Asymptotic Properties of M-estimators with Applications in Discriminant Analysis

by

Tzu-Cheg Kao
Indiana University-Purdue University at Indianapolis

and

George P. McCabe
Purdue University

Technical Report #83-8

July 1983

Purdue University
West Lafayette, IN 47907

AMS 1980 subject classification. 62F12, 62H30.

Keywords and phrases. Consistency, asymptotic normality, M-estimation, general regression approach, logistic regression approach, mixture sampling, stratified sampling, discriminant analysis.

Summary

Asymptotic Properties of M-estimators with
Applications in Discriminant Analysis

Consistency and asymptotic normality for a large class of M-estimators is demonstrated under a set of useful regularity conditions. The results are used to study the asymptotic theory of a class of logistic-type discrimination functions. The normal equal and unequal covariance matrix cases are examined in detail.

Asymptotic Properties of M-estimators with Applications in Disciminant Analysis

by

Tzu-Cheg Kao
Indiana University-Purdue University at Indianapolis

and

George P. McCabe
Purdue University

1. Introduction.

Let Z_i, i=1,...,n be independent, not necessarily identically distributed random vectors with cumulative distribution functions $F_i(z;\eta)$ and densities $f_i(z;\eta)$ with respect to some σ-finite measure λ. Note that λ may be Lebesque measure or counting measure. Thus, we include absolutely continuous and discrete distributions. We assume that η is a p-dimensional parameter vector in Ω_0, a non-empty subset of $\mathbb{R}^p$. Furthermore, we assume that $\eta = (\theta_1',\ldots,\theta_k',\gamma')'$ and that F_i depends upon η only through θ_s and γ for $i = n_{s-1}+1,\ldots,n_{s-1}+n_s$, where $n_0=1$. Clearly, $n = \sum_{s=1}^{k} n_s$. In other words, $\{Z_i\}$ contains k sets of iid random vectors. The sets are of size n_s and the common cumulative distribution function will be denoted by $F(z;\theta_s,\gamma)$ where convenient. The parameter γ is common to all of the random vectors while θ_s refers to the s^{th} set.

Let $\beta = (\beta_1,\ldots,\beta_q)'$ be a vector valued function of η lying in Ω, an open convex subset of R^q, with $q \leq p$. We will be concerned with estimation of the parameter β using the following framework.

Let $g_i(Z_i,\beta)$ be an a.e. positive function and let

$$(1.1) \qquad \phi^n(\beta) = \prod_{i=1}^{n} g_i(Z_i,\beta).$$

An estimator, $\hat{\beta}$, which maximizes $\phi^n(\beta)$ with respect to $\beta \in \Omega$, is called an M-estimator. In the special case where $g_i(z_i,\beta)$ is the density of Z_i, the maximizing $\hat{\beta}$ is called the maximum likelihood estimator.

Several authors, such as Cramér (1946), Chanda (1954), Bradley and Gart (1962), and Tarone (1974) have studied maximum likelihood estimation of β, under the condition that $\theta_1=\ldots=\theta_k$ and $\beta=\eta$. They require regularity conditions involving the third order partial derivatives of g_i. Other authors such as Huber (1967) and Inagaki (1973), have studied M-estimation with regularity conditions not involving second and higher order partial derivatives of g_i.

We have encountered applied problems for which the regularity conditions in the above papers are very difficult to verify. Therefore, we have developed general theorems on the consistency and asymptotic normality of the M-estimator which can be easily verified for the problems which concern us.

Under regularity conditions involving second order partial derivatives, we show the consistency and asymptotic normality of the M-estimator in sections 2 and 3, respectively. In section 4, an area of application is described. Sections 5 and 6 discuss detailed models using the results in the previous sections.

2. Consistency

Let

$$p_i(z,\beta) = \log g_i(z,\beta),$$

$$\varphi^n(\beta) = \log \phi^n(\beta),$$

$$L_r^n(\beta) = \frac{1}{n}\frac{\partial \varphi^n(\beta)}{\partial \beta_r},$$

and

$$L_{rs}^n(\beta) = -\frac{1}{n}\frac{\partial^2 \varphi^n(\beta)}{\partial \beta_r \partial \beta_s},$$

for $n=1,2,\ldots$ and $r,s=1,\ldots,q$. Let $(L_r^n(\beta))$ and $(L_{rs}^n(\beta))$ denote the vector of first derivatives and matrix of second derivatives, respectively. Let η^0 denote the true value of η and let β^0 be the corresponding value of β. In what follows, all probabilities and expected values are calculated under the distribution evaluated at $\eta = \eta^0$.

We now state some regularity conditions.

(C1) The functions $p_i(z,\beta)$ are twice continuously differentiable with respect to β for all $\beta \in \Omega$.

(C2) $(L_r^n(\beta^0)) \xrightarrow{P} 0$.

(C3) $\sup_{\beta \in \Omega} \{L_{rs}^n(\beta)\} = O_p(1)$.

We are now in a position to prove

Theorem 2.1. If a sequence of M-estimators, $\{\hat{\beta}^n\}$, exists with probability tending to 1 as $n \to \infty$ and the regularity conditions (C1) to (C3) hold, then

$$\hat{\beta}^n \xrightarrow{P} \beta^0.$$

Proof. From (C1), we can expand $\left(\frac{\partial \varphi^n(\beta)}{\partial \beta_r}\right)$ in Taylor series about β^0 as

$$(2.1) \qquad \left(\frac{\partial \varphi^n(\beta)}{\partial \beta_r}\right) = \left(\frac{\partial \varphi^n(\beta^0)}{\partial \beta_r}\right) + \left(\frac{\partial^2 \varphi^n(\beta^*)}{\partial \beta_r \partial \beta_s}\right)(\beta - \beta^0),$$

where $\beta^* = \lambda\beta + (1-\lambda)\beta^0$, for some λ, $0 < \lambda < 1$.

By assumption, $\left(\frac{\partial \varphi^n(\beta)}{\partial \beta_r}\right) = 0$ has a root, $\hat{\beta}^n$, with probability

tending to 1 as $n\to\infty$. Letting $\beta = \hat{\beta}^n$ in (2.1) and dividing by n gives

$$(2.2) \qquad (L_{rs}^n(\beta^{*n}))(\hat{\beta}^n - \beta^0) = (L_r^n(\beta^0)),$$

where $\beta^{*n} = \lambda\hat{\beta}^n + (1-\lambda)\beta^0$ for some λ, $0 < \lambda < 1$.

From (C2), the right-hand side of (2.2) converges in probability to zero. Therefore,

$$(2.3) \qquad (L_{rs}^n(\beta^{*n}))(\hat{\beta}^n - \beta^0) \xrightarrow{P} 0.$$

Since Ω is convex, we have $\beta^{*n} \in \Omega$. Applying (C3) gives $\sup(L_{rs}^n(\beta^{*n})) = O_p(1)$ and hence, (2.3) implies $\hat{\beta}^n - \beta^0 \xrightarrow{P} 0$. □

Now suppose that the M-estimators are not unique. For any two of them say $\hat{\beta}^n$ and $\tilde{\beta}^n$, it is straightforward to show that

$$\hat{\beta}^n - \tilde{\beta}^n \xrightarrow{P} 0.$$

Theorem 2.1 remains valid if the condition (C3) is replaced by

$$(C3') \qquad (L_{rs}^n(\beta)) - (L_{rs}(\beta)) = o_p(1)$$

where

$$\sup_{\beta\in\Omega} \{(L_{rs}(\beta))\} \text{ is finite.}$$

This result is a consequence of the fact that (C3') implies (C3).

To obtain strong consistency results, we simply replace convergence in probability by almost sure convergence in the appropriate places. Thus, we obtain,

Theorem 2.2. If a sequence M-estimators, $\{\beta^n\}$, exists with probability tending to 1 as $n\to\infty$ and conditions (C1) to (C3) hold with almost sure convergence in place of convergence in probability for (C3), then

$$\hat{\beta}^n \xrightarrow{a.s.} \beta^0.$$

The proof follows that of Theorem 2.1 with trivial modifications.

Again, suppose that the M-estimators are not unique. For any two of them, say $\hat{\beta}^n$ and $\tilde{\beta}^n$, it is straightforward to show that

$$\hat{\beta}^n - \tilde{\beta}^n \xrightarrow{a.s.} 0.$$

If we replace convergence in probability by almost sure convergence in condition (C3') we get a similar alternate condition for the validity of Theorem 2.2.

