Ann. Inst. Statist. Math.
29 (1977), Part A, 247-257

COVARIANCE ADJUSTED DISCRIMINANT FUNCTIONS

PETER A. LACHENBRUCH

(Received May 7, 1974)

1. Introduction

In 1948, Cochran and Bliss [1] introduced the notion of using covariates in discriminant functions. These were variables that in themselves had no discriminating power, but because they were correlated with other variables, they could be useful in combination with those other variables. They showed that the Mahalanobis distance between the two populations always increased, and thus the power of tests would be increased. The use of covariates could never hurt. The method was simple; one computed the usual covariance adjustment for the discriminators and did a standard linear discriminant analysis on the adjusted variables. Somewhat later Cochran [2] compared the performance of this procedure with doing a discriminant analysis on the complete set of discriminators and covariates. In this study, he found that the covariance technique produced more powerful significance tests, but the gain was trivial for assigning new observations. This is not surprising, for let the densities be $f_1(\boldsymbol{x}, \boldsymbol{z})=f_1(\boldsymbol{x}|\boldsymbol{z})g(\boldsymbol{z})$ in Π_1 and $f_2(\boldsymbol{x}, \boldsymbol{z})=f_2(\boldsymbol{x}|\boldsymbol{z})g(\boldsymbol{z})$ in Π_2. (Π_1 and Π_2 denote which population we are sampling from.) For the discriminant function, we assume $f_i(\boldsymbol{x}, \boldsymbol{z})$ are multivariate normal. The optimal rule is to assign the unknown observation to Π_1 if

$$\frac{f_1(\boldsymbol{x}, \boldsymbol{z})}{f_2(\boldsymbol{x}, \boldsymbol{z})} > \ln \frac{1-p}{p},$$

but this is equivalent to

$$\frac{f_1(\boldsymbol{x}|\boldsymbol{z})}{f_2(\boldsymbol{x}|\boldsymbol{z})} > \ln \frac{1-p}{p}.$$

Cochran assumed a linear relation between $\boldsymbol{x}$ and $\boldsymbol{z}$. Denote this by $\mathrm{E}(\boldsymbol{x}|\boldsymbol{z})=\boldsymbol{\mu}_1+A_1'\boldsymbol{z}$ in Π_1 and $\mathrm{E}(\boldsymbol{x}|\boldsymbol{z})=\boldsymbol{\mu}_2+A_2'\boldsymbol{z}$ in Π_2. If the relationship between $\boldsymbol{z}$ and $\boldsymbol{x}$ is such that $f_1(\boldsymbol{x}, \boldsymbol{z})$ and $f_2(\boldsymbol{x}, \boldsymbol{z})$ have the same covariance matrix, then the optimal rule is the linear discriminant function. This occurs if $A_1=A_2$, that is, the relation differs only with respect to $\boldsymbol{\mu}_i$. Otherwise, the quadratic rule will be optimal. Depending

on the difference between A_1 and A_2, and the magnitude of the components of z, the linear discriminant function may be an excellent approximation to the optimal rule. If the relationship is not linear, then the pair (x, z) cannot be multivariate normal if z is multivariate normal, and x is conditionally multivariate normal.

We have seen one reason why covariance adjusted discriminant functions are used. That is, the model is appropriate, and it is desired to test for significant differences between the two groups. For classifying observations, when there is a linear relation with the covariate, little or no gain accrues from using the covariate adjusted function over the function including all variates and covariates. A second reason for using the covariate adjusted function is that the relation between the covariates and the discriminators may not be linear. If this is so, then the density functions are not multivariate normal, and may be quite messy. If the assumption is made that $f(x|z)$ is normal, the usual form of the linear discriminant function holds with the means in Π_1 and Π_2 replaced by the conditional means in Π_1 and Π_2. It is not necessary that the means have the same functional form in Π_1 and Π_2, but only the marginal distribution of z has to be the same.

In their original article, Cochran and Bliss were not concerned specifically with the form of classification regions. This paper will discuss them in some detail for the case of one discriminator and one covariate. Some of the regions have rather strange shapes which may lead statisticians to prefer different forms for the conditional means than might be assumed otherwise. Some other problems can be subsumed under this model. For example, suppose it is desired to predict survival or non-survival of an operation, but the number of patients may be too small to calculate adequate statistics at any one hospital. By combining data from several hospitals, one can get a more stable estimate of the covariance matrix if it can be assumed constant over hospitals.

