BIOSTATISTICS: Statistics in Biomedical, Public Health
and Environmental Sciences, edited by P.K. Sen




# A COMPARISON OF THE DISCRIMINATION OF DISCRIMINANT ANALYSIS AND LOGISTIC REGRESSION UNDER MULTIVARIATE NORMALITY

Frank E. Harrell, Jr. and Kerry L. Lee

Division of Biometry
Department of Community and Family Medicine
Duke University Medical Center
Durham, North Carolina

When sampling from two multivariate normal populations having equal covariance matrices, both the Fisher linear discriminant function (LDF) and logistic multiple regression model (LRM) can be used to derive valid estimates of the probability that a new observation comes from one of the two populations. In this setting, the LDF has been shown to yield asymptotically smaller relative classification error rates. When assumptions for the LDF are violated, LRM has been shown to be superior. In many situations, one is interested in using more information from a probability model than what is needed to devise a binary classification rule. In this paper we will study the relative performance of the LDF and LRM when all assumptions of the LDF are satisfied, to compare the spectrum of posterior probabilities arising from the two models. The cross-validation predictive accuracy and extent of separation (discrimination) of the posterior probabilities from the two methods will be assessed.

KEY WORDS: Discriminant analysis, logistic regression, predictive accuracy, discrimination, multivariate normality, maximum likelihood estimators, posterior probabilities

## 1. INTRODUCTION

Suppose that a p-dimensional random vector X can be observed from one of two p-variate normal populations with equal covariance matrices,

$$X \sim N_p(\mu_0, \Sigma) \text{ with probability } \Pi_0,$$
$$X \sim N_p(\mu_1, \Sigma) \text{ with probability } \Pi_1. \qquad (1.1)$$

If a random sample under model (1.1) is available, say $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, where $Y_j$ is an indicator of the population being sampled so that $Y_j=1$ with probability $\Pi_1$, Bayes' theorem (Truett et al., 1967) can be used to derive a model for the probability of $Y_j$ conditional on $X_j$:

$$\Pr(Y_j=1|X_j) = \frac{1}{1 + \exp\{-(a + X_j\beta)\}}. \qquad (1.2)$$

Under the assumptions in (1.1), the maximum likelihood estimates of a and $\beta$ are given by

$$\hat{\beta} = S^{-1}(\bar{X}_1,-\bar{X}_0)', \qquad (1.3)$$

$$\hat{a} = \log_e n_1/n_0 - 1/2(X_0+X_1)\ \hat{\beta}$$

where

$$\bar{X}_0 = \sum_{y_j=0} X_j/n_0,\ \bar{X}_1 = \sum_{y_j=0} X_j/n_1,$$

$$S = [\sum_{y_j=0}(X_j-\bar{X}_0)(X_j-\bar{X}_0)' + \sum_{y_j=1}(X_j-\bar{X}_1)(X_j-\bar{X}_1)']/n \qquad (1.4)$$

$$n_1 = \sum_{j=1}^{n} Y_j,\ n_0 = n - n_1.$$

These estimates form the basis of the LDF procedure, with the LDF being $\hat{a} + X\hat{\beta}$. The LDF has been used extensively, especially in biomedical and epidemiologic applications (see Abernathy, et al., 1966, for example). Lachenbruch (1975) has provided an excellent extensive bibliography.

By assuming only random sampling and (1.2), maximum likelihood estimates of a and $\beta$ can be derived conditional on the observed $X_j$ (Walker and Duncan, 1967). These estimates must be computed iteratively; there is no closed-form solution to the likelihood equations arising from (1.2). These conditional maximum likelihood estimates form the basis of the LRM estimation procedure.

