A COMPARISON OF LOGISTIC REGRESSION AND MAXIMUM LIKELIHOOD CLASSIFICATION METHODS

BY

TERENCE J. O'NEILL

TECHNICAL REPORT NO. 12
AUGUST 15, 1975

PREPARED UNDER THE AUSPICES OF
PUBLIC HEALTH SERVICE GRANT 1 RO1 GM21215-01

DIVISION OF BIOSTATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
A COMPARISON OF LOGISTIC REGRESSION AND MAXIMUM LIKELIHOOD CLASSIFICATION METHODS

by

TERENCE J. O'NEILL

TECHNICAL REPORT NO. 12
AUGUST 15, 1975

PREPARED UNDER THE AUSPICES
OF
PUBLIC HEALTH SERVICE GRANT 1 RO1 GM21215-01

DIVISION OF BIOSTATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
A Comparison of Logistic Regression and Maximum Likelihood Classification Methods

by

Terence J. O'Neill
Stanford University

Abstract

This paper discusses classification into one of two members of a general exponential family. The strong consistency and asymptotic distribution of the logistic regression estimates is established. The asymptotic distribution of the probability of misclassification of an observation by an estimate of the optimal boundary is found. This is used to compare logistic regression and maximum likelihood estimation.
A Comparison of Logistic Regression and Maximum Likelihood Classification Methods

by

Terence J. O'Neill*
Stanford University

1. INTRODUCTION AND SUMMARY

Consider a random variable $\tilde{x}$ which arises from one of two populations with densities

$$
\begin{align*}
\tilde{f}_1(\tilde{x}) &= g(\theta_1,\tilde{\eta}) h(\tilde{x},\tilde{\eta}) e^{\theta_1 \tilde{x}} \quad \text{with probability } \pi_1 \\
\tilde{f}_0(\tilde{x}) &= g(\theta_0,\tilde{\eta}) h(\tilde{x},\tilde{\eta}) e^{\theta_0 \tilde{x}} \quad \text{with probability } \pi_0
\end{align*}
$$

(1.1)

Let the associated distribution functions be $F_0$ and $F_1$, and let $\tilde{x}$ have distribution function $F$. Suppose that $\tilde{x}$ has a finite covariance matrix. The random variable $y$ indicates which of the distributions, $F_0$ or $F_1$, gives rise to $\tilde{x}$. Then

$$
y = \begin{cases} 
1 & \text{with probability } \pi_1 \\
0 & \text{with probability } \pi_0
\end{cases},
$$

and

$$
f(\tilde{x}|y) = f_y(\tilde{x}).
$$

Thus

* This research was carried out while the author held a C.S.I.R.O. Postgraduate Studentship in Statistics.
\[ P(y = i|x) = \begin{cases} \pi_1(x), & i = 1 \\ 1 - \pi_1(x) = \pi_0(x), & i = 0 \end{cases} \]

where

\[ \pi_1(x) = \frac{e^{\beta_0 + \beta x}}{1 + e^{\beta_0 + \beta x}} \]

\[ \beta_0 = \log \frac{\pi_1}{\pi_0} + \log \frac{g(\theta_1, \eta)}{g(\theta_0, \eta)} \]

and

\[ \beta = \theta_1 - \theta_0 \cdot \]

The classification problem is to decide which population gives rise to an observed \( x \). If \((\beta_0, \beta')\) is known, then it is shown by Anderson [1] that the classification procedure which minimizes the expected probability of misclassification decides whether \( x \) comes from \( f_1 \) or \( f_0 \) according as \( \pi_1(x) \) is greater than or less than \( \pi_0(x) \). This is equivalent to \( \beta_0 + \beta x \) being greater than or less than \( 0 \). In practice, \((\beta_0, \beta')\) is not usually known, but a random sample from (1.1) is available and it is used to estimate \((\beta_0, \beta')\). One possible estimate is the maximum likelihood estimate. Another is the logistic regression estimate of Cox [3].

Let \( \{(x_1, y_1), \ldots, (x_n, y_n)\} \) be the random sample from (1.1). Then the logistic regression estimates, denoted \((\beta_0, \beta')\), are defined by
\[
\sum_{i=1}^{n} \frac{e^{\beta'_{0} (x_{i} \beta_{0})}}{1 + e^{\beta'_{0} (x_{i} \beta_{0})}} = \sum_{i=1}^{n} \left( \frac{1}{x_{i} \beta_{0}} \right) y_{i}.
\]

It is shown in this paper that (1.2) has a unique solution for \( n \geq p + 1 \) and also that the logistic regression estimates are strongly consistent. Their asymptotic distribution is found.

The maximum likelihood estimates change in form with the exponential family. They will estimate the wrong parameters when the underlying family is not the assumed one. In this case, they cannot be expected to asymptotically minimize the expected probability of misclassification. For example, Fisher's linear discriminant function will not have any asymptotic virtue if the underlying family is not normal with constant covariance matrix. Thus the logistic regression estimates are more robust than the maximum likelihood estimates.