2. The covariance adjusted discriminant function

Let z denote the vector of covariates and let x be the k-dimensional vector of discriminators. Suppose the conditional distribution of x given z is multivariate normal with mean $h_1(z)$ in Π_1 and $h_2(z)$ in Π_2, and covariance matrix Σ in both Π_1 and Π_2. z itself may have a distribution, or it may be a variable that is non-stochastic and under the control of the investigator. Then the optimal classification rule is to assign an unknown observation x to Π_1 if

$$D_c(x|z)=\left(x-\frac{1}{2}(h_1(z)+h_2(z))\right)'\Sigma^{-1}(h_1(z)-h_2(z))>\ln\frac{1-p}{p}$$

where p is the a priori probability of $\boldsymbol{x}$ belonging to Π_1. This is easily seen since the optimal rule assigns $\boldsymbol{x}$ to Π_1 if

$$\ln \frac{f_1(\boldsymbol{x}|z)}{f_2(\boldsymbol{x}|z)} > \ln \frac{1-p}{p} .$$

Now

$$\begin{aligned}
\ln \frac{f_1(\boldsymbol{x}|z)}{f_2(\boldsymbol{x}|z)} &= \frac{\frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{h}_1(z))'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{h}_1(z))\right)}{\frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{h}_2(z))'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{h}_2(z))\right)} \\
&= \ln \exp\left(\boldsymbol{x}'\Sigma^{-1}\boldsymbol{h}_1(z) - \frac{1}{2}\boldsymbol{h}_1(z)'\Sigma^{-1}\boldsymbol{h}_1(z) - \boldsymbol{x}'\Sigma^{-1}\boldsymbol{h}_2(z)\right. \\
&\qquad \left. + \frac{1}{2}\boldsymbol{h}_2(z)'\Sigma^{-1}\boldsymbol{h}_2(z)\right) \\
&= \left(\boldsymbol{x} - \frac{1}{2}(\boldsymbol{h}_1(z)+\boldsymbol{h}_2(z))\right)'\Sigma^{-1}(\boldsymbol{h}_1(z)-\boldsymbol{h}_2(z))
\end{aligned}$$

which yields the desired result.

If the parameters are unknown, they can be estimated by maximum likelihood. No general results can be given regarding estimation of $\boldsymbol{h}_i(z)$, but for specific examples, it is a straightforward matter to estimate parameters of these functions. Estimation of Σ seems fairly simple, if the ML estimates of $\boldsymbol{h}_1(z)$ and $\boldsymbol{h}_2(z)$ are available. The ML estimate of Σ is

$$\hat{\Sigma} = \frac{1}{n_1+n_2}\left\{\sum_{i=1}^{n_1}(\boldsymbol{x}_{1i}-\boldsymbol{h}_1(z_i))(\boldsymbol{x}_{1i}-\boldsymbol{h}_1(z_i))' + \sum_{i=1}^{n_2}(\boldsymbol{x}_{2i}-\boldsymbol{h}_2(z_i))(\boldsymbol{x}_{2i}-\boldsymbol{h}_2(z_i))\right\} .$$

That is, replace the $\boldsymbol{x}$ values by their covariance adjusted values and estimate Σ in the usual manner.

If z has some distribution, say $g(z)$, it may be of some interest to replace $\boldsymbol{h}_i(z)$ by the relation $\boldsymbol{h}_i(z) = \mathrm{E}(\boldsymbol{h}_i(z)) + \boldsymbol{a}_i(z)$. Denote $\mathrm{E}(\boldsymbol{h}_i(z))$ by $\boldsymbol{\mu}_i$. Note that $\mathrm{E}(\boldsymbol{a}_i(z)) = 0$. Then

$$\mathrm{D}_C(\boldsymbol{x}|z) = \left(\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2+\boldsymbol{a}_1(z)+\boldsymbol{a}_2(z))\right)'\Sigma^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2+\boldsymbol{a}_1(z)-\boldsymbol{a}_2(z)) .$$

Now from this representation we can see that if $\boldsymbol{a}_1(z) = \boldsymbol{a}_2(z)$, the coefficients of the covariance adjusted discriminant function are the same as if we had ignored the covariates entirely. However, the constant term is affected, so the covariance adjustment will still have an effect. This particular case states that the relation between $\boldsymbol{x}$ and z is identical in Π_1 and Π_2 except that the intercept is different.

