The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis

Author(s): Bradley Efron

Source: *Journal of the American Statistical Association* , Dec., 1975, Vol. 70, No. 352 (Dec., 1975), pp. 892-898

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: https://www.jstor.org/stable/2285453

The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis

BRADLEY EFRON*

A random vector $\mathbf{x}$ arises from one of two multivariate normal distributions differing in mean but not covariance. A training set $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ of previous cases, along with their correct assignments, is known. These can be used to estimate Fisher's discriminant by maximum likelihood and then to assign $\mathbf{x}$ on the basis of the estimated discriminant, a method known as the normal discrimination procedure. Logistic regression does the same thing but with the estimation of Fisher's disriminant done conditionally on the observed values of $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$. This article computes the asymptotic relative efficiency of the two procedures. Typically, logistic regression is shown to be between one half and two thirds as effective as normal discrimination for statistically interesting values of the parameters.

1. INTRODUCTION AND SUMMARY

Suppose that a random vector $\mathbf{x}$ can arise from one of two p-dimensional normal populations differing in mean but not in covariance,

$$\begin{aligned} \mathbf{x} &\sim \mathfrak{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \quad \text{with} \quad \text{prob}\, \pi_1 \;, \\ \mathbf{x} &\sim \mathfrak{N}_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) \quad \text{with} \quad \text{prob}\, \pi_0 \;, \end{aligned} \tag{1.1}$$

where $\pi_1 + \pi_0 = 1$.

If the parameters $\pi_1, \pi_0 = 1 - \pi_1, \boldsymbol{\mu}_1, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}$ are known, then $\mathbf{x}$ can be assigned to a population on the basis of Fisher's "linear discriminant function" [1].

$$\begin{aligned} \lambda(\mathbf{x}) &= \beta_0 + \boldsymbol{\beta}'\mathbf{x}, \\ \beta_0 &\equiv \log\frac{\pi_1}{\pi_0} - \frac{1}{2}(\boldsymbol{\mu}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0) \;, \\ \boldsymbol{\beta}' &\equiv (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1} \;. \end{aligned} \tag{1.2}$$

The assignment is to population 1 if $\lambda(x) > 0$ and to population 0 if $\lambda(x) < 0$. This method of assignment minimizes the expected probability of misclassification, as is easily shown by applying Bayes theorem. There is no loss of generality in assuming $\boldsymbol{\Sigma}$ nonsingular as we have done, since singular cases can always be made nonsingular by an appropriate reduction of dimension.

In usual practice, the parameters $\pi_1, \pi_0, \boldsymbol{\mu}_0, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}$ will be unknown to the statistician, but a training set $(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), \cdots, (y_n, \mathbf{x}_n)$ will be available, where y_j indicates which population $\mathbf{x}_j$ comes from, so

$$\begin{aligned} y_j = 1 \quad &\text{with} \quad \text{prob}\, \pi_1 \;, \\ 0 \quad &\text{with} \quad \text{prob}\, \pi_0 \;, \end{aligned} \tag{1.3}$$

and, of course,

$$\mathbf{x}_j | y_j \sim \mathfrak{N}_p(\boldsymbol{\mu}_{y_j}, \boldsymbol{\Sigma}) \;. \tag{1.4}$$

The $(y_j, \mathbf{x}_j)$ are assumed independent of each other for $j = 1, 2, \cdots, n$. In this case, maximum likelihood estimates of the parameters are available,

$$\begin{aligned} \hat{\pi}_1 = n_1/n \;, \quad \hat{\pi}_0 = n_0/n \;, \\ \hat{\boldsymbol{\mu}}_1 = \bar{\mathbf{x}}_1 \equiv \sum_{y_j=1} \mathbf{x}_j/n_1 \;, \quad \hat{\boldsymbol{\mu}}_0 = \bar{\mathbf{x}}_0 \equiv \sum_{y_j=0} \mathbf{x}_j/n_0 \;, \end{aligned} \tag{1.5}$$

and

$$\hat{\boldsymbol{\Sigma}} = [\sum_{y_j=1} (\mathbf{x}_j - \bar{\mathbf{x}}_1)(\mathbf{x}_j - \bar{\mathbf{x}}_1)' + \sum_{y_j=0} (\mathbf{x}_j - \bar{\mathbf{x}}_0)(\mathbf{x}_j - \bar{\mathbf{x}}_0)']/n \;,$$

where $n_1 \equiv \sum_{j=1}^{n} y_0$ and $n_0 \equiv n_1 - n_0$ are the number of population 1 and population 0 cases observed, respectively. Substituting these into (1.2) gives a version of Anderson's [1] estimated linear discriminant function, say, $\hat{\lambda}(\mathbf{x}) = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}'\mathbf{x}$, and an *estimated* discrimination procedure which assigns a new $\mathbf{x}$ to population 1 or 0 as $\hat{\lambda}(\mathbf{x})$ is greater than or less than zero. This will be referred to as the "normal discrimination procedure."

Bayes' theorem shows that $\lambda(\mathbf{x})$, as given in (1.2), is actually the *a posteriori* log odds ratio for Population 1 versus Population 0 having observed $\mathbf{x}$,

$$\lambda(\mathbf{x}_j) \equiv \log\frac{\pi_1(\mathbf{x}_j)}{\pi_0(\mathbf{x}_j)}, \quad \pi_i(\mathbf{x}_j) \equiv \text{prob}\,\{y_j = i | \mathbf{x}_j\} \;, \quad i = 1, 0 \;. \tag{1.6}$$

To simplify notation we will also write

$$\pi_{ij} \equiv \pi_i(\mathbf{x}_j) \quad \text{and} \quad \lambda \equiv \log\,(\pi_1/\pi_0) \;. \tag{1.7}$$

Given the values $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$, the y_j are conditionally independent binary random variables,

$$\begin{aligned} \text{prob}\,\{y_j = 1 | \mathbf{x}_j\} &= \pi_{1j} \\ &= \exp\,(\beta_0 + \boldsymbol{\beta}'\mathbf{x}_j)/[1 + \exp\,(\beta_0 + \boldsymbol{\beta}'\mathbf{x}_j)] \;, \end{aligned} \tag{1.8}$$

$$\text{prob}\,\{y_j = 0 | \mathbf{x}_j\} = \pi_{0j} = 1/[1 + \exp\,(\beta_0 + \boldsymbol{\beta}'\mathbf{x}_j)] \,.$$

To estimate $(\beta_0, \boldsymbol{\beta})$, we can maximize the conditional

* Bradley Efron is professor, Department of Statistics, Stanford University, Stanford, Calif. 94305. The author is grateful to Gus Haggstrom of RAND Corporation for helpful comments.