3. Asymptotic Normality

To demonstrate asymptotic normality for the sequence $\hat{\beta}^n$, we use the following regularity conditions.

(N1) $(L^n_{rs}(\beta^0)) \xrightarrow{P} (L_{rs}(\beta^0))$,

where $(L_{rs}(\beta^0))$ is a positive definite matrix.

(N2) For any $\epsilon > 0$ and $\beta = \beta^0$,

$$\frac{1}{n}\sum_{i=1}^{n}\int_{D_{n,i}}\sum_{r=1}^{q}\left[\frac{\partial p_i}{\partial \beta_r}-E\left(\frac{\partial p_i}{\partial \beta_r}\right)\right]^2 dF_i(z_i;\eta^0)=o(1).$$

where

$$D_{n,i}=\left\{\sum_{r=1}^{q}\left[\frac{\partial p_i}{\partial \beta_r}-E\left(\frac{\partial p_i}{\partial \beta_r}\right)\right]^2>\epsilon\ n\right\}.$$

(N3) For all r and s, and $\beta=\beta^0$,

$$(3.1)\qquad \frac{1}{n}\sum_{i=1}^{n}\int\left(\frac{\partial p_i}{\partial \beta_r}-E\left(\frac{\partial p_i}{\partial \beta_r}\right)\right)\left(\frac{\partial p_i}{\partial \beta_s}-E\left(\frac{\partial p_i}{\partial \beta_s}\right)\right)' dF_i(z_i;\eta^0)\to V.$$

In addition, the following lemma is needed.

Lemma 3.1. If

$$(L^n_{rs}(\beta))-(L_{rs}(\beta))=o_p(1),$$

and $(L_{rs}(\beta))$ is positive definite, then $(L^n_{rs}(\beta))$ is positive definite for sufficiently large n with probability 1.

Proof. For each $i=1,\ldots,q$, let $(L^n_{rs}(\beta))_i$ denote the square matrix with i rows, obtained by deleting the last q-i rows and columns of $(L^n_{rs}(\beta))$. The quantities $(L_{rs}(\beta))_i$ are defined similarly. It is sufficient to show that for each i, $\det\{(L^n_{rs}(\beta))_i\}>0$ with probability 1 for sufficiently large n.

Since a determinant is a continuous function of the matrix elements, it follows that

$$\det\{(L^n_{rs}(\beta))_i\}-\det\{(L_{rs}(\beta))_i\}=O_p(1),\ \text{for each } i.$$

By the assumption that $(L_{rs}(\beta))$ is positive definite, $\det\{(L_{rs}(\beta))_i\}>0$ for each i. Hence the result follows immediately. □

The following theorem gives conditions for the asymptotic normality of the M-estimator sequence, $\hat{\beta}^n$. Let $N_q(\mu,\ddagger)$ denote the q-variate normal distribution with mean μ and covariance matrix $\ddagger$.

Theorem 3.1. If a sequence of M-estimators, $\{\hat{\beta}^n\}$ is consistent and conditions (C1), (N1), (N2) and (N3) hold, then

$$\sqrt{n}\,(\hat{\beta}^n-\beta^0) \overset{\mathcal{L}}{\Rightarrow} N_q(L^{-1}u,L^{-1}VL^{-1}) \tag{3.2}$$

where

$$u = \lim_{n\to\infty} \left(\frac{1}{\sqrt{n}} \Sigma_{i=1}^n E\left(\frac{\partial p_i}{\partial \beta_r}\right)\right)\Big|_{\beta^0},$$

$$L = (L_{rs}(\beta^0)),$$

and V is defined by (3.1).

Proof. From (2.2), we have

$$(3.3) \qquad (L^n_{rs}(\beta^{*n}))(\hat{\beta}^n - \beta^0) = (L^n_r(\beta^0)),$$

where $\beta^{*n} = \lambda\hat{\beta}^n + (1-\lambda)\beta^0$ for some λ, $0 < \lambda < 1$.

Also we note that condition (C1) implies that $L^n_{rs}(\beta)$ is uniformly continuous on $H(\beta^0,\delta)$ for $\delta > 0$. Consistency of $\{\hat{\beta}^n\}$ therefore implies

$$(L^n_{rs}(\hat{\beta}^n)) - (L^n_{rs}(\beta^0)) = o_p(1).$$

Combining this fact with condition (N1) gives

$$(L^n_{rs}(\hat{\beta}^n)) - (L_{rs}(\beta^0)) = o_p(1).$$

In a similar fashion, it follows that

$$(3.4) \qquad (L^n_{rs}(\beta^{*n})) - (L_{rs}(\beta^0)) = o_p(1),$$

where β^{*n} is defined by (3.3)

Now, Lemma 3.1 implies that the inverse of $(L^n_{rs}(\beta^{*n}))$ exists for sufficiently large n with probability 1 and hence,

$$(3.5) \qquad (L^n_{rs}(\beta^{*n}))^{-1} - (L_{rs}(\beta^0))^{-1} = o_p(1).$$

Conditions (N2) and (N3) with the multivariate central limit theorem

(see Serfling (1980)) imply

$$(3.6) \qquad \sqrt{n}\,(L_r^n(\beta^0)) \xrightarrow{\mathcal{L}} N_q(u,V),$$

where u is defined above and V is given in (3.1).

Combining (3.3), (3.5) and (3.6) gives the desired result. □

Corollary 3.1. If a sequence of M-estimators exists with probability tending, to 1 as $n\to\infty$, then conditions (C1), (C2), (N1), (N2), (N3) and either (C3) or (C3') imply (3.2).

Proof. This result is a direct consequence of Theorem 2.1, the comment following it, and Theorem 3.1. □

4. Applications

We are concerned with the two sample discriminant analysis problem. Specifically, let Y be a random variable with

$$P(Y=1) = 1 - P(Y=0) = \pi,$$

where $0 < \pi < 1$ and Y indicates from which of the two populations the sample observation is drawn. A k-dimensional random vector X is assumed to have conditional densities, depending upon Y, denoted by $f_y(x;\theta_y,\gamma)$, or $f_y(x;\eta)$, or $f_y(x)$, whichever is more convenient. The parameters θ_y and γ may be vectors, and $\eta = (\theta_0',\theta_1',\gamma')'$.

Given an X from one of the two populations, the expected misclassification probability is minimized by the rule which assigns X to the Y=1 population if and only if

$$(4.1) \qquad \pi f_1(x;\theta_1,\gamma) \geq (1-\pi) f_0(x;\theta_0,\gamma).$$

See Anderson (1958) for details. This rule is also Bayes for a zero-one loss function. An alternative statement of condition (4.1) is

$$(4.2) \qquad P(Y=1|X) \geq P(Y=0|X).$$

Cox (1966) proposed a logistic form for these probabilities, i.e.

$$(4.3) \qquad P(Y=1|X) = \frac{e^{\beta'Z}}{1+e^{\beta'Z}},$$

where $Z = (1,X')'$ and $\beta = (\beta_0,\beta_1,\ldots,\beta_k)'$. His approach is called logistic regression. Note that the usual discriminant analysis approach involves conditioning on Y, i.e. X is the random variable. In logistic regression, however, we condition on X, i.e. Y is the random variable. There is a great deal of confusion in the literature on this point and the phrase maximum likelihood estimation is used without sufficient attention to the exact model which is assumed.

Note that (4.3) holds if the distributions of X given Y are multivariate normal with equal covariance matrices. Lachenbruch (1975) gives other sufficient conditions for this form.

In this section, we assume

$$(4.4) \qquad P(Y=1|X) = G(L(X,\beta))$$

where G is a cumulative distribution function, L is a real-valued function of (X,β) and β is a q-dimensional vector function of θ_0, θ_1 and π. Note that here $\eta^0 = (\theta_0^{0'},\theta_1^{0'},\gamma^{0'})'$ is the true value of η and β^0 is the corresponding value of β. We will consider estimation of the parameter β under two sampling procedures. The first, called mixture sampling (MS), is to take a random sample of size n from the mixture of the two populations giving $\{(Y_1,X_1'),(Y_2,X_2'),\ldots,(Y_n,X_n')\}$. The second, called stratified sampling (SS) is to take random samples of sizes n_1 and n_0 from the two populations, i.e. $\{X_1,\ldots,X_{n_1}\}$ from the population corresponding to Y=1 and $\{X_{n_1+1},\ldots,X_{n_1+n_0}\}$ from the population corresponding to Y=0. We let the total sample size be n, i.e. $n=n_0+n_1$.

