Choosing Between Logistic Regression and Discriminant Analysis

S. James Press; Sandra Wilson

Journal of the American Statistical Association, Vol. 73, No. 364. (Dec., 1978), pp. 699-705.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28197812%2973%3A364%3C699%3ACBLRAD%3E2.0.CO%3B2-0

Journal of the American Statistical Association is currently published by American Statistical Association.

Choosing Between Logistic Regression and Discriminant Analysis

S. JAMES PRESS and SANDRA WILSON*

Classifying an observation into one of several populations is discriminant analysis, or classification. Relating qualitative variables to other variables through a logistic cdf functional form is logistic regression. Estimators generated for one of these problems are often used in the other. If the populations are normal with identical covariance matrices, discriminant analysis estimators are preferred to logistic regression estimators for the discriminant analysis problem. In most discriminant analysis applications, however, at least one variable is qualitative (ruling out multivariate normality). Under nonnormality, we prefer the logistic regression model with maximum likelihood estimators for solving both problems. In this article we summarize the related arguments, and report on our own supportive empirical studies.

KEY WORDS: Logistic regression; Discriminant analysis; Qualitative variables; Classification.

1. INTRODUCTION

We will consider two problems. The first is the one of relating a qualitative dependent variable to one or more independent variables, which may or may not be qualitative. This problem has its multivariate analogs as well. When the dependent and independent variables are related by a logistic distribution functional form, the model is often referred to as a logistic regression.

The second problem under discussion is the one of classification, or discrimination, in which an object of given characteristics is to be classified into one of several alternative populations. The discrimination problem is distinct from the logistic regression problem and, as might be expected, solutions generally proposed for the one are different from those for the other, although they are related. In some situations (such as when at least one variable is qualitative), the differences in solution become substantial (Halperin, Blackwelder, and Verter 1971, p. 128).

The logistic regression model is usually formulated mathematically by relating the probability of some event, E, occurring, conditional on a vector, $\mathbf{x}$, of explanatory variables, to the vector $\mathbf{x}$, through the functional form of a logistic cdf. Thus,

$$p(\mathbf{x}) \equiv \Pr\{E|\mathbf{x}\} = 1/[1 + \exp\{-\alpha - \boldsymbol{\beta}'\mathbf{x}\}] ,$$

where $(\alpha, \boldsymbol{\beta})$ are unknown parameters that are estimated from the data. This model may be used for classifying an object into one of two populations by letting E denote the event that the object belongs to the first population, and letting $\mathbf{x}$ denote a profile vector of attributes of the object to be classified. (See, e.g., Nerlove and Press 1973 for more detail.)

The normal discrimination or classification problem is usually formulated by assuming that the two populations are multivariate normal with equal covariance matrices, $\boldsymbol{\Sigma}$, and that the costs of misclassification are equal. If $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ denote the mean vectors of the two populations, a likelihood ratio test readily yields the classification procedure to classify the object into the first population if

$$(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)'\boldsymbol{\Sigma}^{-1}\mathbf{x} + (\tfrac{1}{2})(\boldsymbol{\theta}_2 + \boldsymbol{\theta}_1)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1) \geq \log (q_2/q_1) ,$$

where (q_1, q_2) denote the prior classification probabilities. The parameters $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\Sigma})$ are estimated from the data, while (q_1, q_2) are assessed from the context. (See, e.g., Press 1972, p. 372 for more detail.)

The article begins with a synthesis of earlier published results on the problem of estimation involving qualitative variables in logistic regression. We contrast the merits of logistic regression maximum likelihood estimators (MLEs) with those of discriminant function estimators. We also illustrate how some trouble spots are often masked. Finally, we present the results of several empirical comparisons of the MLE logistic regression and discriminant analysis estimators in the contexts of (1) studies involving breast cancer, and (2) population changes across states of the U.S.

2. DISCUSSION

Discriminant function estimators have often been used in logistic regression, in both theory and applications (see, for example, Truett, Cornfield, and Kannel 1967). When such estimators were compared empirically with maximum likelihood estimators for logistic regression problems, however, they were found to be generally inferior, although not always by substantial amounts (see Halperin, Blackwelder, and Verter 1971, and D'Agostino et al. 1978). We will show why we prefer alternatives to

* S. James Press is Professor, Department of Statistics, University of California, Riverside, CA 92521. Sandra Wilson is Instructor, School of Business, University of Iowa, Cedar Falls, IA 50613. This research was supported in part by The Rand Corporation, and in part by the National Research Council of Canada under grant A-4050. We would also like to thank Dr. John Rolph of The Rand Corporation who commented on an early draft of this article, Professor S. Fienberg for suggestions that helped make the article more meaningful, and an associate editor whose careful reading and suggestions substantially improved the presentation.

© Journal of the American Statistical Association
December 1978, Volume 73, Number 364
Applications Section

discriminant function estimators for the logistic regression problem, as well as for the nonnormal discriminant analysis problem.