Recently Efron [4] computed the asymptotic relative efficiency of the two procedures in the standard normal case. In this paper, the asymptotic distribution of the probability of error associated with an estimate of \((\beta'_{0}, \hat{\beta'})\) is obtained. This is used to compare the maximum likelihood and logistic regression methods.

2. **STRONG CONSISTENCY OF THE LOGISTIC REGRESSION ESTIMATES**

The following lemmas will be necessary.

**Lemma 1**

Let \( G_{n} \) be the empirical c.d.f. of a sample of size \( n \) from the c.d.f. \( F \) on \( \mathbb{R}^{p} \). Suppose \( \{ h_{\theta}(x), \theta \in \Theta \} \) is a family of real valued
functions on \( \mathbb{R}^p \) such that

\[ |h_\theta(x)| \leq h(x), \text{ for all } x \in \mathbb{R}^p, \]

where \( h(x) \) is bounded on compact sets and

\[ \int h(x) dF(x) < \infty. \]

Suppose further that for any compact set \( \mathcal{A} \subseteq \mathbb{R}^p \),

\[ \left| \sum_{i=1}^p h_\theta(x) \right| \leq k_{A_i} \nu(x, z), \]

where

\[ \nu(x, z) = \prod_{i=1}^p |y_i - z_i|. \]

Then

\[ \sup_{x \in \mathcal{A}} \left| \int h_\theta(x) d(G_n - F)(x) \right| \xrightarrow{a.s.} 0. \]

Proof

The proof will only be done for \( p = 1 \). It follows in exactly the same manner for higher dimensions.

Let \( b_\varepsilon \) satisfy

\[ \left\{ \begin{array}{c}
\int_{k \notin (-b_\varepsilon, b_\varepsilon]} h(x) dF(x) < \varepsilon.
\end{array} \right. \]

Let

\[ i_n = \min \{ i; x_{(i)} \in (-b_\varepsilon, b_\varepsilon] \}, \]

\[ j_n = \max \{ i; x_{(i)} \in (-b_\varepsilon, b_\varepsilon] \}, \]

where \( \{ x_{(1)}, \ldots, x_{(n)} \} \) is the order statistic. Then by (2.1), (2.3) and the strong law of large numbers,
\[
\lim \sup_{n \to \infty} \left| \int h_\theta(x) d(G_n - F)(x) \right| \overset{a.s.}{\leq} \]

\[
\lim \sup_{n \to \infty} \left| \sum_{i=1}^{I_n} \int (x_{(i-1)}, x_{(i)}) h_\theta(x) d(G_n - F)(x) \right|
\]

\[
+ \lim \sup_{n \to \infty} (\sup\{h(x); x \in \{x_{(i-1)}; b \}\})(|G_n - F| \{x_{(i-1)}, -b \})
\]

\[
+ |G_n - F| \{x_{(j_n)}, b \} + 2 \varepsilon.
\]

But

\[
|G_n - F| \{x_{(i-1)}, -b \} + |G_n - F| \{x_{(j_n)}, b \} \overset{a.s.}{\to} 0.
\]

Also

\[
\left| \sum_{i=1}^{I_n} \int (x_{(i-1)}, x_{(i)}) h_\theta(x) d(G_n - F)(x) \right|
\]

\[
\leq \left| \sum_{i=1}^{I_n} (x_{(i-1)}, x_{(i)}) h_\theta(x_{(i)}) d(G_n - F)(x) \right|
\]

\[
+ \sum_{i=1}^{I_n} \int (h_\theta(x_{(i)} - h_\theta(x)) dF(x)
\]

\[
\leq \left| \sum_{i=1}^{I_n} (x_{(i-1)}, x_{(i)}) h_\theta(x_{(i)}) d(G_n - F)(x) \right|
\]

\[
+ \sup_{x, y} \left| h_\theta(x) - h_\theta(y) \right| \quad |x - y| < \delta
\]
where
\[
\delta = \sup_{i = i_n, \ldots, j_n} |x(i) - x(i-1)| \quad \text{a.s.} \quad \rightarrow 0.
\]
Hence by (2.2),
\[
\sup_{x, y} |h_\theta(x) - h_\theta(y)| \xrightarrow{\text{a.s.}} 0.
\]
Now
\[
\sum_{i = i_n}^{j_n} \left( |\sum_{i = i_n}^{j_n} (x(i) - x(i-1)) \, d(G_n^{-F})(x) - h_\theta(x(i))| \right) \leq \sum_{i = i_n}^{j_n} |h_\theta(x(i+1))| - h_\theta(x(i)))(G_n^{-F})(x) \right) + |h_\theta(x(i))|\left( (G_n^{-F})(x(i-1)) \right) + |h_\theta(x(i))|\left( (G_n^{-F})(x(i+1)) \right) - h_\theta(x(i))(G_n^{-F})(x) \right) + |h_\theta(x(i+1))|\left( (G_n^{-F})(x(i)) \right)
\]
But
\[
(G_n^{-F})(x(i)) \xrightarrow{\text{a.s.}} 0, \ i = i_n - 1, j_n.
\]
Thus
\[
|h_\theta(x(i))|\left( (G_n^{-F})(x(i)) \right) + |h_\theta(x(i+1))|\left( (G_n^{-F})(x(i-1)) \right) \xrightarrow{\text{a.s.}} 0.
\]
Also
\[
\limsup_{n \to \infty} \left( \sum_{i = i_n}^{j_n} |h_\theta(x(i+1)) - h_\theta(x(i)))(G_n^{-F})(x) \right) \leq \limsup_{n \to \infty} \sup_{x \in \mathbb{R}} |(G_n^{-F})(x)|
\]
and by (2.2),

\[
\leq 2k_{b, b} \lim_{n \to \infty} \sup_{x \in R} \left| (G_n - F)(x) \right|
\]