3. Some special cases

Suppose an operation is to be performed, and it is desired to predict success or failure of the operation. Because the condition is a rare one, few patients are seen at any one hospital, but several hospitals may be combined to give data. Let z be a J-vector of dummy variables which indicate which hospital the patient comes from (i.e., $z_i=0$ or 1 and $\sum z_i=1$). Then we may write

$$h_i(z)=\mu_i+A_i'z$$

where $A_i=(a_{j'j}^{(i)})$ is a matrix of coefficients which have the property that

$$\sum_{j=1}^{J} a_{j'j}^{(i)}=0 \qquad \text{for } i=1, 2,\ j'=1,\cdots, J'.$$

Here J is the number of covariates (dummy variables) and J' is the number of discriminators. The coefficients $a_{kj}^{(i)}$ represent inter-hospital differences, which may be due to differences in population served, skill or hospital personnel, facilities of the hospitals and so forth. Estimation is quite simple. We have

$$\hat{\mu}_i=\frac{1}{\sum_{l=1}^{L} n_{il}}\sum_{l=1}^{L}\sum_{k=1}^{n_{il}} x_{ilk}=\bar{x}_{i\cdot\cdot} \qquad i=1, 2$$

where x_{ilk} is the kth observation in the lth hospital from the ith population. Define $\bar{x}_{il\cdot}=\frac{1}{n_{il}}\sum_{k=1}^{n_{il}} x_{ilk}$. Then $a_l^{(i)}=\bar{x}_{il\cdot}-\bar{x}_{i\cdot\cdot}$ is the ML estimate satisfying $\sum n_{ij}\hat{a}_j^{(i)}=0$.

For the remainder of this paper we shall be concerned with a set of specific examples where x and z have only one component and we shall give some picture of the assignment regions which result from various different forms of $h_i(z)$. Note first that the covariance adjusted discriminant function has the form assign x to Π_1 if

$$D_C(x\,|\,z)=\left(\left(x-\frac{1}{2}(h_1(z)+h_2(z))\right)(h_1(z)-h_2(z))\right)\Big/\sigma^2>\ln\frac{1-p}{p}\,.$$

We can assume that $\sigma^2=1$ in the following examples. They were chosen to show the wide variety of shapes of classification regions that can be obtained with very simple covariance functions.

The first case is the linear relation

$$h_i(z)=a_i+b_iz\,.$$

Then we have

$$D_C(x|z)=\left(\left(x-\frac{1}{2}(a_1+a_2+z(b_1+b_2)\right)\right)(a_1-a_2+z(b_1-b_2))\right).$$

and we assign x to Π_1 when $D_C(x|z)>\ln((1-p)/p)$. If $b_1\neq b_2$, the boundary of this region is a hyperbola when $p\neq 1/2$, and a degenerate hyperbola (i.e., a pair of straight lines) when $p=1/2$. This can be seen as follows. If $a_1-a_2+z(b_1-b_2)>0$, then we assign x to Π_1 when

$$x-\frac{1}{2}(a_1+a_2+z(b_1+b_2))>\left(\ln\frac{1-p}{p}\right)\Big/(a_1-a_2+z(b_1-b_2))$$

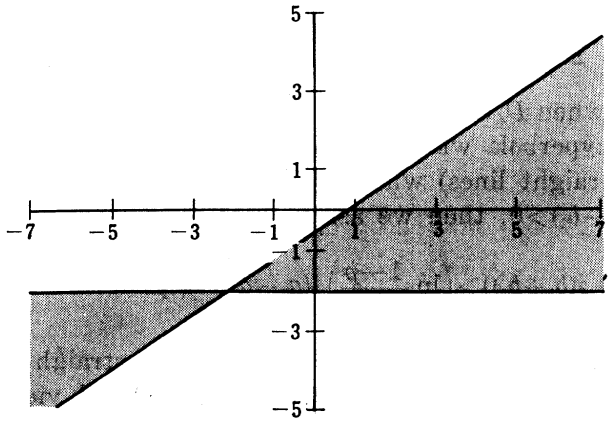
which is the equation of a hyperbola. If $p=1/2$ we get a straight line. The case in which $a_1-a_2+z(b_1-b_2)<0$ is handled similarly and yields a second hyperbola, and another straight line. The case illustrated in Fig. 1-a), b), c) has $h_1(z)=2+2z$, $h_2(z)=z$ and is given for values of p such that $\ln((1-p)/p)=0$ for 1-a), 1 for 1-b) and -1 for 1-c). In this figure the shaded region is the region in which we would assign x to Π_1. This is a quadratic discriminant function. In general if z is univariate normal with mean 0 and variance 1 and x is univariate normal with mean $a+bz$ and variance 1, then the pair $\begin{pmatrix} x \\ z \end{pmatrix}$ has a bivariate normal distribution with mean $\begin{pmatrix} a \\ 0 \end{pmatrix}$ and covariance matrix $\begin{pmatrix} 1+b^2 & b \\ b & 1 \end{pmatrix}$. Thus if $b_1\neq b_2$ one is led to the quadratic function.