© Journal of the American Statistical Association
December 1975, Volume 70, Number 352
Theory and Methods Section

likelihood

$$
\begin{aligned}
f_{\beta_0,\boldsymbol{\beta}}(y_1, \cdots, y_n | \mathbf{x}_1, \cdots, \mathbf{x}_n) \\
&= \prod_{j=1}^{n} \pi_{1j}{}^{y_j} \pi_{0j}{}^{(1-y_j)} , \\
&= \prod_{j=1}^{n} \frac{\exp\,[(\beta_0 + \boldsymbol{\beta}'\mathbf{x}_j) y_j]}{[1 + \exp\,(\beta_0 + \boldsymbol{\beta}'\mathbf{x}_j)]} ,
\end{aligned} \tag{1.9}
$$

with respect to $(\beta_0, \boldsymbol{\beta})$. The maximizing values, call them $(\bar{\beta}_0, \bar{\boldsymbol{\beta}})$, give $\bar{\lambda}(\mathbf{x}) = \bar{\beta}_0 + \bar{\boldsymbol{\beta}}'\mathbf{x}$ as an estimate of the linear discriminant function. The discrimination procedure which chooses Population 1 or 0 as $\bar{\lambda}(\mathbf{x})$ is greater than or less than zero will be referred to as the "logistic regression procedure." An excellent discussion of such procedures is given in Cox's monograph [2].

The logistic regression procedure must be less efficient than the normal discrimination procedure under model (1.1), at least asymptotically, as n goes to infinity, since the latter is based on the full maximum likelihood estimator for $\lambda(\mathbf{x})$. This article calculates the asymptotic relative efficiencies (ARE) of the two procedures. The central result is that, under a variety of situations and measures of efficiency, the ARE is given by

$$\text{ARE} = \frac{1 + \Delta^2\pi_1\pi_0}{(2\pi)^{\frac{1}{2}}} e^{-\Delta^2/8} \int_{-\infty}^{\infty} \frac{e^{-x^2/2}}{\pi_1 e^{\Delta x/2} + \pi_0 e^{-\Delta x/2}} dx \ , \tag{1.10}$$

where

$$\Delta \equiv [(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)]^{\frac{1}{2}} \ , \tag{1.11}$$

the square root of the Mahalanobis distance. Following is a small tabulation of (1.10) for reasonable values of Δ, with $\pi_1 = \pi_0 = \frac{1}{2}$ (the case most favorable to the logistic regression procedure).

Δ	0	.5	1	1.5	2	2.5	3	3.5	4
ARE	1.000	1.000	.995	.968	.899	.786	.641	.486	.343

(1.12)

Why use logistic regression at all if it is less efficient (and also more difficult to calculate)? Because it is more robust, at least theoretically, than normal discrimination. The conditional likelihood (1.9) is valid under general exponential family assumptions on the density $f(\mathbf{x})$ of $\mathbf{x}$,

$$
\begin{aligned}
f(\mathbf{x}) &= g(\boldsymbol{\theta}_1, \boldsymbol{\eta}) h(\mathbf{x}, \boldsymbol{\eta}) \exp\,(\boldsymbol{\theta}_1'\mathbf{x}) \quad \text{with} \quad \text{prob}\ \pi_1 \ , \\
f(\mathbf{x}) &= g(\boldsymbol{\theta}_0, \boldsymbol{\eta}) h(\mathbf{x}, \boldsymbol{\theta}) \exp\,(\boldsymbol{\theta}_0'\mathbf{x}) \quad \text{with} \quad \text{prob}\ \pi_0 \ ,
\end{aligned} \tag{1.13}
$$

where $\pi_1 + \pi_0 = 1$.

Here, $\boldsymbol{\eta}$ is an arbitrary nuisance parameter, like $\boldsymbol{\Sigma}$ in (1.1). Equation (1.13) includes (1.1) as a special case.

Unfortunately, (1.12) shows that the statistician pays a moderately steep price for this added generality, assuming, of course, that (1.1) is actually correct. Just when good discrimination becomes possible, for Δ between 2.5 and 3.5, the ARE of the logistic procedure falls off sharply. The question of how to choose or compromise between the two procedures seems important, but no results are available at this time. Another important unanswered question is the relative efficiency under some model other than (1.1), when we are not playing ball on normal discrimination's home court.

In many situations, the sampling probabilities π_1, π_0 acting in (1.1) may be systematically distorted from their values in the population of interest. For example, if Population 1 is murder victims and Population 0 is all other people, a study conducted in a morgue would have π_1 much larger than in the whole population. Quite often n_1 and n_0 are set by the experimenter and are not random variables at all. These cases are discussed briefly in Section 5.

Technical details relating to asymptotic normality and consistency are omitted throughout the article. These gaps can be filled in by the application of standard exponential family theory, as presented, say, in [5], to (1.1). For another comparison of normal discrimination and logistic regression, the reader is referred to [4]. In that article, and also in [3], the distributions of $\mathbf{x}$ are allowed to have discrete components.

2. EXPECTED ERROR RATE

By means of a linear transformation $\tilde{\mathbf{x}} = \mathbf{a} + \mathbf{A}\mathbf{x}$, we can always reduce (1.1) to the case

$$
\begin{aligned}
\tilde{\mathbf{x}} &\sim \mathfrak{N}_p((\Delta/2)\mathbf{e}_1, \mathbf{I}) \ , \quad \text{with} \quad \text{prob}\ \pi_1 \ , \\
\tilde{\mathbf{x}} &\sim \mathfrak{N}_p(-(\Delta/2)\mathbf{e}_1, \mathbf{I}) \ , \quad \text{with} \quad \text{prob}\ \pi_0 \ ,
\end{aligned} \tag{2.1}
$$

where $\pi_1 + \pi_0 = 1$,

and $\mathbf{e}_1' \equiv (1, 0, 0, \cdots, 0)$; $\mathbf{I}$ is the $p \times p$ identity matrix; and $\Delta = ((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0))^{\frac{1}{2}}$ as before.

The boundary $B \equiv \{\mathbf{x} : \lambda(\mathbf{x}) = 0\}$ between Fisher's optimum decision regions for the two populations transforms to the new optimum boundary in the obvious way,

$$\tilde{B} \equiv \{\tilde{\mathbf{x}} : \tilde{\lambda}(\tilde{\mathbf{x}}) = 0\} = \{\tilde{\mathbf{x}} : \tilde{\mathbf{x}} = \mathbf{a} + \mathbf{A}\mathbf{x}, \mathbf{x} \in B\} \ . \tag{2.2}$$

Moreover, if $\mathbf{x}_1$, $\mathbf{x}_2$, $\cdots$, $\mathbf{x}_n$ is an iid sample from (1.1), and $\tilde{\mathbf{x}}_i = \mathbf{a} + \mathbf{A}\mathbf{x}_i$, $i = 1, 2, \cdots, n$, is the transformed sample, then both estimated boundaries $\hat{B} \equiv \{\mathbf{x} : \hat{\lambda}(\mathbf{x}) = 0\}$ and $\bar{B} \equiv \{\mathbf{x} : \bar{\lambda}(\mathbf{x}) = 0\}$ also transform as in (2.2). In words, then, for both logistic regression and normal discrimination, the estimated discrimination procedure based on the transformed data is the transform of that based on the original data. All of these statements are easy to verify.