This bound holds for all \( \theta \) and it converges almost surely to 0 by the Glivenko Cantelli theorem. This completes the proof of Lemma 1.

**Corollary 1**

Suppose \( \Theta \) is an interval in \( R^p, \{h_{\theta} (x)\} \) are differentiable and

\[
(2.4) \quad \left| \frac{\partial^p h_{\theta} (x)}{\partial \theta_1 \cdots \partial \theta_p} \right| \leq M_A < \infty,
\]

for all \( \theta \in A \), for all compact sets \( A \subset R^p \). Also suppose

\[
|h_{\theta} (x)| \leq h(x)
\]

where

\[
\int h(x) dF < \infty.
\]

Then

\[
\sup_{\theta \in H} \left| \int h_{\theta} (x) d(G_n - F)(x) \right| \xrightarrow{a.s.} 0.
\]

**Proof**

\[
\left| \Delta h_{\theta} (x) \right| = \left| \frac{\partial^p h_{\theta} (x)}{\partial \theta_1 \cdots \partial \theta_p} \right| u(y, z),
\]

for some point \( z \) in the region formed by \( y \) and \( z \) (i.e. the region with vertices \( s \) where \( s_i = y_i \) or \( z_i \)). The result follows by (2.4) and Lemma 1.

**Lemma 2**

Let \( f \) be a continuous function from \( R^p \) to \( R^p \) defined by
\[
    f(x) = \begin{pmatrix}
    f_1(x) \\
    \vdots \\
    f_p(x)
    \end{pmatrix}
\]

Suppose the Jacobian matrix, \( J_f \), where
\[
    J_f = \begin{bmatrix}
    \frac{\partial f_1}{\partial x_j} \\
    \vdots \\
    \frac{\partial f_p}{\partial x_j}
    \end{bmatrix} \quad i = 1, \ldots, p \\
    j = 1, \ldots, p
\]
is symmetric and positive definite. Then \( f \) is one to one, continuous, and is an open mapping. Thus \( f^{-1} \) exists and is continuous on \( f(R^p) \).

**Proof**

Suppose there exist distinct \( x_1, x_2 \in R^p \) such that
\[
    (2.5) \quad f(x_1) = f(x_2)
\]

Let \( \Gamma \) be an orthogonal matrix with first row \( \frac{(x_1' - x_0')/||x_1 - x_0||}{x_2} \).
Define the function \( g : R^p \to R^p \) by
\[
    g(x) = \Gamma f (\Gamma' x + x_0').
\]
Then
\[
    J_g = \Gamma J_f \Gamma' > 0.
\]
Hence
\[
    (2.6) \quad (J_g)'_{11} > 0
\]
But by (2.5),
\[ g \left( \begin{bmatrix} x_1 - x_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \right) = g(0). \]

This contradicts (2.6). Hence \( f \) is one to one. Thus by Apostol [2], page 371, \( f \) is an open mapping. This completes the proof of Lemma 2.

**Corollary 2**

The function

\[ g(\gamma) = E \left[ \frac{\gamma' \left( \frac{1}{x} \right)}{1 + e^{\gamma' \left( \frac{1}{x} \right)}} \right] \]

where \( x \) is distributed according as (1.1) has a continuous inverse on \( g(\mathbb{R}^{p+1}) \).

**Proof**

The continuity of \( g \) follows from the dominated convergence theorem. Clearly \( J_g \) is symmetric. Hence it remains to show it is positive definite. But

\[ a' J_g a = E \left[ \frac{\gamma' \left( \frac{1}{x} \right)}{1 + e^{\gamma' \left( \frac{1}{x} \right)}} \left( a' \left( \frac{1}{x} \right) \right)^2 \right]. \]

But by (1.1), this is positive. Hence \( J_g \) is positive definite and the result follows by Lemma 2.
Corollary 3

For \( n \geq p + 1 \), the function

\[
g_n(\gamma) = \frac{1}{n} \Sigma_{i=1}^{n} \frac{\gamma'(1/x_i)}{1+e^{\gamma'(1/x_i)}} (1/x_i)
\]

has a continuous inverse almost surely. Hence the logistic regression estimates are well defined for \( n \geq p + 1 \).