Whichever estimation method is used, the customary binary classification rule consists of assigning an observation X to population 1 if the estimate of $Pr(Y_j=1|X)$ from (1.2) is greater than 1/2, or to population 0 otherwise. Efron (1975) demonstrated that if (1.1) holds and n increases without bound, the LDF yields lower rates of binary classification errors relative to the LRM. He showed that the relative efficiency of the LRM also decreases when $\Pi_1$ is far from 1/2 or when the separation between populations increases. (This relative efficiency was found to range from about 1/2 to 2/3.) The separation is measured by the square root of the Mahalanobis distance, given by

$$D = [(\mu_1-\mu_0)' \Sigma^{-1} (\mu_1-\mu_0)]^{\frac{1}{2}}. \qquad (1.5)$$

When (1.1) is not satisfied, Halperin, Blackwelder, and Verter (1971) have shown the LDF method yields biased estimates of a and $\beta$, and others (Press and Wilson, 1978, D'Agostino and Pozen, 1982) have shown that the LRM has better classification error rates. For a variety of reasons, Press and Wilson strongly advocate

the use of the LRM when there is any evidence that (1.1) is violated, both in formulating a binary decision rule and in deriving absolute posterior probabilities.

Even when (1.1) holds, we believe there are several unanswered questions concerning the comparison of the LDF and LRM. 1) What is the relative behavior of the two for finite sample sizes? 2) Should one calculate relative error rates, or the difference of absolute error rates? 3) Should one even calculate error rates? Are the two models really only used to derive binary decision rules? Shouldn't a more general measure of predictive discrimination or predictive accuracy be used to judge the performance of the two models?

We believe the questions in 3) are especially important. It is apparent, especially in biological and medical applications, that these models are frequently used to derive probabilities of Y=1 rather than for making binary decisions, which can be quite arbitrary depending on the cutoff value chosen. Even when the user of the model is not accustomed to dealing with absolute probabilities, she often wishes to establish a "gray zone" for application of the model. In diagnostic modeling problems, for example, a physician may classify the disease as "present" if the estimated probability is greater than 0.90, "absent" if less than 0.10, and perform another diagnostic test if the probability is between 0.10 and 0.90. In this example, the trinary decision rule could be studied, but the element of arbitrariness of the 0.10 and 0.90 thresholds detracts from the analysis. Another major disadvantage of using simple error rates is that a predicted probability of say 0.51 carries the same penalty as a probability of 0.99 when the observation arose from population 0.

In this paper we will discuss four measures of comparing the predictive accuracy of the two models and present results of simulation studies that use these measures to compare LDF and LRM when (1.1) holds.

## 2. MEASURES OF PREDICTIVE ACCURACY

There are two important aspects of the predictive accuracy of a model. Reliability refers to the degree of bias of estimates. If a model estimates $Pr(Y=1|X)$ to be 0.80, 80% of observations with like values of X should have Y=1. Discrimination refers to the ability of a model to discriminate or separate values of Y. When Y is binary, discrimination is a measure of the extent to which observations with Y=1 have higher predicted probabilities of Y=1 than do the observations with Y=0. Discrimination is the more important aspect of predictive accuracy, we believe, because good discrimination is necessary for accurate predictions. Reliability can be achieved by calibration without affecting discrimination,

whereas a model that does not discriminate well cannot be fixed.

For each method of assessing accuracy discussed below, it is assumed that the model is derived on a training sample and tested in a separate test sample, to obtain an unbiased estimate of the method's accuracy.

A very general measure of discrimination can be derived from the Kendall-type (Goodman and Kruskal, 1979) rank correlation between predicted probabilities and observed outcomes. The simplest way to state such correlation indices is through the concordance probability which we label c. When predicting the probability that Y=1, this index is defined by the proportion of pairs of observations, one having Y=0 and the other having Y=1, such that the one having Y=1 also had the higher predicted probability. If the two probabilities are tied, that pair receives a score of 1/2 instead of 0 or 1, i.e.