The second case is that of a quadratic relationship, $h_i(z)=a_i+b_iz+c_iz^2$. Again if $b_1=b_2$, $c_1=c_2$, a discriminant function that is linear is both z and x results. Otherwise the covariance adjusted function is linear only in x. Fig. 2 shows classification regions for $h_1(z)=z^2$ and $h_2(z)=2z^2$. Clearly, the mean in Π_2 is always greater than the mean in Π_1, so that large values of x should be assigned to Π_2. Note the behavior at $z=0$. In this case the means are both equal to 0. In Fig. 2-a), x is assigned to Π_1 if it is less than 0. In 2-b) it is always assigned to Π_2 since $\ln((1-p)/p)=1$ implies Π_2 is more likely. In 2-c), it is always assigned to Π_1.

Case three is a step function. In Π_1, $h_1(z)=[z]$, and in Π_2, $h_2(z)=2[z]$. Thus if $z>0$, the mean in Π_2 is greater than the mean in Π_1, so that large values of x are assigned to Π_2. On the other hand, when $z<0$, the mean in Π_1 is greater than the mean in Π_2, so large values of x are assigned to Π_1. Fig. 3-b) is of some interest because of the kinky shape of the assignment region. Note that it is for $\ln((1-p)/p)=2$.

Case four uses $h_1(z)=|z|$, $h_2(z)=2|z|$. Its shape is similar to the regions of case 2, for the same reasons.

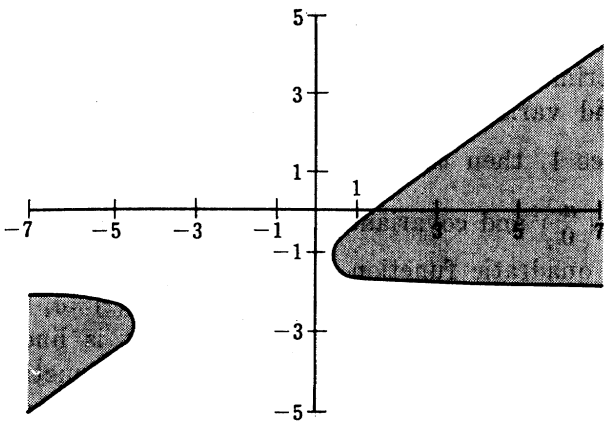
The final special case is the so-called hockey stick function. This is the case where there is a fixed response up to a threshold level, and a