Proof

\( g_n \) is clearly continuous. To show that

\[
\frac{1}{n} \Sigma_{i=1}^{n} \exp \left( \frac{\gamma'(1/x_i)}{1+e^{\gamma'(1/x_i)}} \right) \left( \frac{1}{x_i} \right) (1/x_i)'
\]

is positive definite it suffices to show that

\[ P(a \sim x = 0) = 0 \]

where \( a \sim 0 \) is a fixed vector in \( \mathbb{R}^{p+1} \). But this follows from (1.1). Thus the result follows from Lemma 2.

Theorem 2

\[
\left( \tilde{\beta}_0 \right) \xrightarrow{a.s.} \left( \beta_0 \right).
\]

Proof

Let \( G_n \) be the empirical c.d.f. of \( x_1, \ldots, x_n \). Then by Corollary 1,
\[
\sup_{\gamma \in E} \lim_{p \to \infty} \int_{\mathbb{R}^p} \frac{\exp\left(\frac{\beta_0 + \beta_1 \gamma}{\gamma^*}\right)}{1 + \exp\left(\frac{\beta_0 + \beta_1 \gamma}{\gamma^*}\right)} \left(\frac{1}{\gamma^*}\right) d(G_n - F_0)(\gamma) \quad \text{a.s.} \quad 0.
\]

By the strong law of large numbers,

\[
\sum_{i=1}^{n} \left(\frac{1}{\gamma_i}\right) y_i \quad \text{a.s.} \quad E\left[\frac{1}{\gamma}\right].
\]

But

\[
\int_{\mathbb{R}^p} \frac{\exp\left(\frac{\beta_0 + \beta_1 \gamma}{\gamma^*}\right)}{1 + \exp\left(\frac{\beta_0 + \beta_1 \gamma}{\gamma^*}\right)} \left(\frac{1}{\gamma^*}\right) d\gamma = \sum_{i=1}^{n} \left(\frac{1}{\gamma_i}\right) y_i
\]

Thus by (2.7) and (2.8),

\[
\int_{\mathbb{R}^p} \frac{\exp\left(\frac{\beta_0 + \beta_1 \gamma}{\gamma^*}\right)}{1 + \exp\left(\frac{\beta_0 + \beta_1 \gamma}{\gamma^*}\right)} \left(\frac{1}{\gamma^*}\right) dF(\gamma) \quad \text{a.s.} \quad E\left[\frac{1}{\gamma}\right].
\]

Hence

\[
\lim_{n \to \infty} g\left(\frac{\gamma}{\gamma^*}\right) \quad \text{a.s.} \quad E\left[\frac{1}{\gamma}\right]
\]

\[
= \frac{g\left(\beta_0\right)}{\left(\gamma^*\right)}.
\]

Thus by Corollary 2,
\[ \lim_{n \to \infty} \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_0 \\ \vdots \\ \hat{\beta}_0 \end{pmatrix} \overset{a.s.}{=} \begin{pmatrix} \beta_0 \\ \beta_0 \\ \vdots \end{pmatrix}. \]

This completes the proof of Theorem 1.

3. **The Asymptotic Distribution of the Logistic Regression Estimates**

The following lemma will be necessary.

**Lemma 3**

Let

\[ H(\gamma, n) = \frac{1}{n} \sum_{i=1}^{n} \frac{e^{(1x_i)\gamma}}{(1+e^{(1x_i)\gamma})^2} \begin{pmatrix} 1 \\ \frac{1}{x_i} \end{pmatrix} \begin{pmatrix} 1 \\ \frac{1}{x_i} \end{pmatrix} \]

Suppose

\[ (3.1) \quad \mathbf{U} \overset{a.s.}{\to} \gamma. \]

Then

\[ H(\mathbf{U}, n) \overset{a.s.}{\to} \Sigma, \]

where

\[ \Sigma = \int_{\mathbb{R}^p} \frac{e^{(1x_\widetilde{\gamma})}}{(1+e^{(1x_\widetilde{\gamma})})^2} \begin{pmatrix} 1 \\ \frac{1}{x_\widetilde{\gamma}} \end{pmatrix} \begin{pmatrix} 1 \\ \frac{1}{x_\widetilde{\gamma}} \end{pmatrix} dF(\widetilde{x}). \]

**Proof**

By Corollary 1,
\[(3.2)\quad \int_{\mathbb{R}^p} \frac{(1 \mathbf{x}^\prime \mathbf{U}_n \mathbf{n})}{(1 + e^{-1} \mathbf{x}^\prime \mathbf{U}_n \mathbf{n})^2} \left(\frac{1}{1 + e^{-1} \mathbf{x}^\prime \mathbf{U}_n \mathbf{n}}\right)^{1/2} d(G_{n - F})(\mathbf{x}) \xrightarrow{a.s.} 0.\]

But by (3.1) and the dominated convergence theorem,

\[(3.3)\quad \int_{\mathbb{R}^p} \left(\frac{e^{1/2} \mathbf{x}^\prime \mathbf{U}_n \mathbf{n}}{(1 + e^{-1} \mathbf{x}^\prime \mathbf{U}_n \mathbf{n})^2} - \frac{e^{1/2} \mathbf{x}^\prime \mathbf{Y}}{(1 + e^{-1} \mathbf{x}^\prime \mathbf{Y})^2}\right) \left(\frac{1}{1 + e^{-1} \mathbf{x}^\prime \mathbf{Y}}\right)^{1/2} dF(\mathbf{x}) \xrightarrow{a.s.} 0.\]

Thus by (3.2) and (3.3) the result follows.