Suppose we have the regions R_0 and R_1, a partition of the p-dimensional space E^p, and we decide for population 0 or population 1 as $\mathbf{x}$ falls into R_0 or R_1, respectively. The error rate of such a partition is the probability of misclassification under assumptions (1.1),

$$
\begin{aligned}
\text{Error Rate} \equiv\ & \pi_1\, \text{prob}\ \{\mathbf{x} \in R_0 | x \sim \mathfrak{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})\} \\
& + \pi_0\, \text{prob}\ \{\mathbf{x} \in R_1 | \mathbf{x} \sim \mathfrak{N}_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})\} \ .
\end{aligned} \tag{2.3}
$$

When the partition is chosen randomly, as it is by the logistic regression and normal discrimination procedures, error rate is a random variable. For either procedure, it follows from the preceding that error rate will have the same distribution under (1.1) and (2.1). Henceforth, we will work with the simpler assumptions (2.1), calling this the "standard situation" (with the basic random variable referred to as "$\mathbf{x}$" rather than "$\tilde{\mathbf{x}}$" for convenience).

For the standard situation, Fisher's linear discriminant function (1.2) becomes

$$\lambda(\mathbf{x}) = \lambda + \Delta x_1 \ . \qquad (2.4)$$

The boundary $\lambda(\mathbf{x}) = 0$ is the $(p-1)$-dimensional plane orthogonal to the x_1 axis and intersecting it at the value

$$\tau \equiv -\lambda/\Delta \ . \qquad (2.5)$$

In the figure, the optimal boundary is labeled $B(0, 0)$. The figure also shows another boundary, labeled $B(d\tau, d\alpha)$, intersecting the x_1 axis at $\tau + d\tau$, with normal vector at an angle $d\alpha$ from the x_1 axis. The differential notation $d\tau$ and $d\alpha$ indicates small discrepancies from optimal, which will be the case in the large sample theory. The error rate (2.3) of the regions separated by $B(d\tau, d\alpha)$ will be denoted by $\text{ER}(d\tau, d\alpha)$.

Letting

$$D_1 \equiv (\Delta/2) - \tau \ , \quad D_0 \equiv (\Delta/2) + \tau \ , \qquad (2.6)$$

we see that the error rate of the optimal boundary $B(0, 0)$ is

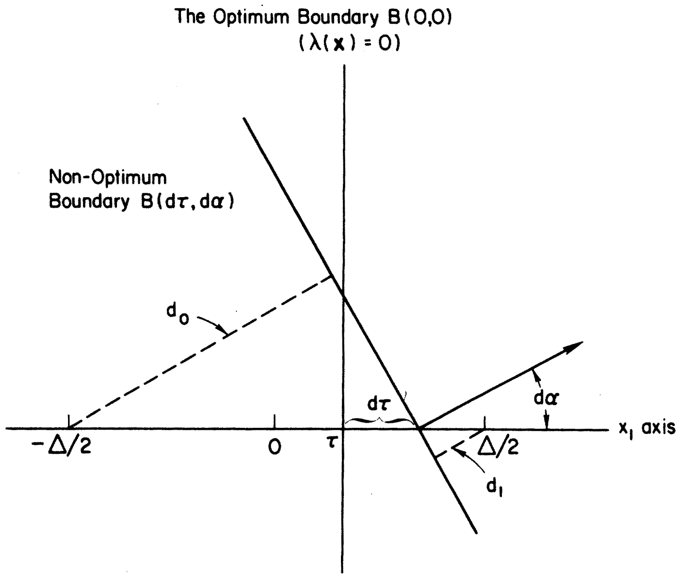
$$\text{ER}\ (0, 0) = \pi_1\Phi(-D_1) + \pi_0\Phi(-D_0) \ , \qquad (2.7)$$

where

$$\Phi(z) \equiv \int_{-\infty}^{z} \varphi(t)dt \quad \text{and} \quad \varphi(t) \equiv (2\pi)^{-\frac{1}{2}} \exp(-t^2/2)$$

as usual. (We are tacitly assuming that the two regions divided by $B(d\tau, d\alpha)$ are assigned to populations 1 and 0, respectively, in the best way.)

Optimum Boundary $\lambda(\mathbf{x}) = 0$ in Standard Situation[a]



[a] Also shown is some other boundary intersecting the x_1 axis at $\tau + d\tau$ and at angle $d\alpha$.

Now, define

$$d_1 \equiv (D_1 - d\tau)\cos(d\alpha) \ , \qquad d_0 \equiv (D_0 + d\tau)\cos(d\alpha) \ , \qquad (2.8)$$

the distances from $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_0$ to $B(d\tau, d\alpha)$. Then,

$$\text{ER}\ (d\tau, d\alpha) = \pi_1\Phi(-d_1) + \pi_0\Phi(-d_0) \ . \qquad (2.9)$$

From the Taylor expansions,

$$\cos(d\alpha) = 1 - (d\alpha)^2/2 + \cdots$$

and

$$\Phi(-D + d\tau) = \Phi(-D) + \varphi(D)d\tau + D\varphi(D)(d\tau)^2/2 + \cdots \ ,$$

we get the following lemma.

Lemma 1: Ignoring differential terms of third and higher orders,

$$\text{ER}\ (d\tau, d\alpha) = \text{ER}\ (0, 0) + (\Delta/2)\pi_1\varphi(D_1)[(d\tau)^2 + (d\alpha)^2] \ . \qquad (2.10)$$

Equation (2.10) makes use of the fact that, by Bayes theorem $\pi_1\varphi(D_1)/\pi_0\varphi(D_0) = 1$, or equivalently,

$$\pi_1\varphi(D_1) = \pi_0\varphi(D_0) \ . \qquad (2.11)$$

Suppose now that the boundary, $B(d\tau, d\alpha)$ is given by those $\mathbf{x}$ satisfying

$$(\lambda + d\beta_0) + (\Delta\mathbf{e}_1 + d\boldsymbol{\beta})'\mathbf{x} = 0 \ , \qquad (2.12)$$

$d\beta_0$ and $d\boldsymbol{\beta} = (d\beta_1, d\beta_2, \cdots, d\beta_p)'$, indicating small discrepancies from the optimal linear function (2.4). Again, ignoring higher-order terms, we have

$$d\tau = (1/\Delta)(-d\beta_0 + (\lambda/\Delta)d\beta_1) \ ,$$

and so

$$(d\tau)^2 = \frac{1}{\Delta^2}\left((d\beta_0)^2 - \frac{2\lambda}{\Delta}d\beta_0 d\beta_1 + \frac{\lambda^2}{\Delta^2}(d\beta_1)^2\right) . \qquad (2.13)$$

Similarly, expansion of

$$d\alpha = \arctan\left[((d\beta_2)^2 + \cdots + (d\beta_p)^2)^{\frac{1}{2}}/(\Delta + d\beta_1)\right]$$

gives

$$(d\alpha)^2 = ((d\beta_2)^2 + (d\beta_3)^2 + \cdots + (d\beta_p)^2)/\Delta^2 \ . \qquad (2.14)$$

Finally, suppose that under some method of estimation, the $(p+1)$ vector of errors $(d\beta_0, d\boldsymbol{\beta})$ has a limiting normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}/n$,

$$\mathcal{L}: \ \sqrt{n}\begin{pmatrix} d\beta_0 \\ \boldsymbol{\beta} \end{pmatrix} \rightarrow \mathfrak{N}_{p+1}(\mathbf{0}, \ \boldsymbol{\Sigma}) \ . \qquad (2.15)$$

The differential term which appears in Lemma 1,

$$(d\tau)^2 + (d\alpha)^2 = \frac{1}{\Delta^2}\left[(d\beta_0)^2 - \frac{2\lambda}{\Delta}d\beta_0 d\beta_1 + \frac{\lambda^2}{\Delta^2}(d\beta_1)^2 + (d\beta_2)^2 + \cdots + (d\beta_p)^2\right], \qquad (2.16)$$

will then have the limiting distribution of $1/n$ times the normal quadratic form

$$(1/\Delta^2)[z_0^2 - (2\lambda/\Delta)z_0z_1 + (\lambda/\Delta)^2z_1^2 + z_2^2 + \cdots + z_p^2] \ ,$$

where $\mathbf{z} \sim \mathfrak{N}_{p+1}(\mathbf{0}, \boldsymbol{\Sigma})$. Assuming moments converge correctly, which turns out to be the case for the logistic regression and normal discriminant procedures, Lemma 1 gives a simple expression for the expected error rate in terms of the elements σ_{ij} of $\boldsymbol{\Sigma}$.

