throughout the paper nonexistence of the maximum likelihood estimate means absence of a finite maximum.

3·2. *Complete separation*

We say there is complete separation in the sample points, if there exists a vector $\alpha \in R^v$, such that for all $i \in E_j$ and for $j, t = 1, \ldots, g$ $(j \neq t)$

$$(\alpha_j - \alpha_t)^{\mathrm{T}} x_i > 0. \tag{4}$$

In other words, there is a vector α that correctly allocates all observations to their group. For $g = 2$, as in Fig. 1a, complete separation means that $\alpha_1^{\mathrm{T}} x_i > 0$, $i \in E_1$ and $\alpha_1^{\mathrm{T}} x_i < 0$, $i \in E_2$.

Denote the set of all vectors α satisfying the above relationship (4) by A^c. Then A^c is a convex cone in R^v. Further, if there is one $\alpha \in A^c$, then $\alpha + \Delta \in A^c$, where $\Delta \neq k\alpha$ may be chosen as small as necessary to satisfy (4). Hence, if A^c contains one ray $k\alpha$, it will contain a continuum of rays. We have the following theorem.

Theorem 1. *If there is complete separation of the data points, the maximum likelihood estimate $\hat{\alpha}$ does not exist, and*

$$\max_{\alpha \in R^v} L(X, \alpha) = 1.$$

Proof. At $\alpha(k) = k\alpha$, $\alpha \in A^c$ and $k > 0$, the log likelihood function (3) becomes

$$\log L(X, k\alpha) = \sum_{j=1}^{g} \sum_{i \in E_j} \log \left[1 \Big/ \sum_{t=1}^{g} \exp \{-k(\alpha_j - \alpha_t)^{\mathrm{T}} x_i\} \right]. \tag{5}$$

Consider the behaviour of (5) as k tends to infinity. Inequalities (4) still hold; hence all exponential terms in (5) tend to zero except the one in each sum over t where $t = j$, which has value unity for all k. Thus, the expression (5) tends to zero, its absolute maximum, as k tends to infinity. We conclude that the absolute maximum of the likelihood function is attained at infinity on the boundary of the parameter space. Hence, $\hat{\alpha}$ is infinite.

Note that the absolute maximum for $\log L(X, \alpha)$ is also achieved by going to infinity along any ray in A^c. Thus there is a continuum of points on the boundary of the parameter space where the absolute maximum is attained, since there is a continuum of rays in A^c.

3·3. *Quasicomplete separation*

If the data set X is not completely separable, another separation concept is needed. The vector $\alpha \in R^v$ is said to give quasicomplete separation of the sample points if for all $i \in E_j$ and for $j, t = 1, \ldots, g$ $(j \neq t)$

$$(\alpha_j - \alpha_t)^{\mathrm{T}} x_i \geqslant 0, \tag{6}$$

with equality for at least one (i, j, t) triplet. We denote by $j(i)$ the value of j for which $i \in E_j$. Let $Q(\alpha)$ denote the set of index values, $i \in E$, satisfying the equality in (6). The corresponding points x_i are said to be quasiseparated with respect to α.

For example, consider two groups, $g = 2$. Assume that the set $Q(\alpha)$ contains $r \neq 0$ elements and denote by X^q the $r \times (p+1)$ matrix of quasiseparated observations. Then by definition $X^q \alpha_1 = 0$. It follows that the rows of X^q are linearly dependent, so that the r points belong to a linear subspace of dimension $d \leqslant p-1$. Actually, $d = \operatorname{rank}(X^q) - 1$. It is obvious that, if $\operatorname{rank}(X^q) = p$, the hyperplane of separation is unique, as in Fig. 1b, otherwise there is a continuum of such planes containing the r points, see Fig. 1b′.

Denote the set of all vectors α satisfying (6) by A^q. We have the following lemma and theorem.

LEMMA 1. *For any data set X for which A^q is not empty*

(i) *A^q is a convex cone;*

(ii) *there is a minimal set Q^m such that $Q^m \subset Q(\alpha)$ for all $\alpha \in A^q$, and $A^{mq} = \{\alpha: \alpha \in A^q$ and $Q(\alpha) = Q^m\}$ is not empty.*

Proof. (i) If α satisfies (6) with equality for at least one (i, j, t) triplet, so does $k\alpha$ for $k > 0$. If α and $\beta \in A^q$, let $\gamma = \lambda\alpha + \mu\beta$ for $\lambda, \mu > 0$ and $\lambda + \mu = 1$. Then γ satisfies also the inequality in (6). Further, γ must satisfy the equality in (6) for at least one (i, j, t) triplet, otherwise γ gives complete separation and this has been excluded by convention. Hence $\gamma \in A^q$.