The estimation procedure, which we call general regression, is the natural generalization of logistic regression, i.e. our estimator $\hat{\beta}^n$ maximizes

$$(4.5) \qquad \phi_n(\beta) = \prod_{i=1}^{n} [G(L(X_i,\beta))]^{Y_i}[1 - G(L(X_i,\beta))]^{1-Y_i},$$

where Y_i may be random (MS) or fixed (SS) depending upon the sampling procedure. Note that this procedure corresponds to that given by (1.1) with

$$(4.6) \qquad g_i(Z_i,\beta) = [G(L(X_i,\beta))]^{Y_i}[1 - G(L(X_i,\beta))]^{1-Y_i},$$

where $Z_i = (Y_i,X_i')'$ for RS and $Z_i = X_i$ for SS. We use the term logistic regression to describe the special case where $G(t) = e^t/(1+e^t)$.

Note that the estimation procedure described above is not maximum likelihood for MS or SS, even if (4.4) holds. For MS, the likelihood function is

$$(4.7) \qquad \ell_{MS}(\theta_0,\theta_1,\gamma,\pi) = \Pi_{i=1}^{n}[\pi f_1(X_i;\theta_1,\gamma)]^{Y_i}[(1-\pi)f_0(X_i;\theta_0,\gamma)]^{1-Y_i} .$$

The MLE procedure is to maximize this function with respect to $(\theta_0,\theta_1,\gamma,\pi)$ and then to use the functional relationship between β and $(\theta_0,\theta_1,\gamma,\pi)$ with the invariance principle to determine the MLE of β. Note that (4.7) can be rewritten as

$$(4.8) \qquad \ell_{MS}(\theta_0,\theta_1,\gamma,\pi) = \{\Pi_{i=1}^{n}[G(L(X_i,\beta))]^{Y_i}[1-G(L(X_i,\beta))]^{1-Y_i}\} \cdot \{\Pi_{i=1}^{n} f(X_i)\},$$

where $f(x_i) = \pi f_1(x_i;\theta_1,\gamma) + (1-\pi)f_0(x_i;\theta_0,\gamma)$, the marginal density of X_i. Estimation of β by maximizing the first part of the right-hand side of (4.8) is called the conditional likelihood approach and has been studied by Efron (1975) and O'Neill (1980). Efron considered the multivariate normal with equal covariance matrices while O'Neill generalized his results to the case where G has logistic form.

Under SS, the likelihood function is

$$(4.9) \qquad \ell_{SS}(\theta_0,\theta_1,\gamma) = [\Pi_{i=1}^{n_1} f_1(X_i;\theta_1,\gamma)] \cdot [\Pi_{i=n_1+1}^{n} f_0(X_i;\theta_0,\gamma)].$$

This can be rewritten as

$$(4.10)\qquad \ell_{SS}(\theta_0,\theta_1,\gamma) = \{[\Pi_{i=1}^{n_1} G(L(X_i,\beta))] \cdot [\Pi_{i=n_1+1}^{n} (1-G(L(X_i,\beta))]\}$$

$$\cdot \{[\Pi_{i=1}^{n} f(X_i)]/(\pi^{n_1}(1-\pi)^{n_0})\}.$$

Under the assumption that the X_i are multivariate Bernoulli random vectors with marginal distributions that do not depend upon β, Anderson (1972) studied the MLE's for this problem. Using the constraints

$$\Sigma_X f(x) = 1,$$

and

$$\Sigma_X G(L(x,\beta))f(x) = \pi,$$

he applied Aitchison and Silvey's (1958) results to obtain the asymptotic covariance matrix of the estimator. For continuous variables, he recommended an approximation based on discretizing the variables.

Efron (1975) suggests that the MS framework can be used for results in the SS model by using the first part of the right-hand side of (4.7) to derive an estimator and by replacing π by n_1/n in all results. Actually, some slight modifications are needed. These will be detailed in section 6.

Recall that X is assumed to be a k-dimensional random vector and β is a q-dimensional parameter vector. The results presented in the next sections apply to the situation where $L(X,\beta)$ is a linear combination of known functions of X. More precisely, we assume

$$L(X,\beta) = \beta' h(X),$$

where

$$h(X) = (h_0(X), h_1(X), \ldots, h_{q-1}(X))',$$

and

$$\beta = (\beta_0, \beta_1, \ldots, \beta_{q-1})'.$$

and the $h_j(X)$ are real-valued functions of the vector X. Here, we let $h_0(X) \equiv 1$.

Let

$$S_n = \{\Sigma_{i=1}^{n} Y_i a_i h(X_i): \ a_i > 0\},$$

and

$$F_n = \{\Sigma_{i=1} (1-Y_i) a_i h(X_i): \ a_i > 0\}.$$

These sets are the relative interiors of the convex cones generated by the (random) X_i vectors corresponding to $Y_i = 1$ and $Y_i = 0$ respectively.

We assume the following regularity conditions

(R1) For any linear subspace $\mathcal{L}$ of dimension less than q,

$$P(h(X)_i \in \mathcal{L}) = 0.$$

(R2) -log G and log(1-G) are convex.

(R3) G is a strictly increasing function with $0 < G(\cdot) < 1$.

(R4) support $f_0(x;\theta_0,\gamma)$ = support $f_1(x;\theta_1,\gamma)$.

(R5) G has continuous derivatives of the second order.

Silvapulle (1981) has studied estimation of β under the condition that $h(X) = (1,X')'$, where $q=k+1$. He assumes a model where the Y_i are the only random variables. A straightforward generalization of his theorem gives

Theorem 4.1. Let (R1) and (R2) hold. Then an M-estimator $\hat{\beta}^n$ exists and the set of such M-estimators is bounded if and only if $S_n \cap F_n \neq \phi$. Furthermore, if (R3) holds, then the M-estimator $\hat{\beta}^n$ exists and is unique if and only if $S_n \cap F_n \neq \phi$.

5. Mixture Sampling

Recall that the model for MS is that $(Y_1,X_1'),\ldots,(Y_n,X_n')$ are iid random vectors. The estimation method is defined by maximizing (4.5). Here we let $Z_i = (Y_i,X_i')'$.

The existence, consistency and asymptotic normality of the sequence of M-estimators, $\{\hat{\beta}^n\}$, is established in the following theorem.

Theorem 5.1. Suppose that regularity conditions (R1) to (R5) hold, the integral defined by (5.1) exists and

$$\pi \int \left\{ \frac{G''(L(x,\beta))}{G(L(x,\beta))} - \left[\frac{G'(L(x,\beta))}{G(L(x,\beta))}\right]^2 \right\} (h(x))(h(x))' f_1(x;\eta^0)dx$$

(5.2)

$$-(1-\pi)\int \left\{ \frac{G''(L(x,\beta))}{1-G(L(x,\beta))} + \left[\frac{G'(L(x,\beta))}{1-G(L(x,\beta))}\right]^2 \right\} (h(x)(h(x))' f_0(x;\eta^0)dx$$

is uniformly bounded above.

Then for mixture sampling and general regression,

(i) the sequence of the M-estimators, $\{\hat{\beta}^n\}$, exists and is unique with probability tending to 1 as $n\to\infty$;

(ii) $\hat{\beta}^n \xrightarrow{a.s.} \beta^0$; and

(iii) $\sqrt{n}\,(\hat{\beta}^n-\beta^0) \xrightarrow{\mathcal{L}} N_q(0,J^{-1})$,

where

$$J = \int \frac{(G'(L(x,\beta^0)))^2(h(x))(h(x))'f(x)}{G(L(x,\beta^0))(1-G(L(x,\beta^0)))}\,dx, \tag{5.1}$$

and

$$f(x) = \pi f_1(x;\eta^0) + (1-\pi)f_0(x;\eta^0).$$

Proof. It is easy to show that $P(S_n \cap F_n \neq \phi) \to 1$ as $n\to\infty$. Part (i) therefore follows from Theorem 4.1.

To establish the strong consistency of $\hat{\beta}^n$ we will apply Theorem 2.2. It is therefore sufficient to verify regularity conditions (C1), (C2) and (C3') for almost sure convergence.

From the definition of $p_i(Z_i,\beta)$ in section 2, it follows that

$$p_i(Z,\beta) = Y\log(G(L(X,\beta))) + (1-Y)\log(1-G(L(X,\beta))), \tag{5.2}$$

for all i. The fact that the $p_i(Z,\beta)$ are twice continuously differentiable (C1) thus follows from the corresponding assumption for G, i.e. (R5).