Theorem 1: Ignoring terms of order less than $1/n$,

$$E\{\text{ER}(d\tau, d\alpha) - \text{ER}(0,0)\} = \frac{\pi_1\varphi(D_1)}{2\Delta n}\left[\sigma_{00} - \frac{2\lambda}{\Delta}\sigma_{01} + \frac{\lambda^2}{\Delta^2}\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp}\right]. \tag{2.17}$$

The quantity $E\{\text{ER}(d\tau, d\alpha) - \text{ER}(0, 0)\}$ is a measure of our expected regret, in terms of increased error rate, when using some estimated discrimination procedure. In Section 3, we evaluate $\boldsymbol{\Sigma}$ for the logistic regression procedure and the normal discriminant procedure and then use Theorem 1 to compare the two procedures.

3. ASYMPTOTIC ERROR RATES OF THE TWO PROCEDURES

First we consider the normal discriminant procedure described after (1.5).

Lemma 2: In the standard situation, the normal discriminant procedure produces estimates $(\hat{\beta}_0, \hat{\boldsymbol{\beta}}') \equiv (\lambda_0, \Delta\mathbf{e}_1') + (d\hat{\beta}_0, d\hat{\boldsymbol{\beta}}')$ satisfying

$$\mathfrak{L}: \quad \sqrt{n}\begin{pmatrix} d\hat{\beta}_0 \\ d\hat{\boldsymbol{\beta}} \end{pmatrix} \rightarrow \mathfrak{N}_{p+1}(\mathbf{0}, \boldsymbol{\Sigma}) \ , \tag{3.1}$$

where

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{\pi_1\pi_0}\begin{bmatrix} 1 + \Delta^2/4 & \frac{-\Delta}{2}(\pi_0 - \pi_1) & 0 \cdots 0 \cdots 0 \\ \frac{-\Delta}{2}(\pi_0 - \pi_1) & 1 + 2\Delta^2\pi_1\pi_0 & 0 \cdots 0 \cdots 0 \\ 0 & 0 & 1 + \Delta^2\pi_1\pi_0 \ 0 \cdots 0 \\ \vdots & \vdots & \ddots \\ 0 \quad \cdots & 0 \quad \cdots & 0 \cdots 0 \ 1 + \Delta^2\pi_1\pi_0 \end{bmatrix}. \tag{3.2}$$

Proof: The density of a single $(y, \mathbf{x})$ pair under (1.3)–(1.4) is

$$f_{\lambda,\mu_1,\mu_0,\Sigma}(y, \mathbf{x}) = \pi_y|\boldsymbol{\Sigma}|^{-\frac{1}{2}} \times \exp\left[-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_y)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_y)\right]\cdot(2\pi)^{-p/2} \ , \tag{3.3}$$

$\lambda \equiv \log \pi_1/\pi_0$, as before.

Let us write the distinct elements of $\boldsymbol{\Sigma}^{-1}$ as a $p(p + 1)/2$ vector $(\sigma^{11}, \sigma^{12}, \cdots, \sigma^{1p}, \sigma^{22}, \sigma^{23}, \cdots, \sigma^{pp})$ and indicate this vector as $(\boldsymbol{\sigma}^{(1)}, \boldsymbol{\sigma}^{(2)})$, where

$$\boldsymbol{\sigma}^{(1)} \equiv (\sigma^{11}, \sigma^{12}, \cdots, \sigma^{1p}) \ , \quad \boldsymbol{\sigma}^{(2)} \equiv (\sigma^{22}, \sigma^{23}, \cdots, \sigma^{pp}) \ . \tag{3.4}$$

Standard results using Fisher's information matrix then give the following asymptotic distributions for the maximum likelihood estimates in the standard situation.

$$\begin{aligned} &\mathfrak{L}: \sqrt{n}(\hat{\lambda} - \lambda) \rightarrow \mathfrak{N}(0, (1/\pi_0\pi_1)) \ , \\ &\mathfrak{L}: \sqrt{n}(\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i) \rightarrow \mathfrak{N}_p(\mathbf{0}, (1/\pi_i)\mathbf{I}), \ i = 1, 0 \ , \\ &\mathfrak{L}: \sqrt{n}(\hat{\boldsymbol{\sigma}}^{(1)} - \boldsymbol{\sigma}^{(1)}) \rightarrow \mathfrak{N}_p(\mathbf{0}, \mathbf{I} + \mathbf{E}_{11}) \ . \end{aligned} \tag{3.5}$$

Moreover, $\hat{\lambda}$, $\hat{\boldsymbol{\mu}}_1$, $\hat{\boldsymbol{\mu}}_0$, $\hat{\boldsymbol{\sigma}}^{(1)}$ and $\hat{\boldsymbol{\sigma}}^{(2)}$ are asymptotically uncorrelated. (We do not need the limiting distribution of $\hat{\boldsymbol{\sigma}}^{(2)}$ for the proof of Lemma 2.) Here, $\lambda = \log \pi_1/\pi_0$, and $\mathbf{E}_{11}$ is the $p \times p$ matrix having upper left element one and all others zero.

Differentiating (1.2) gives

$$\begin{gathered} \frac{\partial\beta_0}{\partial\lambda} = 1 \ , \quad \frac{\partial\beta_0}{\partial\boldsymbol{\mu}_1'} = -\boldsymbol{\mu}_1'\boldsymbol{\Sigma}^{-1} \ , \\ \frac{\partial\beta_0}{\partial\boldsymbol{\mu}_0'} = \boldsymbol{\mu}_0'\boldsymbol{\Sigma}^{-1} \ , \quad \frac{\partial\beta_0}{\partial\sigma^{ij}} = \frac{\mu_{0i}\mu_{0j} - \mu_{1i}\mu_{1j}}{1 + \delta_{ij}} \ , \\ \frac{\partial\boldsymbol{\beta}}{\partial\lambda} = 0 \ , \quad \frac{\partial\boldsymbol{\beta}}{\partial\boldsymbol{\mu}_1'} = \boldsymbol{\Sigma}^{-1} \ , \\ \frac{\partial\boldsymbol{\beta}}{\partial\boldsymbol{\mu}_0'} = -\boldsymbol{\Sigma}^{-1} \ , \quad \frac{\partial\boldsymbol{\beta}}{\partial\sigma^{ij}} = \frac{\mathbf{E}_{ij} + \mathbf{E}_{ji}}{1 + \delta_{ij}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \ , \end{gathered} \tag{3.6}$$