(ii) In the above notation, let $i \in Q(\gamma)$, then x_i satisfies

$$\lambda(\alpha_j - \alpha_t)^{\mathrm{T}} x_i + \mu(\beta_j - \beta_t)^{\mathrm{T}} x_i = 0 \tag{7}$$

for at least one pair (j, t), $j \neq t$, j such that $i \in E_j$. Since $\lambda, \mu > 0$, and α and $\beta \in A^q$, each of the two terms in the above equations are nonnegative. Hence, they must both be zero which implies that $i \in Q(\alpha)$ and $i \in Q(\beta)$. Hence $Q(\gamma) = Q(\alpha) \cap Q(\beta)$. The above result shows that the class $E^q = \{Q(\alpha): \alpha \in A^q\}$ is closed under intersection. Define

$$Q^m = \bigcap_{\alpha \in A^q} Q(\alpha). \tag{8}$$

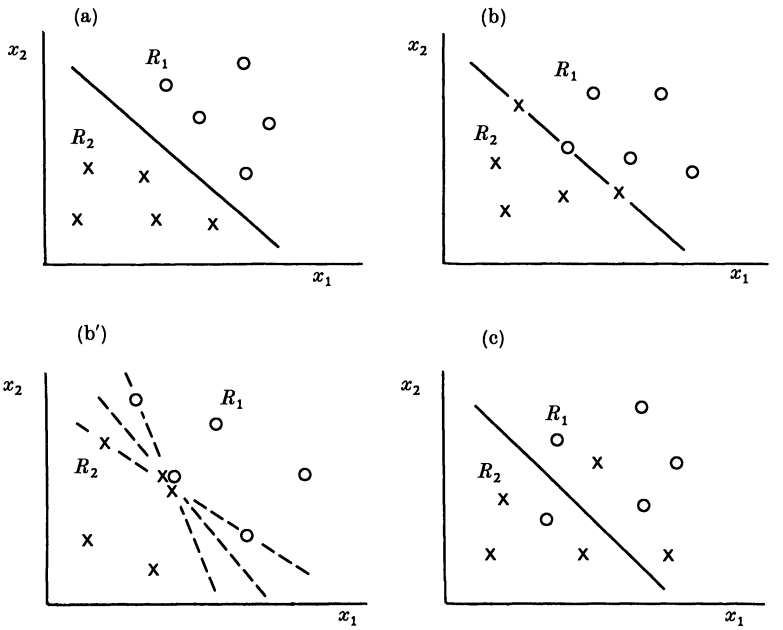

Fig. 1. Possible configurations of sample points in the case of two variables, x_1 and x_2, and two groups, H_1, shown by circles, and H_2, shown by crosses. Regions R_1 and R_2 define corresponding allocation rule. (a) Complete separation. (b) Quasicomplete separation $\rho(X^q) = 2$. (b′) Quasicomplete separation $\rho(X^q) = 1$; at the point of intersection of the lines, there are three observations, one from H_1, and two from H_2. (c) Overlap.

Then by the closure property and the finiteness of each $Q(\alpha)$, there is $\alpha \in A^q$, $Q(\alpha) = Q^m \in E^q$ and A^{mq} is not empty.

THEOREM 2. *If there is quasicomplete separation of the data points* ($A^c = \phi$, $A^q \neq \phi$), *then*
(i) *the maximum likelihood estimate* $\hat{\alpha}$ *does not exist;*

(ii) $$\max_{\alpha \in R^v} L(X, \alpha) = \max_{\gamma \in \bar{A}^q} L(X^{mq}, \gamma) < 1,$$

where X^{mq} *is the matrix of quasiseparated points corresponding to the minimal set* Q^m.