Recall that

$$\varphi^n(\beta) = \Sigma_{i=1}^n p(Z_i,\beta).$$

Note that the subscript i is dropped from p since the function does not depend upon i. By the strong law of large numbers,

$$(L_r^n(\beta^0)) = \frac{1}{n}\left(\frac{\partial \varphi^n(\beta)}{\partial \beta_r}\right)_{\beta^0} \xrightarrow{a.s.} E\left(\frac{\partial p(Z,\beta)}{\partial \beta_r}\right)_{\beta^0} .$$

To prove (C2) for almost sure convergence it remains to show that the latter expected value is zero.

From (4.4) and (R4) we note that

$$\frac{\pi f_1(x;n^0)}{(1-\pi)f_0(x;n^0)} = \frac{G(L(x,\beta^0))}{1-G(L(x,\beta^0))} .$$

The result follows immediately.

To prove (C3') for almost sure convergence, we proceed in a similar manner. Recall that

$$(L_{rs}^n(\beta)) = -\frac{1}{n}\left(\frac{\partial^2 \varphi^n(\beta)}{\partial \beta_r \partial \beta_s}\right) = -\frac{1}{n}\left(\sum_{i=1}^n \frac{\partial^2 p(Z_i,\beta)}{\partial \beta_r \partial \beta_s}\right) .$$

Therefore, by the strong law of large numbers,

$$(5.4) \qquad (L_{rs}^n(\beta)) \xrightarrow{a.s.} -E\left(\frac{\partial^2 p(Z,\beta)}{\partial \beta_r \partial \beta_s}\right).$$

Note that the right-hand side of (5.4) is equal to (5.2). Therefore, (C3') for almost sure convergence follows by assumption and (ii) is proved.

To prove (iii), we will apply Theorem 3.1. Given (i) and (ii), it remains to show that regularity conditions (N1), (N2) and (N3) hold. To demonstrate (N1) it is necessary to show that the matrix on the right-hand side of (5.4) is positive definite. With a little manipulation, this matrix is seen to be the matrix J given in (5.1). Note that $G'(L(X,\beta))$

denotes the derivative of the function $G(\cdot)$ evaluated at $L(X,\beta^0)$. This derivative exists, is positive and is continuous by (R3) and (R5). Furthermore, $h(X)(h(X))'$ is positive definite by (R1). Therefore, the matrix J is positive definite and (N1) is satisfied. Note that the assumption on the second derivatives is used in verifying that the right-hand side of (5.4) is equal to J.

To verify (N2) we first observe that for any $\epsilon > 0$,

$$(5.5) \qquad \frac{1}{n}\sum_{i=1}^{n}\int_{D_{n,i}}\sum_{r=1}^{q}\left[\frac{\partial p_i(Z_i,\beta)}{\partial\beta_r} - E\left(\frac{\partial p_i(Z_i,\beta)}{\partial\beta_r}\right)\right]^2 dF_i$$

$$= \int_{D_n}\sum_{r=1}^{q}\left(\frac{\partial p(Z,\beta)}{\partial\beta_r}\right)^2 dF,$$

where

$$D_{n,i} = \left\{\sum_{r=1}^{q}\left[\frac{\partial p_i(Z_i,\beta)}{\partial\beta_r} - E\left(\frac{\partial p_i(Z_i,\beta)}{\partial\beta_r}\right)\right]^2 > \epsilon\ n\right\};$$

D_n is defined similarly with p_i and X_i replaced by p and X; F_i is the distribution function of Z_i and F is the distribution function of Z. The equality follows because the Z_i are i.i.d., the p_i do not depend upon i, and $E(\partial p_i(Z_i,\beta)/\partial\beta_r)_{\beta 0} = 0$. Note that

$$(5.6) \qquad \left(E\ \frac{\partial p(Z,\beta)}{\partial\beta_r}\cdot\frac{\partial p(Z,\beta)}{\partial\beta_s}\right)_{\beta 0} = -E\left(\frac{\partial^2 p(Z,\beta)}{\partial\beta_r\partial\beta_s}\right)_{\beta 0},$$

which is equal to J and therefore assumed to exist by hypothesis. Thus,

the left-hand side of (5.5) is o(1) and (N2) is established.

Finally, for $\beta = \beta^0$,

$$\left(\frac{1}{n}\sum_{i=1}^{n}\int\left[\frac{\partial p_i(Z_i,\beta)}{\partial\beta_r} - E\left(\frac{\partial p_i(Z_i,\beta)}{\partial\beta_r}\right)\right]\left[\frac{\partial p_i(z_i,\beta)}{\partial\beta_s} - E\left(\frac{\partial p_i(Z_i,\beta)}{\partial\beta_s}\right)\right]'dF_i\right)$$

$$= \left(E\frac{\partial p(Z,\beta)}{\partial\beta_r}\cdot\frac{\partial p(Z,\beta)}{\partial\beta_s}\right),$$

by the same argument used to establish (5.5). Since this matrix is J, (N3) is thus established and the conditions for the validity of Theorem 3.1 are demonstrated.

To complete the proof of the theorem it suffices to note that

$$\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}E\left(\frac{\partial p_i(Z_i,\beta)}{\partial\beta_r}\right)\right)_{\beta^0} = \sqrt{n}\,E\left(\frac{\partial p(Z,\beta)}{\partial\beta_r}\right)_{\beta^0} = 0. \quad \square$$

For logistic regression, i.e. $G(t) = e^t/(1+e^t)$, some of the assumptions in the above theorem are automatically satisfied and the form of the matrix J can be simplified. The result is given in the following theorem.

Theorem 5.2. Suppose that regularity conditions (R1) and (R4) hold, the integral defined by (5.7) exists and

$$\int \frac{\exp(L(X,\beta))}{[1+\exp(L(X,\beta))]^2}\,[h(x)][h(x)]'f(x)dx$$

is uniformly bounded above, where

$$f(x) = \pi f_1(x;n^0) + (1-\pi)f_0(x;n^0).$$

Then, for mixture sampling and logistic regression,

(i) the sequence of the M-estimators, $\{\hat{\beta}^n\}$, exists and is unique with probability tending to 1 as $n\to\infty$;

(ii) $\hat{\beta}^n \xrightarrow{a.s.} \beta^0$; and

(iii) $\sqrt{n}\,(\hat{\beta}^n - \beta^0) \xrightarrow{\mathcal{L}} N_q(0,J^{-1})$,

where

$$J = \Pi(1-\pi)\int \frac{f_1(x;\eta^0)f_0(x;\eta^0)[h(x)][h(x)]'}{f(x)}\,dx. \tag{5.7}$$

Proof. First note that regularity conditions (R2), (R3) and (R5) are satisfied by the logistic function. Furthermore, $G'(t) = G(t)(1-g(t))$. Using this fact and

$$G(L(x,\beta^0)) = \frac{\pi f_1(x;\eta^0)}{\pi f_1(x;\eta^0)+(1-\pi)f_0(x;\eta^0)}.$$

gives the desired result. □

Note that the simplification of (5.1) to (5.7) used the fact that $G'(t) = G(t)(1-G(t))$. This condition is equivalent to the assumption that $G(t)$ is the logistic function.

The following theorem concerns the multivariate normal, equal covariance matrix case. We assume that X given Y=i is $N_k(\mu_i,\ddagger)$, for i=0,1. Thus, letting $\beta = (\beta_0,\delta')'$, we have

$$L(X,\beta) = \beta_0 + \delta'X, \tag{5.8}$$

where

$$\beta_0 = \log\left(\frac{\pi}{1-\pi}\right) - \frac{1}{2}\left(\mu_1'\Sigma^{-1}\mu_1 - \mu_0'\Sigma^{-1}\mu_0\right), \tag{5.9}$$

and

$$\delta = \Sigma^{-1}(\mu_1 - \mu_0). \tag{5.10}$$

Theorem 5.3. Suppose that X given Y is $N_k(u_Y,\Sigma)$. Then for mixture sampling and logistic regression,

(i) the sequence of the M-estimators, $\{\hat{\beta}^n\}$, exists, and is unique with probability tending to 1 as $n\to\infty$;

(ii) $\hat{\beta}^n \xrightarrow{a.s.} \beta^0$; and

(iii) $\sqrt{n}\,(\beta^n-\beta^0) \xrightarrow{\mathcal{L}} N_{k+1}(0,J^{-1})$,

where

$$J = \pi(1-\pi)\int \frac{f_1(x)f_0(x)}{\pi f_1(x)+(1-\pi)f_0(x)}\binom{1}{x}(1,x')dx. \tag{5.11}$$