$$c = \sum_{\substack{i=1 \\ Y_i=0}}^{n} \sum_{\substack{j=1 \\ Y_j=1}}^{n} [I(P_j > P_i) + 1/2\ I(P_j=P_i)]/n_0 n_1, \qquad (2.1)$$

where $P_k$ denotes an estimate of $P(Y_k=1|X_k)$ from (1.2). The quantity 2(c-1/2) is Somer's $D_{YP}$ rank correlation coefficient (Goodman et al., 1979). In this binary Y setting, c is proportional to the Wilcoxon-Mann-Whitney statistic for comparing $P_k$ values from the Y=0 and Y-1 samples. The c index in this case is also the area under a "receiver operating characteristic" curve, a quantity routinely used to measure the diagnostic ability of medical tests (Hanley and McNeil, 1982). The c index takes on a value of 1 for perfect discrimination and 1/2 for random predictions. An advantage of the c index is its ease of interpretation and its generalizability to more complex problems such as ordinal response and censored survival time data (Harrell, et al., 1984).

The c index is purely a measure of discrimination. There are many measures of predictive accuracy that take discrimination into account but also penalize a predictor for being unreliable. One such measure is the logarithmic probability scoring rule (Cox, 1970), stated by Shapiro (1977) as

$$Q = \sum_{i=1}^{n} [1+\log_2 (P_i^{Y_i} (1-P_i)^{1-Y_i})]/n\ . \qquad (2.2)$$

Q obtains a value of 1 for perfect predictions ($P_i=Y_i$ for all i), 0 for random predictions, and values less than 0 for predictions that are worse than random.

Another accuracy score similar to Q is the quadratic score of Brier (1950) considered extensively in meteorologic forecast assessment. We will state this index

here as

$$B = 1 - \sum_{i=1}^{n} (P_i - Y_i)^2/n \qquad (2.3)$$

Perfect predictions receive a score of 1, and perfectly bad predictions receive a score of 0.

A fourth way to quantify the relative accuracy of two predictors, which also takes both reliability and discrimination into account, involves using both predictors as covariates in a model to predict the values of Y in the test sample. We then ask the question: do the predictions of method 1 ($P_i^1$) add information about predicting Y to the predictions of method 2 ($P_i^2$) and vice-versa? A formal statistical test can readily be made for both hypotheses, and the relative predictive ability of each method may be measured by comparing a statistic testing the strength of an individual predictor to a statistic for testing the strength of the best linear combination of the predictors. Since Y is binary, it is natural to use the LRM itself for this purpose, taking as covariates logit ($P_i^1$) and logit ($P_i^2$), where logit (q) = $\log_e[q/(1-q)]$. Then a statistic for testing whether $P_i^1$ adds information to that provided by $P_i^2$ is a likelihood ratio chi-square statistic with one degree of freedom. A measure of the "adequacy" of $P_i^1$ is given by

$$A(P_i^1) = \frac{L(\text{logit } P_i^1)}{L(\text{logit } P_i^1, \text{ logit } P_i^2)}, \qquad (2.4)$$

where $L(\text{logit } P_i^1)$ is the log likelihood due to logit $P_i^1$ alone, and $L(\text{logit } P_i^1$ logit $P_i^2)$ is the log likelihood due to the best linear combination of logit $P_i^1$ and logit $P_i^2$. It should be noted that $L(\text{logit } P_i^1)$ would be a linear translation of Q in (2.2) had the logistic slope coefficient for logit $P_i^1$ been fixed at 1 and the intercept been fixed at 0. However, we are allowing these parameters to be estimated in the test sample. We define $A(P_i^2)$ in a similar way using (2.4). A disadvantage of both the Q and A measure is that if predicted probabilities of 0 or 1 exist, the measures are undefined. Since estimates of $\beta$ can be infinite for the LRM, this can be a problem. However, there is no problem with the LRM predicted probabilities themselves in such cases.

## 3. SIMULATION STUDIES

To estimate the four measures of predictive accuracy for LDF and LRM discussed in section 2 based on independent model validation, a series of simulation experiments was conducted. The value of p (the dimension of X) was fixed at 5 for all studies, and the covariance matrix was fixed at

$$\sum = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ & 1 & 1/2 & 1/2 & 0 \\ & & 1 & 1/2 & 0 \\ & & & 1 & 0 \\ & & & & 1 \end{bmatrix}$$

and $\mu_0$ was fixed at (0, 0, 0, 0, 0). Values of $\mu_1$ were varied to yield different separations (1.5) as shown in Table 1.