Proof. (i) For any parameter vector $\alpha \in R^v$, there is a nearest element β of the convex cone A^q, which is unique. We can write $\alpha = \gamma + \beta$, where $\gamma \in \bar{A}^q$. We consider the sequence of vectors $\alpha(k) = \gamma + k\beta$ ($k > 0$) and show that the likelihood function is increasing with k. Let $\bar{Q}(\beta)$ be the set of indices i of vectors x_i which are not quasicompletely separated, satisfying (6) with the strict inequality. Denote $Q(\beta)$ and $\bar{Q}(\beta)$ by Q and $\bar{Q}$, temporarily. The log likelihood function (3) can be written as

$$\log L\{X, \alpha(k)\} = \sum_{i \in \bar{Q}} \log U_i + \sum_{i \in Q} \log U_i, \tag{9}$$

where

$$U_i = 1 \Big/ \sum_{t=1}^{g} \exp\{(\gamma_t - \gamma_{j(i)})^{\mathrm{T}} x_i + k(\beta_t - \beta_{j(i)})^{\mathrm{T}} x_i\}. \tag{10}$$

It is easily seen that for fixed γ, U_i is a monotone increasing function of k for $i \in Q$ and so is

$$l(Q) = \sum_{i \in Q} \log U_i.$$

Further U_i is a strictly monotone increasing function of k for $i \in \bar{Q}$ and so is

$$l(\bar{Q}) = \sum_{i \in \bar{Q}} \log U_i.$$

Note that if $\bar{Q}$ is empty, then $l(Q)$ is strictly monotone increasing since there exists at least one triplet (i, j, t) satisfying the inequality in (6) for $i \in Q$ and the corresponding U_i is strictly monotone. The inequality follows because we have assumed in § 2 that X is of full rank. Thus whether $\bar{Q}$ is empty or not, $\log L\{X, \alpha(k)\} = l(\bar{Q}) + l(Q)$ is a strictly increasing function of k for every $\alpha \in R^v$. Let $\alpha^b = \lim \alpha(k)$ as $k \to \infty$; then $L(X, \alpha) = L\{X, \alpha(k=1)\} < L(X, \alpha^b)$. The maximum likelihood estimate $\hat{\alpha}$ is thus at infinity on the boundary of R^v.

(ii) By an argument similar to that of Theorem 1, $\lim l(\bar{Q}) = 0$ as $k \to \infty$. To investigate the limiting behaviour of $l(Q)$, define for each x_i, $i \in Q$ the set

$$T_i = \{t\colon (\beta_{j(i)} - \beta_t)\, x_i = 0;\ t \neq j\}. \tag{11}$$

No T_i is empty by quasicomplete separation, and T_i depends on $\beta \in A^q$. Let $T = \{T_i\colon i \in Q\}$. It follows that $(\beta_{j(i)} - \beta_t)^{\mathrm{T}} x_i > 0$ for $i \in Q$ and $t \notin T_i$, and hence, as $k \to \infty$ all the exponential terms in U_i for $i \in Q$ and $t \notin T_i$ tend to zero. For $t \in T_i$, the equality in (11) is satisfied and only the terms in γ remain. Hence

$$\log L(X, \alpha^b) = \lim_{k \to \infty} l(Q) = \sum_{i \in Q} \log [1/\sum_{t \in T_i} \exp\{(\gamma_t - \gamma_{j(i)})^{\mathrm{T}} x_i\}]. \tag{12}$$

This last expression depends on $\gamma \in \bar{A}^q$, and $\beta \in A^q$ only through the sets Q and T. In other words, summation is now restricted to the matrix X^q of quasiseparated points with respect to $\beta \in A^q$. Hence the maximum of $\log L(X, \alpha)$ as α varies in R^v is given by the maximum of $\log L(X, \alpha^b)$ over γ and sets $Q(\beta)$, $\beta \in A^q$. From Lemma 1, there is a minimal set Q^m of quasiseparated points to which also corresponds a minimal set $T^m = \cap\{T(\beta)\colon \beta \in A^q \text{ and } Q(\beta) = Q^m\}$. For fixed γ,

$$\log L(X, \alpha^b) \leqslant \sum_{i \in Q^m} \log [1/\sum_{t \in T_i^m} \exp\{(\gamma_t - \gamma_{j(i)})^{\mathrm{T}} x_i\}], \tag{13}$$

since there are least exponential terms in the denominators for the choice $T = T^m$, and since for $Q = Q^m$ there are fewer terms in the sum over i than for any other choice of Q. Since each term in (12) is nonpositive, it follows immediately that

$$\max_{\alpha \in R^v} L(X, \alpha) = \max_{\gamma \in \bar{A}^q} L(X^{mq}, \gamma). \tag{14}$$

Note that the sample vectors x_i, $i \in Q^m$ cannot be separated, completely or quasicompletely, with respect to γ. This follows from their definition. The maximizing of quantities like $L(X^{mq}, \gamma)$, appropriately reparameterized, is dealt with in the next section. They are shown to have unique maxima, achieved at interior points of the parameter space.