μ_{0i} indicating the ith component of $\boldsymbol{\mu}_0$; likewise for $\boldsymbol{\mu}_1$, with derivatives involving vectors taken componentwise in the obvious way. Moreover, $\delta_{ij} = 1$ or 0 as $i = j$ or $i \neq j$, and $\mathbf{E}_{ij}$ is the matrix with one in the ijth position and zero elsewhere. In the standard situation, we have the differential relationship

$$\begin{pmatrix} d\beta_0 \\ d\boldsymbol{\beta} \end{pmatrix} = \begin{bmatrix} 1 & \frac{-\Delta}{2}\mathbf{e}_1' & \frac{-\Delta}{2}\mathbf{e}_1' & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & -\mathbf{I} & \Delta\mathbf{I} & \mathbf{0} \end{bmatrix} \begin{bmatrix} d\lambda \\ d\boldsymbol{\mu}_1 \\ d\boldsymbol{\mu}_0 \\ d\boldsymbol{\sigma}^{(1)} \\ d\boldsymbol{\sigma}^{(2)} \end{bmatrix} . \tag{3.7}$$

Letting $\mathbf{M}$ be the matrix on the right side of (3.7),

$$\mathfrak{L}: \quad \sqrt{n}\begin{pmatrix} d\hat{\beta}_0 \\ d\hat{\boldsymbol{\beta}} \end{pmatrix} \rightarrow \mathfrak{N}_{p+1}(\mathbf{0}, \mathbf{M}[n\boldsymbol{\Sigma}_{\hat{\lambda},\hat{\mu}_1,\hat{\mu}_0,\hat{\sigma}^{(1)},\hat{\sigma}^{(2)}}]\mathbf{M}') \ , \tag{3.8}$$

where $n\boldsymbol{\Sigma}_{\hat{\lambda},\hat{\mu}_1,\hat{\mu}_0,\hat{\sigma}^{(1)}\hat{\sigma}^{(2)}}$ is the joint limiting covariance matrix of $(\hat{\lambda}, \hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\sigma}}^{(1)}, \hat{\boldsymbol{\sigma}}^{(2)})$, as indicated by (3.5). Evaluation of (3.8) gives the result.

Next consider the logistic regression estimates defined at (1.9).

Lemma 3: In the standard situation, the logistic regression procedure produces estimates $(\bar{\beta}_0, \bar{\boldsymbol{\beta}}') \equiv (\lambda_0, \Delta\mathbf{e}_1') + (d\bar{\beta}_0, d\bar{\boldsymbol{\beta}}')$ satisfying

$$\mathfrak{L}: \quad \sqrt{n}\begin{pmatrix} d\bar{\beta}_0 \\ \boldsymbol{\beta} \end{pmatrix} \rightarrow \mathfrak{N}_{p+1}(\mathbf{0}, \bar{\boldsymbol{\Sigma}}) \ , \tag{3.9}$$

where

$$\bar{\boldsymbol{\Sigma}} = \frac{1}{\pi_1\pi_0}\begin{bmatrix} \frac{A_2}{A_0A_2 - A_1^2} & \frac{-A_1}{A_0A_2 - A_1^2} & 0 \cdots 0 \\ \frac{-A_1}{A_0A_2 - A_1^2} & \frac{A_0}{A_0A_2 = A_1^2} & 0 \cdots 0 \\ 0 & 0 & \frac{1}{A_0} \quad 0 \\ \vdots & \vdots & \ddots \\ 0 \quad \cdots & 0 & \cdots 0 \cdots \frac{1}{A_0} \end{bmatrix} , \tag{3.10}$$

$A_i \equiv A_i(\pi_1, \Delta)$ being defined by

$$A_i(\pi_1, \Delta) \equiv \int_{-\infty}^{\infty} \frac{e^{-\Delta^2/8}x^i\varphi(x)}{\pi_1 e^{\Delta x/2} + \pi_0 e^{-\Delta x/2}}\,dx\ , \qquad i = 0, 1, 2\ . \quad (3.11)$$

Proof: The density (1.9) can be written in exponential family form as

$$\begin{aligned} f_{\beta_0,\boldsymbol{\beta}}(y_1, y_2, \cdots, y_n|\mathbf{x}_1, \cdots, \mathbf{x}_n) &= \exp\left[(\beta_0, \boldsymbol{\beta}')\mathbf{T} - \psi(\beta_0, \boldsymbol{\beta})\right]\ , \\ \mathbf{T} &\equiv \sum_{j=1}^{n} \binom{1}{\mathbf{x}_j} y_j\ , \qquad (3.12) \\ \psi(\beta_0, \boldsymbol{\beta}) &= \sum_{j=1}^{n} \log\left(1 + \exp\left(\beta_0 + \boldsymbol{\beta}'\mathbf{x}_j\right)\right)\ . \end{aligned}$$

The sufficient statistic $\mathbf{T}$ has mean vector and covariance matrix

$$\begin{aligned} E_{\beta_0,\boldsymbol{\beta}}\mathbf{T} = \sum_{j=1}^{n} \pi_{1j}\binom{1}{\mathbf{x}_j} \quad & \mathrm{Cov}_{\beta_0,\boldsymbol{\beta}}\mathbf{T} \\ &= \sum_{j=1}^{n} \pi_{1j}\pi_{0j}\binom{1}{\mathbf{x}_j}(1, \mathbf{x}_j')\ . \quad (3.13) \end{aligned}$$

Let $F^{(n)}$ denote the sample cdf of $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$, and suppose that $\mathcal{L}: F^{(n)} \to F$ as $n \to \infty$. Then,

$$\lim_{n\to\infty} \frac{1}{n}\,\mathrm{Cov}_{\beta_0,\boldsymbol{\beta}}\mathbf{T} = \int_{E^p} \binom{1}{\mathbf{x}}(1, \mathbf{x}')\pi_1(\mathbf{x})\pi_0(\mathbf{x})dF(\mathbf{x})\ , \quad (3.14)$$

where $\pi_0(\mathbf{x}) = 1 - \pi_1(\mathbf{x}) = [1 + \exp - (\beta_0 + \boldsymbol{\beta}'\mathbf{x})]^{-1}$. Exponential family theory says that the mapping from the expectation vector $E_{\beta_0,\boldsymbol{\beta}}\mathbf{T}$ to the natural parameters β_0, $\boldsymbol{\beta}$ has Jacobian matrix $[\mathrm{Cov}_{\beta_0,\boldsymbol{\beta}}\mathbf{T}]^{-1}$. Therefore, the "delta method" gives

$$\lim_{n\to\infty} n\,\mathrm{Cov}_{\beta_0,\boldsymbol{\beta}}\binom{\bar{\beta}_0}{\bar{\boldsymbol{\beta}}} = \left[\int_{E^p} \binom{1}{\mathbf{x}}(1, \mathbf{x}')\pi_1(\mathbf{x})\pi_0(\mathbf{x})dF(\mathbf{x})\right]^{-1}\ . \quad (3.15)$$