Proof. First, (R1) and (R4) are satisfied for the normal case. Also, the posterior G is clearly logistic and it is easy to show that J exists. Therefore, the result follows from Theorem 5.2. □

For the above problem, O'Neill (1980) suggests using Bradley and Gart's (1962) results to obtain the asymptotic distribution theory of the estimator. To apply these results, the following regularity condition must be verified:

$$E_{X|Y=1}\left[\frac{\partial \log G(L(X,\beta))}{\partial \beta_r}\right]_{\beta=\beta^0} = E_{X|Y=0}\left[\frac{\partial \log(1-G(L(X,\beta)))}{\partial \beta_r}\right]_{\beta=\beta^0} = 0, \tag{5.12}$$

where G is the logistic function. Letting

$$\Delta = [(\mu_1-\mu_0)'\Sigma^{-1}(\mu_1-\mu_0)]^{1/2},$$

and

$$\lambda = \log(\pi/(1-\pi)),$$

we can reduce without loss of generality to the case where $\mu_1 = -\mu_0 = (\Delta/2)e_1$ and $\Sigma = I$, where $e_1 = (1,0,\ldots,0)'$. Therefore, $L(X,\beta) = \lambda + \Delta X_1$.

A straightforward calculation gives

$$E_{X|Y=0}\left[\frac{\partial \log G(L(X,\beta))}{\partial \beta_r}\right]_{\beta=\beta 0} = -\pi\,(A_0,A_1,0,\ldots,0)',$$

and

$$E_{X|Y=1}\left[\frac{\partial \log(1-G(L(X,\beta)))}{\partial \beta_r}\right]_{\beta=\beta 0} = (1-\pi)\,(A_0,A_1,0,\ldots,0)',$$

where

$$A_i = \frac{\exp(-\Delta^2/8)}{(2\pi)^{1/2}} \int \frac{t^i \exp(-t^2/2)dt}{(1-\pi)\exp(-\Delta t/2)+\pi\exp(\Delta t/2)},$$

for i=0,1. Since $A_0 > 0$, condition (5.12) is violated.

The results in this section are sufficiently general to include the normal, unequal covariance matrix case. We assume that X given Y=i is $N_k(\mu_i,\Sigma_i)$, for i=0,1. Here,

(5.13) $$L(X,\beta) = \beta_0 + \delta'X + X'BX,$$

where

$$(5.14)\qquad \beta_0 = \log\left(\frac{\pi|\Sigma_0|^{1/2}}{(1-\pi)|\Sigma_1|^{1/2}}\right) - \frac{1}{2}\left(\mu_1'\Sigma_1^{-1}\mu_1 - \mu_0'\Sigma_0^{-1}\mu_0\right),$$

$$(5.15)\qquad \delta = \Sigma_1^{-1}\mu_1 - \Sigma_0^{-1}\mu_0,$$

and

$$(5.16)\qquad B = \frac{1}{2}\left(\Sigma_0^{-1} - \Sigma_1^{-1}\right).$$

The function L is clearly of the form $\beta' h(X)$ where

$$(5.17)\qquad h(X) = (1, X_1, \ldots, X_k, X_1^2, 2X_1X_2, \ldots, 2X_1X_k, X_2^2, 2X_2X_3, \ldots, X_k^2)'.$$

Here $q = 1 + 2k + k(k-1)/2$. The following theorem gives the result.

Theorem 5.4. Suppose that X given Y is $N_k(\mu_Y, \Sigma_Y)$. Then, for mixture sampling and logistic regression,

(i) the sequence of the M-estimators, $\{\hat{\beta}^n\}$, exists and is unique with probability tending to 1 as $n \to \infty$;

(ii) $\hat{\beta}^n \xrightarrow{a.s.} \beta^0$; and

(iii) $\sqrt{n}\,(\hat{\beta}^n - \beta^0) \xrightarrow{\mathcal{L}} N_q(0, J^{-1})$,

where

$$(5.18)\qquad J = \pi(1-\pi)\int \frac{f_1(x)f_0(x)[h(x)][h(x)]'}{\pi f_1(x) + (1-\pi)f_0(x)}\,dx,$$

and $h(x)$ is given by (5.17).

Proof. The result follows from Theorem 5.2 in the same way that Theorem 5.3 is proved. □

and $h(x)$ is given by (5.17).

Proof. The result follows from Theorem 5.2 in the same way that Theorem 5.3 is proved.

6. Stratified Sampling

Recall that the model for SS is that $X_1,\ldots,X_{n_1}$ are i.i.d. according to $f_1(x;\theta_1,\gamma)$ and $X_{n_1+1},\ldots,X_{n_1+n_0}$ are i.i.d. according to $f_0(x;\theta_0,\gamma)$ and the two samples are independent. The Y_i are non-random indicators which denote the two populations. The total sample size is $n=n_1+n_0$. The estimation method is that defined by maximizing (4.5). Since Y is not a random variable, we will define

$$(6.1) \qquad G(L(x,\beta)) = \frac{\pi f_1(x;\theta_1,\gamma)}{\pi f_1(x;\theta_1,\gamma)+(1-\pi)f_0(x;\theta_0,\gamma)} \quad .$$

A complication present with this sampling scheme is that the sample fraction n_1/n is non-random and does not necessarily relate to the parameter π. Therefore, it will be necessary to either assume that π is known or that a separate estimator for this parameter is available. In what follows, we let

$$\pi^* = \lim_{n\to\infty} \frac{n_1}{n} ,$$

and we assume

$$(\frac{n_1}{n} - \pi^*) = o(n^{-1/2}).$$

Let β^* be the parameter which would correspond to a model with true probability π^*. Hence we have

$$(6.2) \qquad G(L(x,\pi^*)) = \frac{\pi^* f_1(x;m^0)}{\pi^* f_1(x;\eta^0)+(1-\pi^*)f_0(x;\eta^0)} .$$

The following theorem gives results related to estimation of β^*.

Theorem 6.1. Suppose that regularity conditions (R1) to (R5) hold, the integrals defined by (6.3) and (6.4) exist and

$$\pi^* \int \left\{ \frac{G''(L(x,\beta))}{G(L(x,\beta))} - \left[\frac{G'(L(x,\beta))}{G(L(x,\beta))}\right]^2 \right\} (h(x))(h(x))' f_1(x;\eta^0)dx$$

$$-(1-\pi^*) \int \left\{ \frac{G''(L(x,\beta))}{1-G(L(x,\beta))} + \left[\frac{G'(L(x,\beta))}{1-G(L(x,\beta))}\right]^2 \right\} (h(x))(h(x))' f_0(x;\eta^0)\, dx$$

is uniformly bounded above. Then, for stratified sampling and general regression,

(i) the sequence of the M-estimators, $\{\hat{\beta}^n\}$, exists, and is unique with probability tending to 1 as $n\to\infty$;

(ii) $\hat{\beta}^n \xrightarrow{a.s.} \beta^*$, and

(iii) $\sqrt{n}(\hat{\beta}^n-\beta^*) \xrightarrow{\mathcal{L}} N_q(0,J_S)$;

where

$$(6.2) \qquad J_S = L^{-1} - L^{-1}KL^{-1},$$

$$(6.3) \qquad L = \int \frac{(G'(L(x,\beta^*)))^2}{G(L(x,\beta^*))(1-G(L(x,\beta^*))} [h(x)][h(x)]'f(x)dx,$$

and

$$(6.4) \qquad K = \frac{1}{\pi^*(1-\pi^*)} (\int G'(L(x,\beta^*))h(x)f(x)dx)(\int G'(L(x,\beta^*))h(x)f(x)dx)',$$

and

$$f(x) = \pi^* f_1(x;\eta^0) + (1-\pi^*)f_0(x;\eta^0).$$

Proof. The proof of this theorem follows from arguments similar to those used in the proof of theorem 5.1. Part (i) follows from the independence of X_i, condition (R4) and Theorem 4.1.