a) $h_1=2+2z$

$h_2=z$

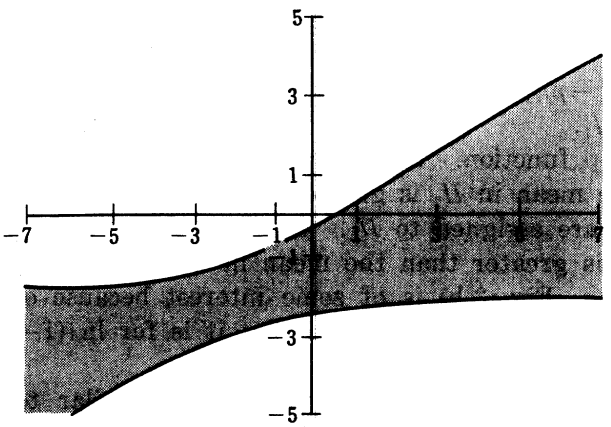
$\ln\left(\frac{1-p}{p}\right)=0$



b) $h_1=2+2z$

$h_2=z$

$\ln\left(\frac{1-p}{p}\right)=1$

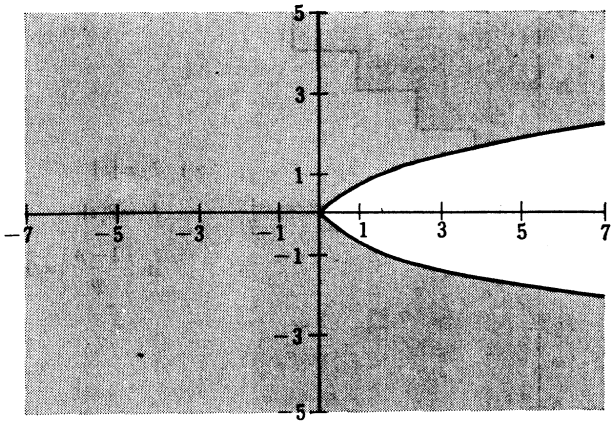


c) $h_1=2+2z$

$h_2=z$

$\ln\left(\frac{1-p}{p}\right)=-1$

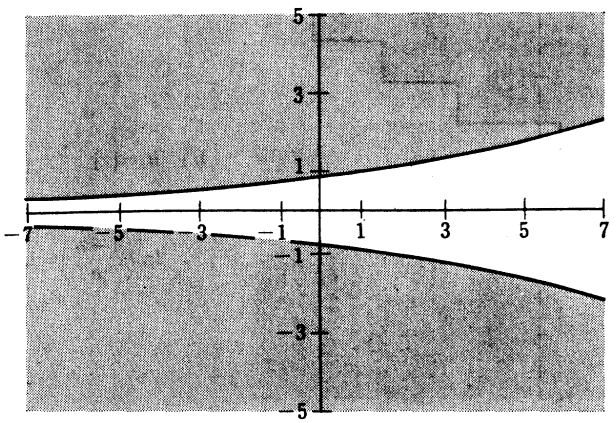
Fig. 1



a) $h_1=z^2$

$h_2=2z^2$

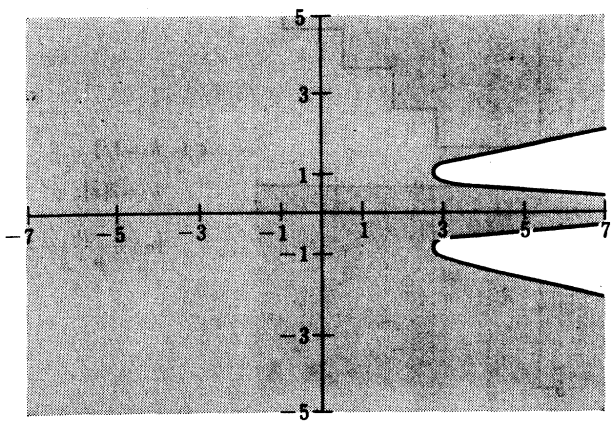
$\ln\left(\frac{1-p}{p}\right)=0$



b) $h_1=z^2$

$h_2=2z^2$

$\ln\left(\frac{1-p}{p}\right)=1$

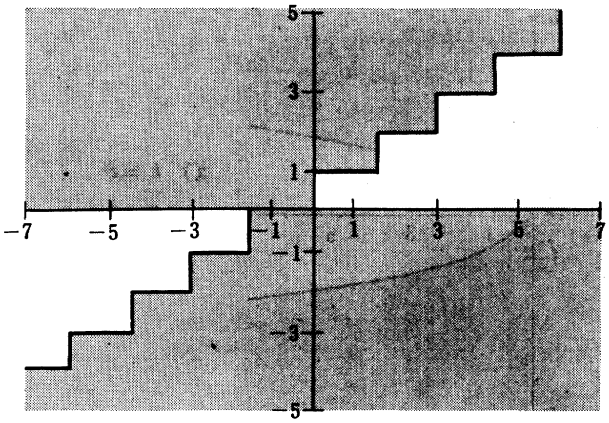


c) $h_1=z^2$

$h_2=2z^2$

$\ln\left(\frac{1-p}{p}\right)=-1$

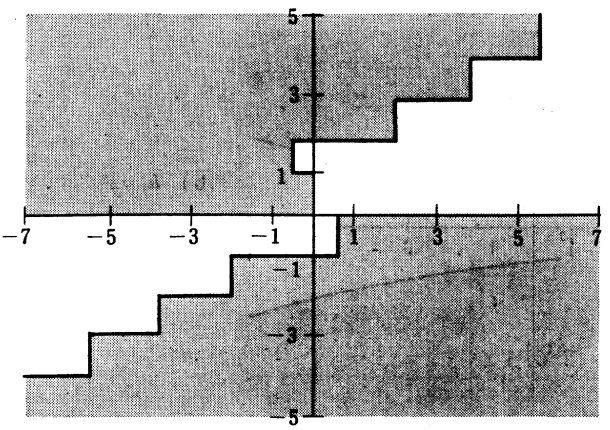
Fig. 2



a) $h_1=[z]$

$h_2=2[z]$

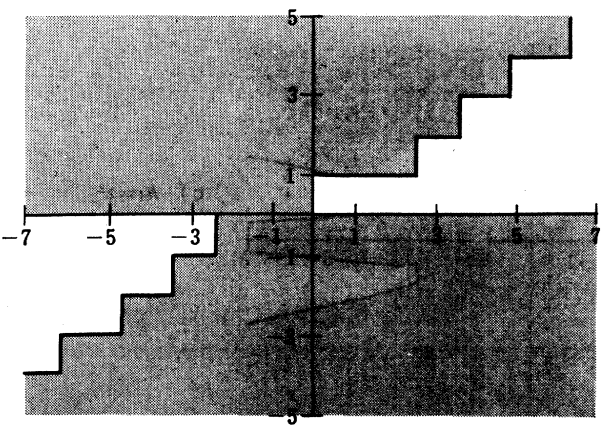
$\ln\left(\frac{1-p}{p}\right)=0$



b) $h_1=[z]$

$h_2=2[z]$

$\ln\left(\frac{1-p}{p}\right)=1$

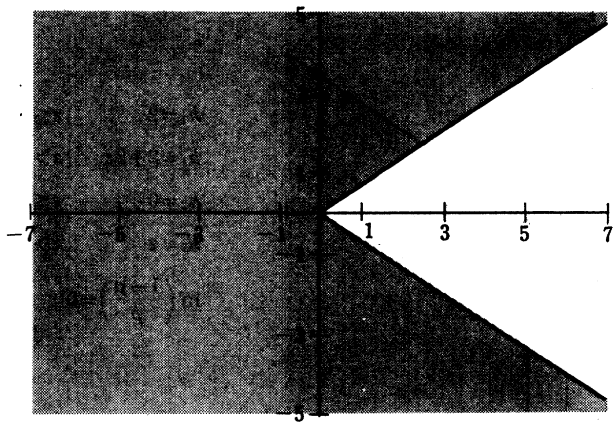


c) $h_1=[z]$

$h_2=2[z]$

$\ln\left(\frac{1-p}{p}\right)=-1$

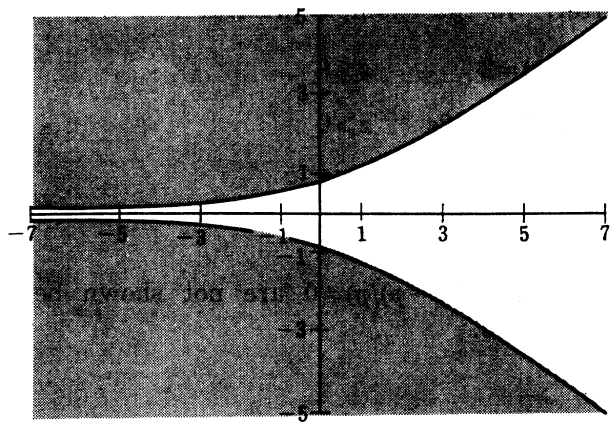
Fig. 3



a) $h_1=|z|$

$h_2=2|z|$

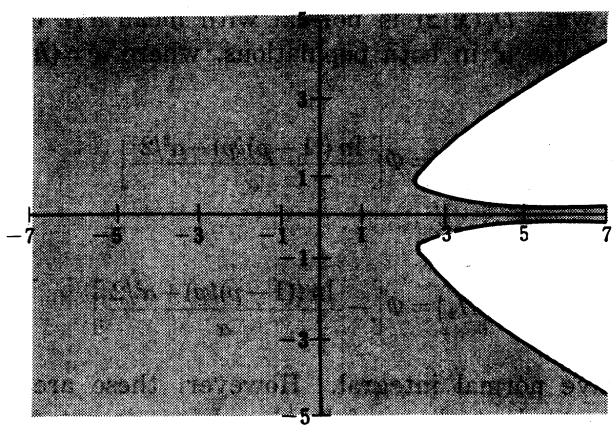
$\ln\left(\frac{1-p}{p}\right)=0$



b) $h_1=|z|$

$h_2=2|z|$

$\ln\left(\frac{1-p}{p}\right)=1$



c) $h_1=|z|$

$h_2=2|z|$

$\ln\left(\frac{1-p}{p}\right)=-1$