**Theorem 2**

\[
\sqrt{n} \left(\frac{\mathbf{b}_0}{\mathbf{b}_0^\prime} - \frac{\mathbf{b}_0^0}{\mathbf{b}_0^0}\right) \xrightarrow{d} N_{p+1}(0, \Sigma_L)
\]

where

\[(3.4)\quad \Sigma_L = \int_{\mathbb{R}^p} \frac{e^{1/2} \mathbf{x}^\prime \mathbf{U}_n \mathbf{n}}{(1 + e^{-1} \mathbf{x}^\prime \mathbf{U}_n \mathbf{n})^2} \left(\frac{1}{1 + e^{-1} \mathbf{x}^\prime \mathbf{Y}}\right)^{1/2} dF(\mathbf{x})\]

**Proof**

Let

\[
n \mathbf{r}_n = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} y_i
\]

Then

\[
E[\mathbf{r}_n | x_1, \ldots, x_n] = g_n \left(\begin{array}{c}
\mathbf{b}_0^0
\mathbf{b}_0^0
\end{array}\right).\]

Hence using Taylor's theorem,
\begin{equation}
\left( r_{\sim n} - g_{\sim n} \left( \beta_0 \right) \right) = H_{\sim n, n} \left( \beta_0 \right) - \left( \beta_0 \right),
\end{equation}

where

\[ || u_{\sim n} - \left( \beta_0 \right)|| \leq || \left( \beta_0 \right) - \left( \beta_0 \right)||. \]

Thus by Theorem 1,

\[ u_{\sim n} \xrightarrow{\text{a.s.}} \left( \beta_0 \right). \]

Hence by Lemma 3,

\begin{equation}
H_{\sim n, n} \xrightarrow{\text{a.s.}} \Sigma_{\sim L}^{-1}.
\end{equation}

Also, by the central limit theorem,

\[ \sqrt{n} \left( r_{\sim n} - g_{\sim n} \left( \beta_0 \right) \right) \xrightarrow{\mathcal{D}} N_{p+1}(0, \Sigma_{\sim L}^{-1}). \]

Thus from (3.5) and (3.6),

\[ \sqrt{n} \left( \left( \beta_0 \right) - \left( \beta_0 \right) \right) \xrightarrow{\mathcal{D}} N_{p+1}(0, \Sigma_{\sim L}). \]

This completes the proof of Theorem 2.

4. THE ASYMPTOTIC DISTRIBUTION OF ERROR RATES

Lemma 4

Let \( g_{\sim} : \mathbb{R}^p \to \mathbb{R}^p \) be a twice continuously differentiable function satisfying

\begin{equation}
\nabla g_{\sim} (\beta) = 0.
\end{equation}

Suppose \( t_{\sim n} \) is a statistic such that

\[ \sqrt{n} (t_{\sim n} - \theta) \xrightarrow{\mathcal{D}} N_p(0, \Sigma). \]

Then
\[ n(\hat{g}_n(t_n) - g_0(\theta)) \overset{p}{\to} Az, \]

where

\[ z \sim N_p(0, \Sigma), \]

and

\[ A = \frac{1}{2} \left( \frac{\partial^2 g_0(y)}{\partial \gamma_1 \partial \gamma_j} \bigg| \theta \right) \]

Proof

The result follows by using a Taylor series expansion of \( g_0(t_n) \) about \( \theta \) and (4.1).

Let \( \text{ER} \left( \left[ \begin{array}{c} Y_0 \\ Y \end{array} \right] \right) \) be the probability of misclassification associated with using \( \left[ \begin{array}{c} Y_0 \\ Y \end{array} \right] \) as the parameters of the classification boundary. Then

\[ \text{ER} \left( \left[ \begin{array}{c} Y_0 \\ Y \end{array} \right] \right) = \pi_0 \int_{Y'x+Y^T0 > 0} f_0(x) dx + \pi_1 \int_{Y'x+Y^T0 \leq 0} f_1(x) dx. \]

Theorem 3

Assume \( \beta_p \neq 0 \). Suppose

\[ \sqrt{n} \left( \left[ \begin{array}{c} Y_0 \\ Y \end{array} \right] - \left( \begin{array}{c} \beta_0 \\ \beta \end{array} \right) \right) \rightarrow N_{p+1}(0, \Sigma) \]