In sum, we have shown that the value of the maximum of $L(X, \alpha)$, $\alpha \in R^v$ is equal to the maximum of the likelihood function restricted to the minimal set of quasiseparated points X^{mq} with respect to $\gamma \in R^v - A^q$. The maximum of $L(X, \alpha)$ is attained on the boundary of R^v at a unique point, if A^{mq} contains a single ray, or at a continuum of boundary points if not.

3·4. Overlap

If neither complete nor quasicomplete separation exists in the sample points, these necessarily overlap in the sense that for any vector $\alpha \in R^v$, there exists a triplet (i, j, t), where $j, t \in \{1, \ldots, g\}$, $j \neq t$, $i \in E_j$ and

$$(\alpha_j - \alpha_t)^{\mathrm{T}} x_i < 0. \tag{15}$$

This situation is illustrated in Fig. 1c for two groups.

Theorem 3, *If there is overlap of the data points, the maximum likelihood estimate $\hat{\alpha}$ exists and is unique.*

Proof. The proof of this theorem is given both by Silvapulle (1981) for binomial response models and also by Haberman (1974, Chapter 8) in his work on log linear models, which includes binomial logistic regression. In his Theorem 8·1 ($g = 2$), λ plays the role of $X\alpha$, and his second condition is either the complete or quasicomplete separation condition.

For multinomial logistic regression model, the proof follows by routine use of Haberman's arguments. The log likelihood has limit $-\infty$ at infinity and is strictly concave. Note that Wedderburn's (1976) results are still not generally applicable here. For example, in observational studies it is quite likely that there is at most one observation at each point x of the observation space. In this case, the prerequisites for Wedderburn's theorems are not satisfied and his results are not applicable.

4. Separation detection

In general, the estimates $\hat{\alpha}$ which maximize the likelihood (3) do not have explicit forms. Iterative solutions are required and Newton–Raphson of quasi-Newton optimization methods usually work well with starting values of zero for the parameters, particularly if the elements of x, excluding x_0, are transformed to zero mean and unit standard deviation.

We have demonstrated in § 3 that computational difficulties in finding the maximum are to be expected when either complete or quasicomplete separation occurs in the observed samples. Without special checks, the standard numerical optimization techniques have no way of detecting this and carry on iterating until the bound on the number of iterations is reached.

The question of categorization of data sets into completely separated, quasicompletely separated or overlapped can be approached in two distinct ways, algebraic or empirical. These will now be considered in turn.

The algebraic approach uses ideas of linear programming; there is related unpublished work in the thesis of J. Burridge. Suppose that there are two groups, then there is complete or quasicomplete separation if there exists $\alpha_1 \neq 0$, such that

$$X^* \alpha_1 \leqslant 0, \tag{16}$$

where X^* is the $n \times (p+1)$ matrix with rows $-x_i$, $i \in E_1$, and x_i, $i \in E_2$. For quasicomplete separation, the equality must hold for at least one value of i. If (16) cannot be satisfied with $\alpha_1 \neq 0$, then the data set is overlapped.

The equation (16) may be rewritten as

$$(X^*, I_n)(\alpha_1, t) = 0, \tag{17}$$

where I_n is the $n \times n$ identity matrix and t is a row vector of n slack variables. There is: (a) complete separation if there exists a solution of (17) with all $t_i > 0$ $(i = 1, \ldots, n)$; (b) quasicomplete separation if there is a solution with $t_i \geqslant 0$ $(i = 1, \ldots, n)$, with equality for at least one value of i; (c) overlap if there is no solution satisfying (a) or (b) above, that is, all the solutions have some t_i positive and some negative.