Table 1. Values of $\mu_1$ and Corresponding Separations

| $\mu_1$ | | | | | D |
|---|---|---|---|---|---|
| .5 | .5 | .5 | .5 | .5 | .94 |
| 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.87 |
| 1.4 | 1.4 | 1.4 | 1.4 | 1.4 | 2.62 |
| 1.75 | 1.75 | 1.75 | 1.75 | 1.75 | 3.27 |
| 2.0 | 2.0 | 2.0 | 2.0 | 2.0 | 3.74 |
| 2.5 | 2.5 | 2.5 | 2.5 | 2.5 | 4.68 |

The sample sizes consisted of n=50 and n=130. The prior probability $\Pi_1$, was set at 0.65 and 0.85 because the results of Efron (1975) suggest that LRM should have decreased efficiency as $\Pi_1$ moves farther from 0.5. For each combination of n, $\mu_1$, and $\Pi_1$, a training sample and test sample each of size n was generated using the RANNOR and RANUNI random number generators in SAS (Ray, 1982a). The LDF coefficients were estimated using the SAS DISCRIM procedure (Ray, 1982b), and the LRM parameters were fitted using the LOGIST procedure (Harrell, 1983). Posterior probabilities were then calculated on the test sample and c, B, and Q were calculated for the test sample. These c, Q, and B values were averaged over 60 samples (replications) for each combination of parameters. For the evaluation which used LRM likelihood ratio statistics in (2.4), all test samples were combined for a particular combination of parameters and only one set of summary statistics was calculated based on [illegible] observations.

## 4. RESULTS

The results of the simulation study are shown in Table 2. It can readily be seen that the discrimination ability of both the LDF and LRM increases smoothly with increasing D. In the situations studied, LDF is superior to LRM, but by quite a

small amount, judged by c, Q and B. The index A, where computable, quantifies the "adequacy" of LRM at 90% on the average.

Table 2. Discrimination and Predictive Accuracy Logistic Regression Model vs. Linear Discrimination Function

| | | | c | | A | | Q | | B | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| D | $\Pi_1$ | n' | LDF | LRM | LDF | LRM | LDF | LRM | LDF | LRM | d |
| .94 | .65 | 50 | .69 | .69 | .99 | .85 | .11 | .07 | .79 | .79 | .06 |
| | | 130 | .72 | .72 | 1.00 | .95 | .16 | .15 | .80 | .80 | .06 |
| | .85 | 50 | .67 | .67 | -- | -- | -- | -- | .87 | .86 | .06 |
| | | 130 | .71 | .71 | 1.00 | .88 | .44 | .42 | .88 | .88 | .03 |
| 1.87 | .65 | 50 | .89 | .88 | 1.00 | .90 | -- | -- | .87 | .86 | .06 |
| | | 130 | .90 | .90 | 1.00 | .95 | .45 | .42 | .88 | .88 | .04 |
| | .85 | 50 | .88 | .85 | -- | -- | -- | -- | .91 | .89 | .06 |
| | | 130 | .90 | .89 | .99 | .86 | .62 | .57 | .92 | .92 | .03 |
| 2.62 | .65 | 50 | .97 | .94 | -- | -- | -- | -- | .93 | .89 | .08 |
| | | 130 | .96 | .96 | 1.00 | .93 | .65 | .59 | .93 | .92 | .03 |
| | .85 | 50 | .94 | .92 | -- | -- | -- | -- | .94 | .91 | .08 |
| | | 130 | .96 | .96 | 1.00 | .91 | .77 | .71 | .95 | .95 | .03 |
| 3.27 | .65 | 50 | .99 | .97 | -- | -- | -- | -- | .96 | .93 | .06 |
| | | 130 | .99 | .98 | -- | -- | -- | -- | .96 | .95 | .03 |
| | .85 | 50 | .98 | .97 | -- | -- | -- | -- | .97 | .95 | .05 |
| | | 130 | .98 | .97 | -- | -- | -- | -- | .97 | .95 | .05 |
| 3.74 | .65 | 50 | .99 | .97 | -- | -- | -- | -- | .97 | .94 | .05 |
| 4.68 | .65 | 50 | 1.00 | 1.00 | -- | -- | -- | -- | .99 | .98 | .02 |