Then

\[ n \left( \text{ER} \left( \left[ \begin{array}{c} Y_0 \\ Y \end{array} \right] \right) - \text{ER} \left( \left[ \begin{array}{c} \beta_0 \\ \beta \end{array} \right] \right) \right) \overset{p}{\to} Az Bz \]

where

\[ z \sim N_{p+1}(0, \Sigma), \]

\[ B = \frac{\pi_1}{2|\beta_p|} \int_{R^{p-1}} \left( \begin{array}{c} 1 \\ h_{\beta_0}(\mu) \end{array} \right) \left( \begin{array}{c} 1 \\ h_{\beta}(\mu) \end{array} \right) \left( \begin{array}{c} f_0_{\beta_0}(\mu) \\ f_1_{\beta}(\mu) \end{array} \right) d\mu, \ldots, d\mu \]

15
and
\[
\left(\begin{array}{c}
\beta_1 \\
\vdots \\
\beta_p
\end{array}\right) \sim \left(\begin{array}{c}
u_1, \ldots, u_{p-1} \\
\beta_0 - \beta_1 u_1 - \cdots - \beta_{p-1} u_{p-1}
\end{array}\right).
\]

Proof
\[
\mathbb{E}(\gamma_0) - \mathbb{E}(\beta_0) = \pi_1 \left\{ \int_{\gamma_0 + \gamma, x < 0} f_1(x)dx \right\}
\]
\[
- \left\{ \beta_0 + \int_{\gamma, x = 0} f_1(x)dx \right\}
\]
\[
+ \pi_0 \left\{ \int_{\gamma_0 + \gamma, x > 0} f_0(x)dx - \int_{\beta_0 + \gamma, x > 0} f_0(x)dx \right\}
\]
\[
= \int_{\gamma_0, x < 0} f(x)dx - \int_{\beta_0, x < 0} f(x)dx,
\]
where
\[
f(x) = \pi_1 f_1(x) - \pi_0 f_0(x).
\]
Let
\[
\mathbb{E}(\beta_0)(\gamma_0) = \int_{\gamma_0, x < 0} f(x)dx.
\]
Then
\[
\mathbb{V}(\beta_0) = -\frac{1}{|\beta_p|} \int_{R^{p-1}} \left(\begin{array}{c}
h_{\beta_0}(u) \\
\beta_0
\end{array}\right) f(h_{\beta_0}(u))du_1, \ldots, du_{p-1},
\]
But

16
\[ f(h_{\beta_0}(u)) = \pi_1 f_1(z) - \pi_0 f_0(z) \]

where \( z \) is a point on the optimum classification boundary. Thus

\[ f(h_{\beta_0}(u)) = 0. \]

Hence (4.1) holds. But

\[ \frac{1}{2} \left( \sum_{i,j} \left( \frac{\partial^2 g_i(\beta_0)}{\partial \gamma_i \partial \gamma_j} \right) \right) = B. \]

The result follows from Lemma 4.

Suppose \( \tilde{s} \) and \( \tilde{t} \) are estimates of \( \left( \beta_0 \right) \) satisfying

\[ n \left( ER(\tilde{s}_n) - ER(\left( \beta_0 \right)) \right) \tilde{s}_n \tilde{z}_s, \]

\[ n \left( ER(\tilde{t}_n) - ER(\left( \beta_0 \right)) \right) \tilde{t}_n \tilde{z}_t, \]

where \( z_s \sim N_{p+1}(0, A_s) \)

\[ z_t \sim N_{p+1}(0, A_t) \]

Then define the asymptotic relative efficiency of \( \tilde{s} \) with respect to \( \tilde{t} \) by

\[ \text{Eff}_{p, \tilde{s}, \tilde{t}} = \frac{E[z_s, z_t]}{E[z_t, z_s]} \]

\[ = \frac{\text{tr} A_t}{\text{tr} A_s}. \]
If
\[ \sqrt{n} \left( s_n - \begin{pmatrix} \beta_0 \\ \beta \end{pmatrix} \right) \sim \mathcal{N}_{p+1}(0, \Sigma) , \]
\[ \sqrt{n} \left( t_n - \begin{pmatrix} \beta_0 \\ \beta \end{pmatrix} \right) \sim \mathcal{N}_{p+1}(0, \Sigma) , \]
then by Theorem 3,
\[ \text{Eff}_{p}(s, t) = \frac{\text{tr} \, \Sigma}{\text{tr} \, \Sigma} . \]

5. EXAMPLES

Example 1 Standard Normal Case

Consider the case where \( f_0 \) and \( f_1 \) are normal densities with the same covariance matrix, but different means. Then since the probability of error does not change under linear transformations, this can be transformed to the "standard situation" of Efron [4],

\[ f_1(x) = N_p \left( \frac{\Delta}{2} \frac{e_p}{p} , I_p \right) \text{ with probability } \pi_1 \]
\[ f_0(x) = N_p \left( -\frac{\Delta}{2} \frac{e_p}{p} , I_p \right) \text{ with probability } \pi_0 \]

where
\[ e_p' = (0, 0, \ldots, 0, 1) , \]
and
\[ \Delta > 0 . \]

Then
\[ \beta_0 = \log \frac{\pi_1}{\pi_0} = \lambda \]
and
\[ \beta' = (0, \ldots, 0, \Delta) . \]