Fig. 4

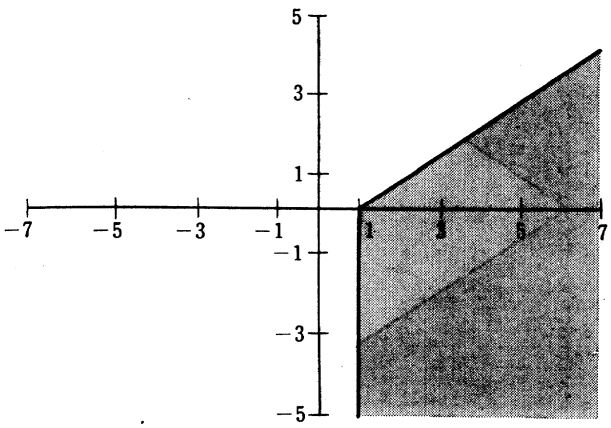


Fig. 5

linear regression thereafter. In the example illustrated in Fig. 5,

$$\begin{aligned} h_1(z)&=2 && z\leqq 0\\ h_1(z)&=2+2z && z>0\\ h_2(z)&=0 && z\leqq 0\\ h_2(z)&=z && z>0\,. \end{aligned}$$

The corresponding figures for $\ln((1-p)/p)\neq 0$ are not shown because they differ minutely from Fig. 5.

4. Error rates

For a fixed z, it is quite easy to determine the error rate when the parameters are known. $D_C(\boldsymbol{x}|z)$ is normal with mean $\alpha^2/2$ in Π_1, $-\alpha^2/2$ in Π_2 and has variance α^2 in both populations, where $\alpha^2=(\boldsymbol{h}_1(z)-\boldsymbol{h}_2(z))'\Sigma^{-1}(\boldsymbol{h}_1(z)-\boldsymbol{h}_2(z))$. Thus

$$P_1=\mathrm{P}\left(D_C(\boldsymbol{x}|z)<\ln\frac{1-p}{p}\,\middle|\,\Pi_1\right)=\Phi\left[\frac{\ln((1-p)/p)-\alpha^2/2}{\alpha}\right]$$

and

$$P_2=\mathrm{P}\left(D_C(\boldsymbol{x}|z)>\ln\frac{1-p}{p}\,\middle|\,\Pi_2\right)=\Phi\left[-\frac{\ln((1-p)/p)+\alpha^2/2}{\alpha}\right]$$

where Φ is the cumulative normal integral. However, these are conditional on z. They are likely to be useful only when z is not a random variable. For example, in the surgery example, each hospital would want to know what its error rate was.