D defined in (1.5)
$\Pi_1$ = prior probability
c defined in (2.1)
A defined in (2.4)
A defined in (2.2)
B defined in (2.3)
d = average $|P^1 - P^2|$
-- = could not be estimated due to at least one infinite estimate of $\beta$

The asymptotic results of Efron (1975) suggest that the relative performance of LRM should suffer for larger D. There is a small indication of this in Table 2. For example, c differs by more than 0.01 in many cases for larger D. However, as D exceeds 2, both LDF and LRM appear to offer excellent predictive accuracy.

Another way to study the relative quality of posterior probabilities arising from LDF and LRM is to assess how the two sets of probabilities differ on the average. Over all cases studied, the average absolute difference between $P^1$ and $P^2$ was 0.06 when the prior probability was 0.65 and 0.04 for a prior probability of 0.85. Thus the two methods yield very similar probability estimates on the average.

Examining the distributions of absolute differences according to the level of LDF posterior probability sheds further light. For small to intermediate D, LRM probabilities agreed well with LDF for all levels of LDF probabilities. Table 3 displays this distribution for a large value of D (4.68) and for n=50, $\Pi_1$ = 0.65. For this case, the overall average difference was 0.02, and 2858 of the 3000 test observations had an LDF posterior probability less than 0.05 or exceeding 0.95. Of these, the average absolute difference between LDF and LRM probabilities was 0.007. Of the remaining 142 observations, the average disagreement was 0.26. Thus, a problem with LRM may lie in an instability of estimates in the middle probability range. In samples with large separations, one or more of the maximum likelihood estimates of regression coefficients may be infinite. For such situations, some large finite regression estimates must be used, and there will not be many estimated probabilities intermediate betwen zero or one. The ones that do fall in this intermediate range are likely to be unstable. This is probably related to Efron's finding for large D regarding a dropoff in relative accuracy for LRM using a rule based on exceeding the 0.5 predicted probability cutoff. However, it is important to note that in this situation, the LRM is able to separate Y=0 from Y=1 very well with extreme probabilities. Also, the proportion of correct classifications (using a 0.5 probability cutoff) for this extreme separation case is virtually as good with LRM as with LDF (0.986 vs. 0.976). Here the ratio of correct classification rates is near one although the relative efficiency of LRM as measured by the ratio of classification errors is 0.58.

Table 3. Distribution of Absolute Differences Between LDF and LRM Probability Estimates n=50, D=4.68, $\Pi_1$=.65 All Test Samples Combined

| $P^1$ | Frequency | Average $\lvert P^1-P^2 \rvert$ |
|---|---|---|
| 0 - .05 | 971 | .01 |
| .05 - .15 | 32 | .17 |
| .15 - .25 | 11 | .32 |
| .25 - .35 | 11 | .36 |
| .35 - .45 | 12 | .47 |
| .45 - .55 | 8 | .34 |
| .55 - .65 | 8 | .47 |
| .65 - .75 | 14 | .31 |
| .75 - .85 | 8 | .21 |
| .85 - .95 | 38 | .16 |
| .95 -1.00 | 1887 | .005 |
| Overall | 3000 | .02 |

## 5. DISCUSSION AND CONCLUSIONS

In assessing the predictive accuracy of a given method, it is important to study the entire spectrum of predicted values. The purpose of a probability model is

generally to predict absolute probabilities. In this light, the properties of predicted probabilities from the LRM or LDF in relation to the dependent variable Y should be used to gauge the usefulness of a LRM or LDF. Relative classification rates, especially misclassification rates, are not adequate measures of relative performance for probability models.