Thus
\[ f_1(h \begin{pmatrix} \beta_0 \\ \beta \end{pmatrix}) = \left( \prod_{i=1}^{p-1} \phi(x_i) \right) \phi(h \beta) . \]
where $\phi$ is the standard normal density function

and

$$D_1 = \frac{A}{2} + \frac{\lambda}{\Delta}.$$

Thus

$$B = \frac{1}{2 D_1} \left( I_{p} \begin{pmatrix} \frac{-\lambda}{\Delta} \\ 0 \end{pmatrix}, \begin{pmatrix} -\frac{\lambda}{\Delta} \\ 0 \end{pmatrix} \right),$$

Suppose that $\left( \gamma_0 \right)_n$ is an estimate of $\left( \beta_0 \right)_n$ such that

$$\sqrt{n} \left( \left( \gamma_0 \right)_n - \left( \beta_0 \right)_n \right) \sim N_{p+1}(0, \Sigma).$$

Then by Theorem 3,

$$n \left( ER \left[ \left( \gamma_0 \right)_n \right] - ER \left[ \left( \beta_0 \right)_n \right] \right) \sim \frac{1}{2 D_1} \left[ z_0^2 + \ldots + z_{p-1}^2 + \left( \frac{\lambda}{\Delta} \right)^2 z_p^2 - \frac{2\lambda}{\Delta} z_0 z_p \right],$$

where

$$z \sim N_{p+1}(0, \Sigma).$$

This agrees with the result of Efron [4] which was obtained by a different method.

**Example 2** Two Dimensional Exponential Distribution

Let

$$f_{1, 2}(x) = 2 \prod_{j=1}^{2} \theta_{1j} e^{-\theta_{1j} x}, i = 0, 1, j = 0, 1.$$

where

$$\theta_{20} - \theta_{21} > 0.$$
Then
\[
\beta_0 = \log \frac{\pi_1}{\pi_0} + \log \frac{\theta_{11} \theta_{21}}{\theta_{10} \theta_{20}}
\]
\[
\beta = \theta_0 - \theta_1.
\]

Then, letting \(\hat{\beta}_0\) denote the maximum likelihood estimate,
\[
\hat{\beta}_0 = \log \frac{n_1}{n_0} + \log \frac{x_{10}}{x_{11}} \frac{x_{20}}{x_{21}}
\]
\[
\hat{\beta} = \begin{pmatrix} x_{10-1} \\ x_{11-1} \\ x_{20-1} \\ x_{21-1} \end{pmatrix}
\]
\[
\begin{pmatrix} x_{10} \\ x_{11} \\ x_{20} \\ x_{21} \end{pmatrix}
\]

By standard theory,
\[
\sqrt{n} \left( \left( \begin{pmatrix} \beta_0 \\ \beta \end{pmatrix} - \left( \begin{pmatrix} \beta_0 \\ \beta \end{pmatrix} \right) \right) \right) \xrightarrow{d} N_3(0, \Sigma_M),
\]

where
\[
\Sigma_M =
\begin{pmatrix}
\frac{3}{\pi_1 \pi_0} & -\frac{\theta_{10}}{\pi_0} & -\frac{\theta_{11}}{\pi_1} & -\frac{\theta_{20}}{\pi_0} & -\frac{\theta_{21}}{\pi_1} \\
-\frac{\theta_{10}}{\pi_0} & -\frac{\theta_{11}}{\pi_1} & \frac{\theta_{11}^2}{\pi_1} & \frac{\theta_{10}^2}{\pi_0} & 0 \\
-\frac{\theta_{20}}{\pi_0} & \frac{\theta_{21}}{\pi_1} & 0 & \frac{\theta_{21}^2}{\pi_1} & \frac{\theta_{20}^2}{\pi_0}
\end{pmatrix}
\]

By Theorem 2,
\[
\sqrt{n} \left( \left( \begin{pmatrix} \beta_0 \\ \beta \end{pmatrix} - \left( \begin{pmatrix} \beta_0 \\ \beta \end{pmatrix} \right) \right) \right) \xrightarrow{d} N_3(0, \Sigma_L),
\]

20
where \( \Sigma_L \) is defined by (3.4). Figure 1 illustrates the cases which need to be considered in determining the matrix \( B \) of Theorem 3.

Figure 1. The Four Different Possibilities for the Slope and Intercept of the Optimum Boundary.

**Case 1**

\[
f_1(h, x_1) = \begin{cases} 0, & x_1 < 0 \\ ce^{-dx}, & x_1 \geq 0 \end{cases}
\]

where

\[
c = \pi_1 \theta_{11} \theta_{22} \beta_2 e
\]

21
and
\[ d = (\theta_{11} \theta_{20} - \theta_{21} \theta_{10})/\beta_2. \]

Thus
\[
B_1 = \frac{c}{2d\beta_2^2} \left( \begin{array}{ccc}
1 & \frac{1}{d} & -\frac{\beta_0}{d\beta_2} - \frac{\beta_1}{d^2\beta_2} \\
\frac{1}{d} & \frac{2}{d^2} & -\frac{\beta_0}{d\beta_2} - \frac{\beta_1}{d^2\beta_2} \\
-\frac{\beta_0}{2} - \frac{\beta_1}{d\beta_2} & -\frac{\beta_0}{d\beta_2} - \frac{\beta_1}{d^2\beta_2} & \frac{1}{(d\beta_2)^2} (2\beta_1^2 + 2d\beta_0 \beta_1 + d^2\beta_0^2) \\
\end{array} \right)
\]