Note that the case of collinearity, where $t_i = 0$ $(i = 1, \ldots, n)$, was excluded in §2 by taking X and hence X^* to be of full rank. In practice, this case can occur but can be recognized and dealt with using standard methods for collinearity in regression problems.

Enumerating the set of all solutions of (17) satisfying (a) or (b) is equivalent to finding the set of feasible solutions of a linear programming problem. The question of whether these sets are empty can be answered using adaptations of standard methods.

The above approach gives a complete, algebraic categorization of data sets into the three classes and can be extended in the obvious way to g (> 2) groups. Besides linear programming techniques, other more empirical methods are also of interest.

Quite a good empirical approach is available for complete separation. Anderson (1972) proved that any convergent method of maximizing the likelihood function (3) must yield a solution giving complete separation, if such a solution exists. Day & Kerridge (1967) proved this for $g = 2$. Hence, it is possible to insert a stopping rule in the iterative algorithm which is activated if complete separation is found. For the current iteration, denote by n_{jj} the number of observations from H_j correctly allocated with no ties, satisfying (4) for $j = 1, \ldots, g$. If $n_{jj} = n_j$ for all j, there is complete separation and the iterations cease. This check has been implemented in a series of programs available from the authors and has been found to be very effective.

For quasicomplete separation, the situation is more difficult but we know from Theorem 2(ii) that, at least for some, although not all, points, as the process diverges, the probability of belonging to the correct group rapidly grows to one. Therefore, at each iteration $t \geqslant T$, we look for the point x with the largest probability of correct allocation across the data set, and denote this maximum by $\mathrm{pr}^m(\alpha_t)$, then first print a warning message if for some suitable ε

$$\mathrm{pr}^m(\alpha_t) > \min\{1-\varepsilon, \mathrm{pr}^m(\alpha_{t-1})\}, \tag{18}$$

and then continue the iteration.

Inequality (18) states that for at least one observation x in the data set, the probability of correct allocation has become extremely close to 1. Now there are two possibilities for this. First, there is overlap in the data set and x simply is an atypical observation in its own group, far away from the mean; in this case the warning is unnecessary and the process, being allowed to continue, will stop whenever the maximum is reached. Secondly, there is quasicomplete separation in the data set, and x is among the completely separated points $(x \in \bar{Q})$; in this case the asymptotic dispersion matrix is unbounded. It is recommended that the program be rerun with all elements of the observation vectors standardized to zero mean and unit variance. Then the process can be stopped if any diagonal element of the dispersion matrix exceeds 10^3.

This seemingly cumbersome procedure has been extensively tested by the authors and has worked remarkably well in all cases tested. We recommend activation of the warning procedure only after several iterations, for example $T = 7$ or 8, in order to avoid early false warnings.

5. Discussion

The difficulties associated with complete and quasicomplete separation are small sample problems. With increasing sample size, the probability of observing a set of separated data points tends to zero, no matter what the sampling scheme. Complete separation may occur with any type of data but it is unlikely that quasicomplete separation will occur with truly continuous data.

The question of what to do if there is separation of the data points does not have a simple answer. In a prediction context where we wish to allocate further sample points x to the most likely group H_j, it may be sufficient to use the boundary value α^b of α giving the global maximum. Values x not falling on the quasiseparating hyperplanes are allocated using any ray $\gamma + k\beta$ which tends to α^b. Remaining points are allocated using the optimal γ; see Theorem 2.

If increasing the sample size is a possibility, it would be helpful to sample conditionally from the regions of quasiseparating hyperplanes, if this is possible. These regions can be specified in terms of x and it is possible sometimes to stratify the sampling to give extra weight to these regions. An extreme example is in bioassay, where the dose levels x can be fixed as desired. This is the $H \mid x$ sampling mentioned at the end of § 2.

The theorems proved here have been derived from the logistic model (1) but they apply in other situations. For example, the likelihood function (3) is very similar to the partial likelihood function for Cox's (1972) proportional hazards regression model and hence similar results to those of § 3 apply to that model. Another situation covered by the Theorems of § 3 is the regression model of § 2 with $g = 2$ and the logistic function replaced by any other continuous distribution function $h(t)$. All the previous results hold provided that sampling is from (H, x) or $H \mid x$, and that $h(.)$ satisfies concavity conditions, guaranteeing the uniqueness of the maximum likelihood solution at interior points of the parameter space. For the g-group case and separate sampling from $x \mid H$, the position is more complex and further work is required to establish conditions for the existence of the maximum likelihood solutions. Further work is also needed on the general regression models of Anderson (1979) and the ordered logistic models of Anderson & Philips (1981).