Under the sampling scheme (2.1), F will be the mixture of the normal populations $\mathfrak{N}_p((\Delta/2)\mathbf{e}_1, \mathbf{I})$ and $\mathfrak{N}_p((-\Delta/2)\mathbf{e}_1, \mathbf{I})$ in proportions π_1, π_0. In the standard situation, $\pi_0(\mathbf{x}) = 1 - \pi_1(\mathbf{x}) = [1 + \exp - (\lambda + \Delta x_1)]^{-1}$. We get

$$\lim_{n\to\infty} \frac{1}{n}\,\mathrm{Cov}_{\beta_0,\boldsymbol{\beta}}\mathbf{T} = \pi_1\pi_0 \begin{bmatrix} A_0 & A_1 & 0 \cdots 0 \\ A_1 & A_2 & 0 \cdots 0 \\ 0 & 0 & A_0 \quad 0 \\ \vdots & & \ddots \\ 0 \cdots & 0 \cdots & 0 \quad A_0 \end{bmatrix} \quad (3.16)$$

from (3.14). The covariance matrix (3.10) for $(d\bar{\beta}_0, d\bar{\boldsymbol{\beta}}')$ follows from (3.15). The fact that $(\bar{\beta}_0, \bar{\boldsymbol{\beta}}_i')$ is consistent for $(\beta_0, \boldsymbol{\beta}')$ and asymptotically normal, which is the remainder of Lemma 3, is not difficult to show, given the structure (2.1). Like most of the other regularity properties, it will not be demonstrated here.

We can now compute the relative efficiency of logistic regression to normal discrimination by Theorem 1. Denote the errors for the two procedures by $(d\bar{\tau}, d\bar{\alpha})$ and $(d\hat{\tau}, d\hat{\alpha})$, respectively, and define the efficiency measure,

$$\mathrm{Eff}_p\,(\lambda, \Delta) \equiv \lim_{n\to\infty} \frac{E\{\textsc{er}\ (d\hat{\tau}, d\hat{\alpha}) - \textsc{er}\ (0, 0)\}}{E\{\textsc{er}\ (d\bar{\tau}, d\bar{\alpha}) - \textsc{er}\ (0, 0)\}}\cdot \quad (3.17)$$

Theorem 1, and Lemmas 2 and 3 then give

$$\mathrm{Eff}_p = (Q_1 + (p - 1)Q_2)/(Q_3 + (p - 1)Q_4)\ , \quad (3.18)$$

where

$$\begin{aligned} Q_1 &\equiv \left(1, \frac{\lambda}{\Delta}\right)\begin{bmatrix} 1 + \dfrac{\Delta^2}{4} & (\pi_0 - \pi_1)\dfrac{\Delta}{2} \\ (\pi_0 - \pi_1)\dfrac{\Delta}{2} & 1 + 2\pi_0\pi_1\Delta^2 \end{bmatrix}\begin{bmatrix} 1 \\ \dfrac{\lambda}{\Delta} \end{bmatrix}, \\ Q_2 &\equiv 1 + \pi_1\pi_0\Delta^2\ , \\ Q_3 &\equiv \left(1, \frac{\lambda}{\Delta}\right)\frac{1}{A_0A_2 - A_1^2}\begin{pmatrix} A_2 & A_1 \\ A_1 & A_0 \end{pmatrix}\begin{bmatrix} 1 \\ \dfrac{\lambda}{\Delta} \end{bmatrix}, \\ Q_4 &\equiv \frac{1}{A_0}\cdot \end{aligned} \quad (3.19)$$

Rewriting (3.18) gives a simple expression for $\mathrm{Eff}_p\,(\lambda, \Delta)$ as a weighted average of the relative efficiencies when $p = 1$ and $p \to \infty$.

Theorem 2: The relative efficiency of logistic regression to normal discrimination is

$$\mathrm{Eff}_p\,(\lambda, \Delta) = \frac{q(\lambda, \Delta)\,\mathrm{Eff}_1\,(\lambda, \Delta) + (p - 1)\,\mathrm{Eff}_\infty\,(\lambda, \Delta)}{q(\lambda, \Delta) + (p - 1)}, \quad (3.20)$$

where

$$\mathrm{Eff}_1\,(\lambda, \Delta) \equiv Q_1/Q_3 \quad \text{and} \quad \mathrm{Eff}_\infty\,(\lambda, \Delta) \equiv Q_2/Q_4\ , \quad (3.21)$$

as defined in (3.19), are the relative efficiencies when $p = 1$ and $p = \infty$, respectively, and

$$q(\lambda, \Delta) \equiv Q_3/Q_4\ . \quad (3.22)$$

It is obvious from (3.18) that $\mathrm{Eff}_\infty\,(\lambda, \Delta) = Q_2/Q_4$ really is the asymptotic efficiency as $p \to \infty$. For $p = 1$, (3.18) gives $\mathrm{Eff}_1\,(\lambda, \Delta) = Q_1/Q_3$. This follows from Lemma 1 because $d\alpha$ can always be taken equal to zero when $p = 1$.

The case $\lambda = 0$ gives a particularly simple answer (since then $A_1 = 0$).

Corollary: When $\lambda = 0$, i.e., when $\pi_1 = \pi_0 = \frac{1}{2}$,

$$\mathrm{Eff}_p\,(\lambda, \Delta) = \mathrm{Eff}_\infty\,(\lambda, \Delta) = A_0(1 + \Delta^2/4)\ , \quad (3.23)$$

for all values of p.

Table 1 gives numerical values for the quantities involved in Theorem 2. It is worth noting that usually $\mathrm{Eff}_1\,(\lambda, \Delta) > \mathrm{Eff}_\infty\,(\lambda, \Delta)$ when $\lambda \neq 0$. However, q is near unity, so (3.20) shows that $\mathrm{Eff}_p\,(\lambda, \Delta)$ will be nearer $\mathrm{Eff}_\infty\,(\lambda, \Delta)$ than $\mathrm{Eff}_1\,(\lambda, \Delta)$, for $p \geq 3$.

Relative Efficiencies of Logistic Regression to Normal Discrimination[a]

π_1 (or π_0)	Δ	Eff_∞	Eff_1	q	A_0	A_1	A_2
.5	2	.899	.899	1	.450	0	.266
.6	2	.892	.906	1.024	.458	−.038	.273
.667	2	.879	.913	1.070	.465	−.067	.287
.75	2	.855	.915	1.177	.488	−.108	.319
.9	2	.801	.804	1.697	.589	−.253	.487
.95	2	.801	.706	2.233	.674	−.375	.667
.5	2.5	.786	.786	1	.307	0	.154
.6	2.5	.778	.794	1.013	.311	−.025	.158
.667	2.5	.762	.806	1.038	.319	−.044	.167
.75	2.5	.733	.819	1.096	.337	−.074	.188
.9	2.5	.660	.750	1.379	.423	−.181	.304
.95	2.5	.650	.637	1.671	.501	−.282	.441
.5	3	.641	.641	1	.197	0	.084
.6	3	.633	.649	1.008	.200	−.016	.087
.667	3	.618	.662	1.023	.206	−.027	.092
.75	3	.589	.682	1.057	.219	−.046	.104
.9	3	.511	.667	1.225	.282	−.117	.175
.95	3	.492	.588	1.400	.344	−.189	.265
.5	3.5	.486	.486	1	.120	0	.044
.6	3.5	.479	.493	1.005	.122	−.009	.045
.667	3.5	.467	.505	1.014	.125	−.016	.048
.75	3.5	.442	.526	1.035	.134	−.027	.055
.9	3.5	.370	.550	1.142	.176	−.070	.095
.95	3.5	.348	.516	1.252	.220	−.116	.147
.5	4	.343	.343	1	.069	0	.022
.6	4	.338	.348	1.003	.070	−.005	.022
.667	4	.328	.358	1.009	.072	−.009	.024
.75	4	.309	.375	1.024	.077	−.014	.027
.9	4	.252	.416	1.094	.103	−.039	.048
.95	4	.230	.416	1.168	.131	−.065	.076

[a] See (3.17), (3.19), (3.20), (3.21) for definition of terms.