To prove (ii) we verify the regularity conditions (C1), (C2) and (C3') for almost sure convergence and apply Theorem 2.2. Condition (C1) follows directly from (R5). Here,

$$(6.5) \qquad p_1(X_i,\beta) = \log(G(L(X_i,\beta))),$$

and

$$(6.6) \qquad p_0(X_i,\beta) = \log(1 - G(L(X_i,\beta))).$$

Note that here $Z_i = X_i$. Furthermore,

$$\varphi^n(\beta) = \Sigma_{i=1}^{n_1} \log(G(L(X_i,\beta))) + \Sigma_{i=n_1+1}^{n} \log(1-G(L(X_i,\beta))),$$

and

$$(L_r^n(\beta)) = (\frac{n_1}{n})(\frac{1}{n_1} \sum_{i=1}^{n_1} (\frac{\partial p_1(X_i,\beta)}{\partial \beta_r})) + (\frac{n_0}{n})(\frac{1}{n_0} \sum_{i=n_1+1}^{n} (\frac{\partial p_0(X_i,\beta)}{\partial \beta})).$$

By the strong law of large numbers,

$$\frac{1}{n_1} \sum_{i=1}^{n_1} (\frac{\partial p_1(X_i,\beta)}{\partial \beta_r})_{\beta^*} \xrightarrow{a.s.} E_0(\frac{\partial p_0(X,\beta)}{\partial \beta_r})_{\beta^*},$$

and

$$\frac{1}{n_0} \sum_{i=n_1+1}^{n} (\frac{\partial p_0(X_i,\beta)}{\partial \beta_r})_{\beta^*} \xrightarrow{a.s.} E_0(\frac{\partial p_0(X,\beta)}{\partial \beta_r})_{\beta^*}.$$

Therefore,

$$(6.7) \qquad (L_r^n(\beta^*)) \xrightarrow{a.s.} \pi^* E_1(\frac{\partial p_1(X,\beta)}{\partial \beta_r})_{\beta^*} + (1-\pi^*)E_0(\frac{\partial p_0(X,\beta)}{\partial \beta_r})_{\beta^*}.$$

Now,

$$E_1(\frac{\partial p_1(X,\beta)}{\partial \beta_r})_{\beta^*} = \int \frac{G'(L(x,\beta^*))}{G(L(x,\beta^*))} h(x) f_1(x;\eta^0)dx,$$

and

$$E_0(\frac{\partial p_0(X,\beta)}{\partial \beta_r})_{\beta^*} = -\int \frac{G'(L(x,\beta))}{1-G(L(x,\beta^*))} h(x) f_0(x;\eta^0)dx.$$

Therefore, by (6.2), the right-hand side of (6.7) is zero and condition (C2) for almost sure convergence is established.

To demonstrate (C3') for almost sure convergence, we first note that

$$-(L^n_{rs}(\beta)) = \frac{1}{n}\left(\frac{\partial^2 \ell^n(\beta)}{\partial\beta_r\partial\beta_s}\right)$$

$$= \left(\frac{n_1}{n}\right)\left(\frac{1}{n_1}\sum_{i=1}^{n_1}\left(\frac{\partial^2 p_1(X_i,\beta)}{\partial\beta_r\partial\beta_s}\right)\right) + \frac{n_0}{n}\left(\frac{1}{n_0}\sum_{i=n_1+1}^{n}\left(\frac{\partial^2 p_0(X_i,\beta)}{\partial\beta_r\partial\beta_s}\right)\right).$$

Using the strong law of large numbers, it follows that

$$(6.8)\qquad -(L^n_{rs}(\beta)) \xrightarrow{a.s.} \pi^* E_1\left(\frac{\partial^2 p_1(X,\beta)}{\partial\beta_r\partial\beta_s}\right) + (1-\pi^*)E_0\left(\frac{\partial^2 p_0(X,\beta)}{\partial\beta_r\partial\beta_s}\right) = (L_{rs}(\beta)).$$

Note that the right-hand side of (6.8) is the function which is to be uniformly bounded. Therefore, (C3') for almost sure convergence follows and (ii) is proved.

To prove (iii) we apply Theorem 3.1. To verify (N1) we need to show that $(L_{rs}(\beta^*))$ is positive definite. First, we note that

$$E_1\left(\frac{\partial^2 p_1(X,\beta)}{\partial\beta_r\partial\beta_s}\right)_{\beta^*} = \int \left[\frac{G''(L(x,\beta^*)}{G(L(x,\beta^*))} - \left(\frac{G'(L(x,\beta^*))}{G(L(x,\beta^*))}\right)^2\right][h(x)][h(x)]'$$

$$\cdot f_1(x;n^0)dx,$$

and

$$E_0\left(\frac{\partial^2 p_0(X,\beta)}{\partial\beta_r\partial\beta_s}\right)_{\beta^*} = -\int \left[\frac{G'(L(x,\beta^*))}{1-G(L(x,\beta^*))} + \left(\frac{G'(L(x,\beta^*))}{1-G(L(x,\beta^*))}\right)^2\right][h(x)][h(x)]'$$

$$\cdot f_0(x;n^0)dx.$$

Combining with (6.8) and simplifying by (6.2) gives

$$(L_{rs}(\beta^*)) = \int \frac{(G'(L(x,\beta^*)))^2}{G(L(x,\beta^*))(1-G(L(x,\beta^*))}[h(x)][h(x)]'f(x)dx,$$

which is equal to (6.3). Positive definiteness follows from (R1), (R3) and (R5) and thus (N1) is satisfied.

To establish (N2) we first note that for any $\epsilon > 0$,

$$(6.9) \quad \frac{1}{n}\sum_{i=1}^{n}\int_{D_{n,i}}\sum_{r=1}^{q}\left[\frac{\partial p_i(x_i,\beta)}{\partial\beta_r} - E\left(\frac{\partial p_i(X_i,\beta)}{\partial\beta_r}\right)\right]^2 dF_i$$

$$= \frac{n_1}{n}\int_{D_{n,1}}\sum_{r=1}^{q}\left[\frac{\partial p_1(x,\beta)}{\partial\beta_r} - E\left(\frac{\partial p_1(X,\beta)}{\partial\beta_r}\right)\right]^2 f_1(x;\pi^0)dx$$

$$+ \frac{n_0}{n}\int_{D_{n,0}}\sum_{r=1}^{q}\left[\frac{\partial p_0(x,\beta)}{\partial\beta_r} - E\left(\frac{\partial p_0(X,\beta)}{\partial\beta_r}\right)\right]^2 f_0(x;\pi^0)dx$$

where

$$D_{n,j} = \left\{\sum_{r=1}^{q}\left[\frac{\partial p_j(X,\beta)}{\partial\beta_r} - E\left(\frac{\partial p_j(X,\beta)}{\partial\beta_j}\right)\right]^2 > \epsilon\ n\right\}$$

for j=0,1. Furthermore,

$$\pi^* E_1\left(\left(\frac{\partial p_1}{\partial\beta_r}\right)\left(\frac{\partial p_1}{\partial\beta_s}\right)\right)_{\beta^*} + (1-\pi^*)E_0\left(\left(\frac{\partial p_0}{\partial\beta_r}\right)\left(\frac{\partial p_0}{\partial\beta_s}\right)\right)_{\beta^*} = (L_{rs}(\beta^*))$$

which is equal to L and assumed to exist by hypothesis. Therefore, the left-hand side of (6.9) is o(1) and (N2) is established

In a similar manner, it follows that

$$(6.10) \quad \frac{1}{n}\sum_{i=1}^{n}\int\left(\frac{\partial p_i(x_i,\beta)}{\partial\beta_r} - E_i\frac{\partial p_i(X_i,\beta)}{\partial\beta_r}\right)\left(\frac{\partial p_i(x_i,\beta)}{\partial\beta_s} - E_i\frac{\partial p_i(X_i,\beta)}{\partial\beta_s}\right)' dF_i$$

$$\to \pi^*\int\left(\frac{\partial p_1(x,\beta)}{\partial\beta_r} - E_1\frac{\partial p_1(X,\beta)}{\partial\beta_r}\right)\left(\frac{\partial p_1(x,\beta)}{\partial\beta_s} - E_1\frac{\partial p_1(X,\beta)}{\partial\beta_s}\right)' f_1(x;\pi^0)dx$$

$$+ (1-\pi^*)\int\left(\frac{\partial p_0(x,\beta)}{\partial\beta_r} - E_0\frac{\partial p_0(X,\beta)}{\partial\beta_r}\right)\left(\frac{\partial p_0(x,\beta)}{\partial\beta_s} - E_0\frac{\partial p_0(X,\beta)}{\partial\beta_s}\right)' f_0(x;\pi^0)dx.$$

Let the right-hand side of (6.10), evaluated at β^*, be denoted by $V = (V_{rs}(\beta^*))$. Then

$$(V_{rs}(\beta^*)) = (L_{rs}(\beta^*)) - (K_{rs}(\beta^*)),$$

where

$$(K_{rs}(\beta^*)) = \pi^* E_1\left(\frac{\partial p_1(X,\beta)}{\partial \beta_r}\right)_{\beta^*} E_1\left(\frac{\partial p_1(X,\beta)}{\partial \beta_s}\right)'_{\beta^*}$$

$$+ (1-\pi^*) E_0\left(\frac{\partial p_0(X,\beta)}{\partial \beta_r}\right)_{\beta^*} E_0\left(\frac{\partial p_0(X,\beta)}{\partial \beta_s}\right)'_{\beta^*}$$

A straightforward calculation reveals that $(K_{rs}(\beta^*))$ is the matrix K given in (6.4). Clearly, (N3) is satisfied.