Finally, we believe our results open promising perspectives in defining general rules for existence of maximum likelihood estimates in log linear models for frequency tables, a problem that Haberman (1974, Appendix B) reports to be difficult and unresolved for high dimensional tables. The results already obtained can be used to obtain a partial answer since a logistic model of type (2) can always be derived from a log linear model by finding the probabilities for one of the dimensions conditional on all the others. The parameters of this logistic model are functions of those of the log linear model and we conjecture that the maximum likelihood estimates of the latter model do not exist if the conditional maximum likelihood estimates of the logistic parameters do not exist.

For example, using Haberman's notation, let $\{n_{ijk}: <i, j, k> \in \bar{I}, \bar{J}, \bar{K}\}$ be a three-dimensional frequency table. By definition, $\bar{I} = \{1, 2, \ldots, I\}$ and similarly for $\bar{J}$ and $\bar{K}$. If one decides to regard the subscript k, for example, as a group membership indicator variable ($k \in \bar{K}$), and the subscripts i and j as two independent variables ($i \in \bar{I}$, $j \in \bar{J}$), then in the two-dimensional space (i, j), the frequency table above can be viewed as observation points from $\bar{K}$ distinct populations, exactly as in the regression model. If the $\bar{K}$ subsamples of size $n_{..k}$ are completely or quasicompletely separated, then according to our results we can be assured that the maximum likelihood logistic estimates do not exist.

The application of our method to the frequency tables displayed in Haberman's

Appendix B led to the following results: Tables B.6, B.7, B.11 and B.15 are 'overlap cases' and thus have a finite maximum likelihood estimate. Tables B.8 and B.10, however, are 'quasicomplete separation cases', and therefore do not admit a maximum likelihood solution. Thus our method confirms Haberman's results but whereas Haberman's approach falls short for higher dimensional tables, our approach does not. In fact, it seems likely that the methods of § 3 can be modified to give direct existence criteria for the log linear and exponential families. Further elucidation is required.

The authors wish to thank the referees for their helpful comments on an earlier version of the paper. Partial support for this work was provided by the U.S. National Institutes of Health.

REFERENCES

ALBERT, A. & ANDERSON, J. A. (1981). Probit and logistic discriminant functions. *Comm. Statist.* A **10**, 641–57.

ANDERSON, J. A. (1972). Separate sample logistic discrimination. *Biometrika* **59**, 19–35.

ANDERSON, J. A. (1974). Diagnosis by logistic discriminant function: Further practical problems and results. *Appl. Statist.* **23**, 397–404.

ANDERSON, J. A. (1979). Robust inference using logistic models. *Bull. Int. Statist. Inst.* **48**, 35–53.

ANDERSON, J. A. & BLAIR, V. (1982). Penalized maximum likelihood estimation in logistic regression and discrimination. *Biometrika* **69**, 123–36.

ANDERSON, J. A. & PHILIPS, P. R. (1981). Regression, discrimination and measurement models for ordered categorical variables. *Appl. Statist.* **30**, 22–31.

COX, D. R. (1970). *The Analysis of Binary Data*. London: Methuen.

COX, D. R. (1972). Regression models and life tables (with discussion). *J. R. Statist. Soc.* B **34**, 187–220.

DAY, N. E. & KERRIDGE, D. F. (1967). A general maximum likelihood discriminant. *Biometrics* **23**, 313–23.

HABERMAN, S. J. (1974). *The Analysis of Frequency Data*. University of Chicago Press.

MCCULLAGH, P. (1980). Regression models for ordinal data (with discussion). *J. R. Statist. Soc.* B **42**, 109–27.

PRENTICE, R. L. & PYKE, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–11.

SILVAPULLE, M. J. (1981). On the existence of maximum likelihood estimates for the binomial response models. *J. R. Statist. Soc.* B **43**, 310–3.

WEDDERBURN, R. W. M. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika* **63**, 27–32.

[*Received August* 1981. *Revised June* 1983]