4. ANGLE AND INTERCEPT ERROR

The terms "Eff$_\infty$ (λ, Δ)" and "Eff$_1$ (λ, Δ)" which appear in (3.20), Theorem 2, have another interpretation. Eff$_\infty$ (λ, Δ) is the asymptotic relative efficiency of logistic regression to normal discrimination for estimating the angle of the discriminant boundary,

$$\mathrm{Eff}_\infty(\lambda, \Delta) = \lim_{n\to\infty} \frac{\mathrm{Var}\,(d\hat{\alpha})}{\mathrm{Var}\,(d\bar{\alpha})}\cdot \tag{4.1}$$

(See the figure and the definitions preceding (3.17).) Likewise, Eff$_1$ (λ, Δ) is the asymptotic relative efficiency for estimating the intercept of the discriminant boundary,

$$\mathrm{Eff}_1(\lambda, \Delta) = \lim_{n\to\infty} \frac{\mathrm{Var}\,(d\hat{\tau})}{\mathrm{Var}\,(d\bar{\tau})}\cdot \tag{4.2}$$

These results follow immediately from (2.13), (2.14), (3.2), (3.10) and (3.21).

Comparing (2.14) with Lemmas 2 and 3 shows that

$$\begin{aligned} &\mathcal{L}:\ n\cdot(d\hat{\alpha})^2 \to \frac{1}{\pi_1\pi_0\Delta^2}(1+\Delta^2\pi_1\pi_0)\chi^2{}_{p-1}\,, \\ &\mathcal{L}:\ n\cdot(d\bar{\alpha})^2 \to \frac{1}{\pi_1\pi_0\Delta^2}\,\frac{1}{A_0}\,\chi^2{}_{p-1}\,. \end{aligned} \tag{4.3}$$

The asymptotic relative efficiency of logistic regression to normal discrimination in terms of angular error is thus

$$\begin{aligned} \textsc{are} &= (1+\Delta^2\pi_1\pi_0)A_0 \\ &= \frac{1+\Delta^2\pi_1\pi_0}{(2\pi)^{\frac{1}{2}}}\, e^{-\Delta^2/8}\int_{-\infty}^{\infty} \frac{e^{-x^2/2}}{\pi_1 e^{\Delta x/2}+\pi_0 e^{-\Delta x/2}}\,dx\ , \end{aligned} \tag{4.4}$$

in the strong sense that a sample of size $\bar{n}$ using logistic regression produces asymptotically the same angular error distribution as a sample of size $\hat{n} = \textsc{are}\cdot\bar{n}$, using normal discrimination. From (1.12), we see that if $\lambda = 0$, $\Delta = 2.5$, for example, $\bar{n} = 1{,}000$ is approximately equivalent to $\hat{n} = 786$. ("Eff$_\infty$" in the table is also "ARE" as given by (4.4).)

The corresponding statement for intercept error is *not* true because the two matrices involved in the definition of Q_1 and Q_3, (3.19), are not proportional. We have to settle for the weaker second-moment efficiency statement (4.2). However, when $\lambda = 0$, i.e., when $\pi_1 = \pi_0 = \frac{1}{2}$, (2.13) and Lemmas 2 and 3 show that

$$\begin{aligned} &\mathcal{L}:\ n\cdot(d\hat{\tau})^2 \to (4/\Delta^2)(1+\Delta^2/4)\chi^2{}_1\,, \\ &\mathcal{L}:\ n\cdot(d\bar{\tau})^2 \to (4/\Delta^2)(1/A_0)\chi^2{}_1\,. \end{aligned} \tag{4.5}$$

In this case, (4.4) with $\pi_1 = \pi_0 = \frac{1}{2}$ again gives the ARE in the strong sense of asymptotically equivalent sample sizes.

Combining (4.3) and (4.5) with Lemma 1 shows that when $\lambda = 0$ (and so $D_1 = \Delta/2$),

$$\begin{aligned} &\mathcal{L}:\ n\{\textsc{er}\,(d\hat{\tau}, d\hat{\alpha}) - \textsc{er}\,(0,0)\} \to \\ &\qquad\qquad (\varphi(\Delta/2)/\Delta)(1+\Delta^2/4)\chi^2{}_p \\ &\mathcal{L}:\ n\{\textsc{er}\,(d\bar{\tau}, d\bar{\alpha}) - \textsc{er}\,(0,0)\} \to (\varphi(\Delta/2)/\Delta)(1/A_0)\chi^2{}_p\,. \end{aligned} \tag{4.6}$$

Thus, error rates for samples of size $\bar{n}$ and $\hat{n} = \textsc{are}\cdot\bar{n}$ will have asymptotically equivalent distributions, with ARE given by (4.4), $\pi_1 = \pi_0 = \frac{1}{2}$. This is not true for π_1, $\pi_0 \neq \frac{1}{2}$, but as the dimension p gets large, it is. That is, error rates for the two procedures will have the same asymptotic distribution if $\hat{n} = \textsc{are}\cdot\bar{n}$, ARE given by (4.4), when $p \to \infty$ and $\bar{n}/p \to \infty$. A simple proof of this follows from (2.16) and Lemmas 2 and 3.

The angular error, $d\alpha$, unlike the error rate, is not invariant under linear transformations. Formulas (4.1), (4.3), and (4.4) refer to a "standardized angular error" defined *after* we have made the linear transformations, which take the general model (1.1) into the standard situation (2.1). However, it is easy to show that (4.1) and (4.4) (but not (4.3)) also hold for the true, unstandardized, angular error. This true error will be some quadratic form in the standardized coordinates $d\beta_2$, $d\beta_3$, $\cdots$, $d\beta_p$, not depending on which procedure is used. The result follows, because for both procedures, $(d\beta_2, \cdots, d\beta_p)$ has a limiting normal distribution with covariance matrix proportional to the identity. (Actually (4.3) holds with "$\chi^2{}_{p-1}$" replaced by a certain weighted sum of independent $\chi^2{}_1$ variates.)