Table 1 Expected error rates

$h_1(z)$	$h_2(z)$	p	$\ln \frac{1-p}{p}$	$E(P_1)$	$E(P_2)$	Total error
$\|z\|$	$2\|z\|$	.731	−1	.032	.793	.237
		.5	0	.348	.348	.348
		.269	1	.793	.032	.237
		.119	2	.902	.006	.113
z^2	$2z^2$	.731	−1	.027	.743	.220
		.5	0	.351	.351	.351
		.269	1	.743	.027	.220
		.119	2	.832	.006	.104

If z has a normal distribution we can integrate over z to find expected error rates. This was done numerically for two cases with results as given in Table 1. It was assumed that z was $N(0, 1)$.

The error rates are symmetrical about values of $p=.5$ and they behave as expected. The total error reaches a maximum at $p=.5$. $E(P_1)$, the expected error rate in Π_1, is a monotonically decreasing function of p while $E(P_2)$ is monotonically increasing.

UNIVERSITY OF CALIFORNIA AND UNIVERSITY OF NORTH CAROLINA

REFERENCES

[1] Cochran, W. G. and Bliss, C. I. (1948). Discriminant functions with covariance, *Ann. Math. Statist.*, 18, 151-176.

[2] Cochran, W. G. (1964). Comparison of two methods of handling covariates in discriminatory analysis, *Ann. Inst. Statist. Math.*, **16**, 43-53.