To apply the results of Theorem 3.1, we first note that, by (6.2)

$$u = \lim_{n\to\infty} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} E\left(\frac{\partial p_i}{\partial \beta_r}\right)_{\beta^*}$$

$$= \lim_{n\to\infty} \sqrt{n}\left(\frac{n_1}{n} E_1\left(\frac{\partial p_1(X,\beta)}{\partial \beta_r}\right)_{\beta^*} + \frac{n_0}{n} E_0\left(\frac{\partial p_0(X,\beta)}{\partial \beta_r}\right)_{\beta^*}\right) = 0.$$

The last equality follows from the assumption that $(\pi^* - n_1/n) = O(n^{-1/2})$. Finally, the covariance matrix is

$$J_S = L^{-1} V L^{-1} = L^{-1} - L^{-1} K L^{-1}.$$

This concludes the proof of the theorem. □

For logistic regression, i.e. $G(t) = e^t/(1+e^t)$, some of the assumptions in the above theorem are automatically satisfied and the form of the matrix J_S can be simplified. The result is given in the following theorem.

Theorem 6.2. Suppose that regularity conditions (R1) and (R4) hold, the integral defined by (6.12) exists and

$$\int \frac{\exp(L(x,\beta))}{[1+\exp(L(x,\beta))]^2} [h(x)][h(x)]' f(x)dx$$

is uniformly bounded above. Here

$$f(x) = \pi^* f_1(x;\eta^0) + (1-\pi^*) f_0(x;\eta^0).$$

Then, for stratified sampling and logistic regression,

(i) a sequence of the M-estimators, $\{\hat{\beta}^n\}$, exists and is unique with probability tending to 1 as $n\to\infty$;

(ii) $\hat{\beta}^n \xrightarrow{a.s.} \beta^*$; and

(iii) $\sqrt{n}\,(\hat{\beta}^n-\beta^*) \to N_q(0,J_S)$,

where

$$(6.11) \qquad J_S = L^{-1} - \frac{1}{\pi^*(1-\pi^*)} E_{11},$$

$$(6.12) \qquad L = \pi^*(1-\pi^*) \int \frac{f_1(x;\eta^0) f_0(x;\eta^0)[h(x)][h(x)]'}{\pi^* f_1(x;\eta^0) + (1-\pi^*) f_0(x;\eta^0)}\, dx,$$

and E_{11} is the qxq matrix with one in the (1,1) position and zero elsewhere.

Proof. This theorem follows from the previous one in the same way that Theorem 5.2 follows from Theorem 5.1. It remains to verify the simplified form for the covariance matrix.

First, using the fact that $G' = G(1-G)$, (6.3) can be reduced to

$$L = \int G'(L(x,\beta^*))[h(x)][h(x)]' f(x)dx.$$

which is equal to (6.12). Let the columns of L be denoted by ℓ_r, r=1,...,q. Then the matrix K in (6.4) can be rewritten as

$$(6.13) \qquad K = \frac{1}{\pi^*(1-\pi^*)}\, \ell_1 \ell_1'.$$

Therefore,

$$L^{-1}K = \frac{1}{\pi^*(1-\pi^*)} \begin{pmatrix} \ell_1' \\ 0 \end{pmatrix}$$

and

$$L^{-1}KL^{-1} = \frac{1}{\pi^*(1-\pi^*)}\, E_{11}.$$

Substitution into the formula for J_S given in Theorem 6.1 completes the proof of this theorem. □

As noted by Anderson (1972), the bias in the SS estimator can be removed by adjusting the coefficient of the constant term. The following two theorems give the asymptotic distribution theory for the adjusted estimators.

Theorem 6.3. Suppose that π is known, regularity conditions (R1) and (R4) hold, the integral defined by (6.12) exists and

$$\int \frac{\exp(L(x,\beta))}{[1+\exp(L(x,\beta))]^2}\, [h(x)][h(x)]'\, f(x)dx$$

is uniformly bounded above. Here,

$$f(x) = \pi^* f_1(x;\eta^0) + (1-\pi^*) f_0(x;\eta^0).$$

Let

$$(6.14) \qquad \tilde{\beta}^n = \hat{\beta}^n + \tilde{\beta}_0$$

where

$$\tilde{\beta}_0 = (\log \frac{\pi(1-\pi_n^*)}{(1-\pi)\pi_n^*}, 0,\ldots,0),$$

$$\pi^* = \lim_{n\to\infty} (n_1/n),$$

and

$$\pi^* - (n_1/n) = o(n^{-1/2}).$$

Then, for stratified sampling and logistic regression,

(i) the sequence of M-estimators, $\{\tilde{\beta}^n\}$, exists and is unique with probability tending to 1 as $n\to\infty$;

(ii) $\tilde{\beta}^n \xrightarrow{a.s.} \beta^0$; and

(iii) $\sqrt{n}\,(\tilde{\beta}^n - \beta^0) \xrightarrow{\mathcal{L}} N_q(0, J_S)$,

where J_S is given by (6.11).

Proof. Since G is the logistic function,

$$\frac{\pi f_1(x)}{(1-\pi)f_0(x)} = e^{(\beta^0)'h(x)}$$

Equivalently,

$$\log(\tfrac{\pi}{1-\pi}) + \log \frac{f_1(x)}{f_0(x)} = (\beta^0)'h(x).$$

A similar formula holds for (π^*,β^*). Therefore,

$$\log \frac{\pi(1-\pi^*)}{(1-\pi)\pi^*} = (\beta^0-\beta^*)'h(x).$$

Since, $h_0(X) \equiv 1$ and assumption (R1) holds,

$$\beta^0 - \beta^* = \tilde{\beta}_0.$$

The conclusions of the theorem follow directly from the fact that the adjustment $\tilde{\beta}_0$ is a constant, i.e., non-random. □

In many practical applications, π is not known but is estimated from an independent random sample for which the X's are not measured. We assume that $Y_1,\ldots,Y_m$ are i.i.d. Bernoulli random variables with parameter π and that this sample is independent of the SS. Let $\hat{\pi} = \bar{Y}$. We further assume that

$$\lim_{n\to\infty} \frac{n}{m} = R$$

where R is finite. The following theorem gives the adjusted estimator and its asymptotic theory.

Theorem 6.4. Suppose that regularity conditions (R1) and (R4) hold, the integral defined by (6.12) exists and the uniform boundedness condition of the previous theorem holds. Furthermore, suppose that π is estimated by $\hat{\pi}$ from a independent sample as described above. Let

(6.15) $$\tilde{\beta}^n = \hat{\beta}^n + \tilde{\beta}_0$$

where

$$\tilde{\beta}_0 = \left(\log \frac{\hat{\pi}(1-\pi_n^*)}{(1-\hat{\pi})\pi_n^*}, 0,\ldots,0\right)$$

and $\pi_n^* = n_1/n$. Then, for stratified sampling and logistic regression,

(i) the sequence of M-estimators, $\{\tilde{\beta}^n\}$, exists and is unique with probability tending to 1 as $n\to\infty$;

(ii) $\tilde{\beta}^n \xrightarrow{a.s.} \beta^0$; and

(iii) $\sqrt{n}\,(\tilde{\beta}^n-\beta^0) \xrightarrow{\mathcal{L}} N_q(0,J_S + \frac{R}{\pi(1-\pi)} E_{11})$,

where J_S is given by (6.11).

Proof. Note that the adjustment $\tilde{\beta}_0$ is random. It converges almost surely to the "true" adjustment given in the previous theorem. The obvious adjustment to the asymptotic covariance matrix is a consequence of the independence assumption.

The following theorem concerns the multivariate normal, equal covariance matrix case. The notation is the same as that given in section 5.