There are two good reasons to be interested in angular error. First, under the fixed sampling proportion setup of Section 5, it is the only error of interest. Second, there is the well-known fact that minimizing $\sum_{i=1}^{n}[y_i - (a + \mathbf{b}'\mathbf{x}_i)]^2$ over all choices of the constant a and vector $\mathbf{b}$ gives $\mathbf{b}$ equal to $\hat{\boldsymbol\beta}$. (But a does not equal $\hat{\beta}_0$.) This connects normal discrimination with ordinary least squares analysis and provides some justification, or at least rationale, for using $\hat{\boldsymbol\beta}$ outside the framework (2.1).

Other efficiency comparisons between the two procedures, e.g., in estimating the slope $\|\boldsymbol\beta\|$ of the discriminant function, can be obtained from Lemmas 2 and 3.

5. DISTORTED SAMPLING PROPORTIONS

It may happen that the true probabilities $\tilde{\pi}_1$ and $\tilde{\pi}_0$ for populations 1 and 0 are distorted in a known way to different values π_1 and π_0 by the nature of the sampling scheme employed. Letting $\lambda \equiv \log \pi_1/\pi_0$, $\tilde{\lambda} \equiv \log \tilde{\pi}_1/\tilde{\pi}_0$, suppose that for some known constant c,

$$\lambda = \tilde{\lambda} + c \ . \tag{5.1}$$

For example, experimental constraints might cause the statistician to randomly exclude from his training set nine out of ten population 0 members, in which case $c = \log 10$. The normal discrimination procedure described at (1.5) is then modified in the obvious way. A new $\mathbf{x}$ is assigned to Population 1 or 0 as $\hat{\lambda}(\mathbf{x})$ is greater or less than c. The logistic regression procedure (1.9) is similarly modified.

Theorem 2 remains true as stated except for the following modification. The vector $(1, \lambda/\Delta)$ (and its transpose), which appears in the definitions of Q_1 and Q_3 in (3.19), is replaced by $(1, \tilde{\lambda}/\Delta)$. The constants $A_i \equiv A_i(\lambda, \Delta)$, which appear in Q_3, are *not* changed to $A_i(\tilde{\lambda}, \Delta)$. The proof of this is almost exactly the same as the proof of Theorem 2.

$\text{Eff}_\infty(\lambda, \Delta)$, the angular efficiency, remains unchanged, which is not surprising, since the discrimination boundary for any choice of c is parallel to that for $c = 0$. Only the intercept is changed. When $\pi_1 = \pi_0 = .5$, the effect of choosing $c \neq 0$ is to reduce $\text{Eff}_1(\lambda, \Delta)$, the intercept efficiency of logistic regression compared to normal discrimination, as shown in the following tabulation.

	$\Delta = 2,\ \pi_1 = .5$				$\Delta = 3,\ \pi_1 = .5$				
c	0	±1	±2	±3	0	±1	±2	±3	
Eff_1	.899	.869	.836	.819	.641	.604	.550	.516	(5.2)

Eff_1 for other values of c, π_1, Δ can be obtained using the entries A_0, A_1, A_2 in the table.

Most frequently, the sample sizes n_1 and n_0 are set by the statistician and are not random variables at all. The usual procedure in this situation is to estimate only the angle, not the intercept, of the discrimination boundary. In terms of the figure, the statistician uses the data to select a family of parallel boundaries $B(\cdot, d\alpha)$. The value of the intercept $d\tau$ is chosen on *a priori* grounds by just guessing what λ is, or may not be formally selected at all.

Either normal discrimination or logistic regression may be used to estimate the vector $\boldsymbol\beta$ in (1.2). It can be shown that $d\hat{\boldsymbol\beta}$ and $d\tilde{\boldsymbol\beta}$ still have the limiting distributions indicated in Lemmas 2 and 3, with π_1 and π_0 replaced by $r_1 \equiv n_1/n$ and $r_0 \equiv n_0/n$. In terms of angular error, the ARE (4.4) still gives the asymptotic relative efficiency of logistic regression to normal discrimination in the strong sense of Section 4. The quantities π_1, π_0 in (4.4) are replaced by $r_1 = n_1/n$, $r_0 = n_0/n$, where these proportions are assumed to exist and do not equal zero in the limit.

The estimates $\hat{\boldsymbol\mu}_1$, $\hat{\boldsymbol\mu}_0$, $\hat{\boldsymbol\Sigma}$, given in (1.5), are maximum likelihood, whether n_1, n_0 are fixed or random. It follows that $\hat{\boldsymbol\beta}' = (\hat{\boldsymbol\mu}_1 - \hat{\boldsymbol\mu}_0)'\hat{\boldsymbol\Sigma}^{-1}$, which we can still call the normal discrimination estimate, is maximum likelihood in either case. Standard maximum likelihood arguments, similar to the proof of Lemma 2, show that $d\hat{\boldsymbol\beta}$ is distributed as stated in Lemma 2, with π_1, π_0 replaced by r_1, r_0.

Let $T_1 \equiv \sum_{j=1}^{n} y_j$ be the first coordinate of $\mathbf{T}$ in (3.12), and let $\mathbf{T}_2$ be the remaining p coordinates. Given that $T_1 = n_1$, the conditional density of $y_1, y_2, \cdots, y_n$ is an exponential family with natural parameter $\boldsymbol\beta$ and sufficient statistic $\mathbf{T}_2$,

$$f_{\boldsymbol\beta}(y_1, y_2, \cdots, y_n \mid T_1 = n_1, \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n) = \exp[\boldsymbol\beta'\mathbf{T}_2 - \psi_{n_1}(\boldsymbol\beta)] \ , \tag{5.3}$$

where $\psi_{n_1}(\boldsymbol\beta)$ is chosen to make (5.3) sum to unity over all choices of $y_1, \cdots, y_n$ with $\sum_{j=1}^{n} y_j = n_1$. The analog of the logistic regression procedure is to select $\tilde{\boldsymbol\beta}$ to maximize the likelihood (5.3). A modification of the proof of Lemma 3, which will not be presented, shows that $d\tilde{\boldsymbol\beta}$ is distributed as stated there, with π_1, π_0 replaced by r_1, r_0.

In practice, the simplest way to apply logistic regression when n_1 and n_0 are fixed is simply to ignore this fact. The standard programs maximize (1.9) over the possible choices of β_0, $\boldsymbol\beta$, and then present the maximizer $\tilde{\boldsymbol\beta}$ as the estimate of $\boldsymbol\beta$. This method can be shown to be asymptotically equivalent to the conditional maximum likelihood estimator based on (5.3).

[*Received December 1974. Revised March 1975.*]

REFERENCES

[1] Anderson, T.W., *An Introduction to Multivariate Statistical Analysis*, New York: John Wiley & Sons, Inc., 1958.

[2] Cox, D.R., *Analysis of Binary Data*, London: Chapman and Hall, Ltd., 1970.

[3] Dempster, A., "Aspects of the Multinomial Logit Model," *Multivariate Analysis*, 3 (March 1973), 129–42.

[4] Halperin, M., Blackwelder, W.C. and Verter, J.I., "Estimation of the Multivariate Logistic Risk Function; A Comparison of the Discriminant Function and Maximum Likelihood Approaches," *Journal of Chronic Diseases*, 24 (January 1971), 125–58.

[5] Lehmann, E., *Testing Statistical Hypotheses*, New York: John Wiley & Sons, Inc., 1959.