If one accepts this premise, one must choose a measure of discrimination or predictive accuracy. All four such measures studied here indicate that even when the conditions under which LDF was optimized are satisfied, the performance of LRM is very nearly as good as that of LDF for reasonable sample sizes and values of D. Predicted probabilities from the two methods also have a small absolute difference on the average when multivariate normality holds. Since others (Halperin et al., 1971) have shown that LDF can yield arbitrarily biased probability estimates when its assumptions are violated (e.g. one of the predictor variables is dichomotous), we argue that LRM is the tool of first choice among these two competitors. With the availability of efficient computers and computer programs, the issue of the computational requirements of the LRM becomes unimportant.

ACKNOWLEDGEMENTS: This work was supported by Research Grant HL-17670 from the National Heart, Lung, and Blood Institute, Bethesda, Maryland; Research Grants HS-03834 and HS-04873 from the National Center for Health Services Research, OASH, Hyattsville, Maryland; Training Grant LM-07003 and Research Grants LM-03373 and LM-00042 from the National Library of Medicine, Bethesda, Maryland; and grants from the Prudential Insurance Company of America, Newark, New Jersey; the Kaiser Family Foundation, Palo Alto, California, and the Andrew W. Mellon Foundation, New York, New York.

REFERENCES:

[1] Abernathy, J.R., Greenberg, B.G., and Donnelly, J.F., Application of discrimination functions in perinatal death and survival, Am. J. Obstetrics and Gynecology 95:860-7 (1966).

[2] Brier, G.W., Verification of forecasts expressed in terms of probability, Monthly Weather Rev. 78:1-3 (1950).

[3] Cox, D.R., The Analysis of Binary Data, London, Metheun (1970).

[4] D'Agostino, R.B. and Pozen, M.W., The logistic function as an aid in the detection of acute coronary disease in emergency patients (a case study), Stat. in Med. 1:41-8 (1982).

[5] Efron, B., The efficiency of logistic regression compared to normal discriminant analysis, J. Amer. Statist. Assoc. 70:892-8 (1975).

[6] Goodman, L.A. and Kruskal, W.H., Measures of Association for Cross-Classifications, New York, Springer-Verlag (1979).

[7] 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, J. Chron. Dis. 24:125-158 (1971).

[8] Hanley, J.A. and McNeil, B.J., The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143:29-36 (1982).

[9] Harrell, F.E., Lee, K.L., Califf, R.M., Pryor, D.B., and Rosati, R.A., Regression modeling strategies for improved prognostic prediction, Stat. in Med. 3:143-152 (1984).

[10] Harrell, F.E., The LOGIST Procedure. In: SUGI Supplemental Library User's Guide, 1983 Edition, Joyner, S.P. (ed). Cary, NC, SAS Institute, 181-202 (1983).

[11] Lachenbruch, P.A., Discriminant Analysis, New York, Hafner Press, 96-126 (1975).

[12] Press, S.J. and Wilson, S., Choosing between logistic regression and discriminant analysis, JASA 73:699-705 (1978).

[13] Ray, A.A. (ed), SAS User's Guide: Basics, Cary, NC, SAS Institute (1982a).

[14] Ray, A.A. (ed), SAS User's Guide: Statistics, Cary, NC, SAS Institute 461-73 (1982b).

[15] Shapiro, A.R., The evaluation of clinical predictions. A method and initial application, New Engl. J. Med. 296:1509-1514 (1977)

[16] Truett, J., Cornfield, J., and Kannel, W.B., A multivariate analysis of the risk of coronary heart disease in Framingham, J. Chron. Dis. 20:511-24 (1967).

[17] Walker, S.H. and Duncan, D.B., Estimation of the probability of an event as a function of several independent variables, Biometrika 54:167-79 (1967).