Theorem 6.5. Suppose that X given Y is $N_k(\mu_Y,\Sigma)$. Then, for stratified sampling and logistic regression,

(i) the sequence of the M-estimators, $\{\hat{\beta}^n\}$, exists and is unique with probability tending to 1 as $n\to\infty$;

(ii) $\hat{\beta}^n \xrightarrow{a.s.} \beta^*$, and

(iii) $\sqrt{n}\,(\hat{\beta}^n-\beta^*) \xrightarrow{\mathcal{L}} N_q(0,J_S)$,

where

$$(6.16) \qquad J_S = L^{-1} - \frac{1}{\pi^*(1-\pi^*)} E_{11},$$

$$L = \pi^*(1-\pi^*) \int \frac{f_1(x)f_0(x)}{\pi^* f_1(x) + (1-\pi^*)f_0(x)} \binom{1}{x}(1,x')dx,$$

$$\beta^* = (\beta^*, \delta')'$$

$$\beta_0^* = \log\left(\frac{\pi^*}{1-\pi^*}\right) - \frac{1}{2}(\mu_1' \Sigma^{-1} \mu_1 - \mu_0' \Sigma^{-1} \mu_0),$$

and

$$\delta = \Sigma^{-1}(\mu_1 - \mu_0).$$

Proof. The proof follows from that of Theorem 5.3 with the obvious modifications. □

The following two theorems concern adjustments to the estimator for the cases where π is known and unknown.

Theorem 6.6. Suppose that π is known and X given Y is $N_k(\mu_Y, \Sigma)$. Let $\tilde{\beta}^n$ be defined by (6.14). Then, for stratified sampling and logistic regression,

(i) the sequence of M-estimators, $\{\tilde{\beta}^n\}$, exists and is unique with probability tending to 1 as $n \to \infty$;

(ii) $\tilde{\beta}_n \xrightarrow{a.s.} \beta^0$; and

(iii) $\sqrt{n}(\tilde{\beta}^n - \beta^0) \xrightarrow{\mathcal{L}} N_q(0, J_S)$

where J_S is given by (6.16).

Proof. The theorem follows from Theorem 6.3 using arguments given in the proof of Theorem 5.3. □

Theorem 6.7. Suppose that X given Y is $N_k(\mu_Y, \ddagger)$ and that π is estimated by $\hat{\pi}$ as described previously. Let $\tilde{\beta}^n$ be defined by (6.15). Then, for stratified sampling and logistic regression,

(i) the sequence of M-estimators, $\{\tilde{\beta}^n\}$, exists and is unique with probability tending to 1 as $n \to \infty$;

(ii) $\tilde{\beta}^n \xrightarrow{a.s.} \beta^0$; and

(iii) $\sqrt{n}\,(\tilde{\beta}^n - \beta^0) \xrightarrow{\mathcal{L}} N_q(0, J_S + \frac{R}{\pi(1-\pi)} E_{11})$,

where J_S is given by (6.16).

Proof. The theorem follows from Theorem 6.4 using arguments given in the proof of Theroem 5.3. □

The following three theorems concern the normal, unequal covariance matrix case. The proofs follow the arguments given in Theorem 5.4 and Theorems 6.2, 6.3 and 6.4, respectively.

Here,

$$L(X,\beta^*) = \beta_0^* + \delta'X + X'BX,$$

$$L(X,\beta) = \beta_0 + \delta'X + X'BX,$$

$$\beta_0^* = \log \frac{\pi^*|\Sigma_0|^{1/2}}{(1-\pi^*)|\Sigma_1|^{1/2}} - \frac{1}{2}(\mu_1'\Sigma_1^{-1}\mu_1 - \mu_0\Sigma_0^{-1}\mu_0),$$

$$\delta = \Sigma_1^{-1}\mu_1 - \Sigma_0^{-1}\mu_0,$$

$$B = \frac{1}{2}(\Sigma_0^{-1} - \Sigma_1^{-1}),$$

$$h(X) = (1, X_1, \ldots, X_k, X_1^2, 2X_1X_2, \ldots, 2X_1X_k, X_2^2, 2X_2X_3, \ldots, X_k^2),$$

and

$$q = 1 + 2k + k(k-1)/2.$$

Theorem 6.8. Suppose that X given Y is $N_k(\mu_Y, \Sigma_Y)$. Then for stratified sampling and logistic regression

(i) the sequence of M-estimators, $\{\hat{\beta}^n\}$, exists and is unique with probability tending to 1 as $n\to\infty$;

(ii) $\hat{\beta}^n \xrightarrow{a.s.} \beta^*$; and

(iii) $\sqrt{n}(\hat{\beta}^n - \beta^*) \xrightarrow{\mathcal{L}} N_q(0, J_S)$,

where

$$(6.17) \qquad J_S = L^{-1} - \frac{1}{\pi^*(1-\pi)} E_{11},$$

and

$$L = \pi^*(1-\pi^*) \int \frac{f_1(x)f_0(x)][h(x)][h(x)]'}{\pi^* f_1(x) + (1-\pi^*)f_0(x)} dx.$$

Theorem 6.9. Suppose that π is known and that X given Y is $N_k(\mu_Y, \Sigma_Y)$. Let $\tilde{\beta}^n$ be defined by (6.14). Then, for stratified sampling and logistic regression,

(i) the sequence of M-estimators, $\{\tilde{\beta}^n\}$, exists and is unique with probability tending to 1 as $n \to \infty$;

(ii) $\tilde{\beta}^n \xrightarrow{a.s.} \beta^0$; and

(iii) $\sqrt{n}\,(\tilde{\beta}^n - \beta^0) \xrightarrow{\mathcal{L}} N_q(0, J_S)$,

where J_S is defined by (6.17).

Theorem 6.10. Suppose that X given Y is $N_k(\mu_Y, \Sigma_Y)$ and that π is estimated by $\hat{\pi}$ as described previously. Let $\tilde{\beta}^n$ be defined by (6.15). Then, for stratified sampling and logistic regression,

(i) the sequence of M-estimators, $\{\tilde{\beta}^n\}$, exists and is unique with probability tending to 1 as $n \to \infty$;

(ii) $\tilde{\beta}^n \xrightarrow{a.s.} \beta^0$; and

(iii) $\sqrt{n}\,(\tilde{\beta}^n - \beta^0) \xrightarrow{\mathcal{L}} N_q(0, J_S + \frac{R}{\pi(1-\pi)} E_{11})$,

where J_S is defined by (6.17).

REFERENCES

Aitchison, J. & Silvey, S. D. (1958). Maximum likelihood estimation of parameters subject to restraints. Ann. Math. Statist. 29, 813-28.

Anderson, J. A. (1972). Separate sample logistic discrimination. Biometrika, 59, 19-35.

Anderson, T. W. (1958). An Introduction to Multivariate Statistical Analysis. Wiley, New York.

Bradley, R. A. and Gart, J. J. (1962). The asymptotic properties of maximum likelihood estimators when sampling from associated populations. Biometrika, 49, 205-213.

Chanda, K. C. (1954). A note on the consistency and maxima of the roots of likelihood equations. Biometrika, 41, 56-61.

Cox, D. R. (1966). Some procedures associated with the logistic qualitative response curve. In "Research Papers on Statistics: Festschrift for J. Neyman", ed. by F. N. David, pp. 55-71. Wiley, New York.

Cramér, H. (1946). Mathematical Methods of Statistics. Princeton University Press.

Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. J. Amer. Statist. Assoc. 70, 892-898.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. Proc. Fifth Berkeley Symp. Math. Statist. Prob., 1, 221-233.

Inagaki, N. (1973). Asymptotic relations between the likelihood estimating function and the maximum likelihood estimator. Annals of the Institute of Statistical Mathematics. 25, 1-26.

Lachenbruch, P.A. (1975). Discriminant Analysis. Hafner, New York.

O'Neill, J. J. (1980). The general distribution of the error rate of a classification procedure with application to logistic regression discrimination. J. Amer. Statist. Assoc., 75, 156-160.

Serfling, R. J. (1980). Approximation Theorems of Mathematical Statistics. Wiley, New York.

Silvapulle, M. J. (1981). On the existence of maximum likelihood estimators for the binomial response models. J. R. Statist. Soc. B., 43, 310-313.

Tarone, R. E. (1974). Asymptotic Properties of Maximum Likelihood Estimators in the General Sampling Framework and Some Results in Non-normal Linear Regression. Ph.D. thesis, Department of Mathematics, U. of California at Davis.