Case 2

\[
f_1(h_{\beta_0}(x_1)) = \begin{cases} 
0, \ x < -\frac{\beta_0}{\beta_1} \\
-\frac{dx_1}{\beta_0}, \ x \geq -\frac{\beta_0}{\beta_1}
\end{cases}
\]

Thus
\[
B_2 = \frac{c}{2d\beta_2^2} \beta_1 \left( \begin{array}{ccc}
1 & \frac{\beta_0}{\beta_1} + \frac{1}{d} & -\frac{\beta_1}{d\beta_2} \\
\frac{\beta_0}{\beta_1} + \frac{1}{d} & \frac{\beta_0^2}{\beta_1^2} - \frac{2\beta_0}{d\beta_1} + \frac{2}{d^2} & \frac{\beta_0}{d\beta_2} - \frac{2\beta_1}{d^2\beta_2} \\
-\frac{\beta_1}{d\beta_2} & \frac{\beta_0}{d\beta_2} - \frac{2\beta_1}{d^2\beta_2} & \frac{2\beta_1}{d^2\beta_2} \\
\end{array} \right)
\]
Case 3

\[ f_1(h_{\beta_0}(x_1)) = \begin{cases} 
-\frac{d_1}{2}, & 0 < x_1 \leq -\frac{\beta_0}{\beta_1} \\
0, & \text{otherwise}
\end{cases} \]

If \( d = 0 \) then

\[ B_3 = \frac{c}{2\beta_2^2} \begin{pmatrix}
1 & -\frac{\beta_0}{2\beta_1} & \frac{\beta_0^2}{2\beta_1\beta_2} \\
-\frac{\beta_0}{2\beta_1} & -\frac{\beta_0^3}{3\beta_1^3} & -\frac{\beta_0^3}{6\beta_2\beta_1^2} \\
\frac{\beta_0^2}{2\beta_1\beta_2} & -\frac{\beta_0^3}{6\beta_2\beta_1^2} & -\frac{\beta_0^3}{3\beta_1^2}\end{pmatrix}.
\]

Otherwise

\[ B_3 = B_1 - B_2. \]

Case 4

\[ f_1(h_{\beta_0}(x_1)) = 0, \ x_1 > 0. \]

Thus

\[ B_4 = 0. \]

Let

\[ \Delta_1 = \frac{\theta_{11}}{\theta_{10}}, \]

\[ \Delta_2 = \frac{\theta_{21}}{\theta_{20}}. \]

The probability of misclassification is unchanged under scale transformation. Thus it can be assumed that
\[
\begin{align*}
\begin{array}{c}
\theta_{11} = \theta_{21} = 1, \\
\theta_{10} = \Delta_1, \\
\theta_{20} = \Delta_2 > 1.
\end{array}
\end{align*}
\]

Table 1 gives numerical values of the relative efficiency for varying \( \pi_1, \Delta_1 \) and \( \Delta_2 \). The logistic regression estimates compare well with the maximum likelihood estimates over the range considered.

**Table 1. Relative Efficiency of Logistic Regression to Maximum Likelihood Discrimination in Bivariate Exponentials.**

<table>
<thead>
<tr>
<th>( \Delta_1 )</th>
<th>( \pi_1 = .25 )</th>
<th>( \pi_1 = .5 )</th>
<th>( \pi_1 = .75 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>.25</td>
<td>.68</td>
<td>.65</td>
<td>.75</td>
</tr>
<tr>
<td>.33</td>
<td>.71</td>
<td>.68</td>
<td>.79</td>
</tr>
<tr>
<td>.5</td>
<td>.74</td>
<td>.70</td>
<td>.82</td>
</tr>
<tr>
<td>.67</td>
<td>.76</td>
<td>.72</td>
<td>.88</td>
</tr>
<tr>
<td>.75</td>
<td>.77</td>
<td>.72</td>
<td>.89</td>
</tr>
</tbody>
</table>
| 1.             | .79 | .73 | .90 | .85 | *   |   *
| 1.33           | .79 | .72 | .92 | .85 | *   |   *
| 1.5            | .77 | .71 | .92 | .85 | *   |   *
| 2.             | .72 | .67 | .89 | .84 | .83 | .78 |
| 3.             | .67 | .62 | .84 | .80 | .78 | .75 |
| 4.             | .63 | .58 | .79 | .76 | .75 | .75 |

* denotes that both procedures satisfy \( n \left( \operatorname{ER} \left( \frac{\gamma_0}{\gamma} \right) \right) \rightarrow 0 \) and so \( \text{Eff}_p \) is not defined.

24
ACKNOWLEDGMENT

I wish to thank Bradley Efron for suggesting the problem.

REFERENCES