It has been common practice to use discriminant function estimators as starting values in iterative maximum likelihood estimation and in exploratory data analysis, for the purpose of fitting logistic regression models. Other starting and exploratory estimators that have been suggested include "reverse Taylor series approximations," and "conditional estimators" (Nerlove and Press 1973). "Conditional estimators" are obtained by maximizing the conditional likelihood (conditional on the explanatory variables). "Reverse Taylor series approximations" arise from the logistic cdf,

$$F(x) = 1/[1 + e^{-(a+bx)}] \;, \quad b \neq 0 \;, \quad -\infty < x < \infty \;.$$

Expanding about $x = \bar{x}$ (the sample mean) in a Taylor series, we get

$$F(x) = \left\{\frac{1}{1 + e^{-(a+b\bar{x})}} - \frac{b\bar{x}e^{-(a+b\bar{x})}}{[1 + e^{-(a+b\bar{x})}]^2}\right\} + \left\{\frac{be^{-(a+b\bar{x})}}{[1 + e^{-(a+b\bar{x})}]^2}\right\} x + R(x) \;,$$

where $R(x)$ denotes a remainder containing terms of order $O(x - \bar{x})^2$. Neglecting $R(x)$, this may be interpreted as the linear function $A + Bx$, where

$$A = \frac{1}{1 + e^{-(a+b\bar{x})}} - B\bar{x} \;, \quad B = \frac{be^{-(a+b\bar{x})}}{[1 + e^{-(a+b\bar{x})}]^2} \cdot$$

Solving these equations for a and b (in reverse from the usual direction), we find

$$b = B/[(A + B\bar{x})(1 - A - B\bar{x})]$$

and

$$a = -b\bar{x} - \log\left(\frac{1}{A + B\bar{x}} - 1\right)$$

as the reverse Taylor series approximation. The results are easily generalized when x, b, and B are vectors.

We prefer the reverse Taylor series estimators to the discriminant function estimators since the former are appropriate regardless of the underlying distribution of explanatory variables, while the latter are really appropriate and justifiable only under (a) multivariate normality of the explanatory variables (a difficult assumption to satisfy in practice), and (b) complete equality of all of the underlying covariance matrices. (Transformations to induce multivariate normality will not typically induce equality of covariance matrices.) In any case we are speaking only of initial values to get the iterative maximum likelihood estimation process started.

There are really two general questions (relating to the logistic regression problem) that need to be addressed. The first is, why use a logistic formulation rather than some other functional form? The second is, how should the parameters of the model be estimated? We now examine both questions.

2.1 Functional Form

The rationale for a logistic formulation of the relationship between qualitative and other variables, rather than a normal (probit analysis), angular (such as arcsine), or other relationship, has been discussed extensively in the literature and is summarized in the excellent book by Cox (1970). We do not repeat it here. To provide additional support for the logistic formulation, however, we note that Anderson (1972) pointed out that it results from a wide variety of underlying assumptions about the explanatory variables. In particular, the logistic formulation results not only from assuming that the explanatory variables are multivariate normally distributed with equal covariance matrices, but also from assuming that the explanatory variables are independent and dichotomous zero-or-one variables, or that some are multivariate normal and some dichotomous. Thus, one advantage of using the logistic model for discriminant analysis (rather than a linear discriminant function) is that it is relatively robust; i.e., many types of underlying assumptions lead to the same logistic formulation. The linear discriminant analysis approach, by contrast, is strictly applicable only when the underlying variables are jointly normal with equal covariance matrices.

Another advantage of logistic modeling relates to its use as an alternative to contingency table analysis. Gordon (1974) pointed out that logistic regression models have played a major role in biological and medical applications where cross-classified tables with large numbers of cells (and usually too few observations per cell) are typically replaced by a logistic or log-linear relationship among the variables, thus obviating the need for the table. In spite of how attractive the logistic formulation appears, however, Gordon cautions that the linear combination of variables in a multivariate logistic formulation is not always an appropriate model, in that some types of interaction may not be expressible in that form. Keeping in mind the possible hazards, however, the logistic function can be appropriately used in many such applications.

2.2 Estimation

The second question of fundamental interest centers around the problem of estimation. In their comparison of maximum likelihood estimation and linear discriminant function estimation (for a logistic regression), Halperin, Blackwelder, and Verter (1971) used an IBM-360-50 and -65 and found that "the times required for compilation and execution of the programs were higher for the maximum likelihood method than for the discriminant function method by factors ranging approximately from 1.3 to 2." Factors of economy in particular systems, and at particular times, however, will depend upon the relative efficiency of algorithms which may be developed. Economy of computation should not usually be the dominant consideration. Estimation efficiency is generally more important. Efron (1975) has shown that logistic regression estimators are between one-half and two-thirds as efficient as discriminant function estimators when the data are multivariate normal with equal covariance matrices. Thus, as long as the data are strictly normal with equal covariance matrices, linear discriminant function estimators are more economical to calculate and are more efficient than logistic regression MLEs. But, "Another important unanswered question is the relative efficiency under some model other than [multivariate normality] ..." (Efron 1975, p. 893). Simulation might be used to determine the relative efficiency of the two estimators under nonnormality, but it would not be surprising to find the sufficient estimator (maximum likelihood estimation) dominant.

On the other side of the question of estimation, however, there are many arguments which strongly militate against the general use of discriminant function estimators:

1. When the explanatory variables don't follow a multivariate normal distribution with equal covariance matrices for each state of the dependent variables, discriminant function estimators of the slope coefficients in the logistic regression will not be consistent. Thus, even in large samples there is no guarantee that good fits or good prediction will be obtained by this method. This means, in particular, that if the explanatory variables are binary, we cannot expect, with discriminant function estimators, to predict accurately the probability that the dependent variables will be in a given state, even with an infinite amount of data! Since many situations commonly encountered are of this type, having at least one dummy explanatory variable, the practical solution is to use a consistent method of estimation, such as MLE. The results on inconsistency are carefully and extensively proven, for various cases, in Halperin, Blackwelder, and Verter (1971). This argument really negates the use of discriminant function estimators in large samples (under nonnormality).

2. Discriminant function estimation can give misleading results regarding significance of the logistic regression coefficients when the normality condition is violated. That is, under nonnormality of the explanatory variables, a slope coefficient which is really zero will tend to be estimated as zero by MLE in large samples, but not necessarily by the discriminant function method; so when underlying normality is violated, meaningless variables will tend to be erroneously included in logistic regressions estimated by discriminant functions.

3. Numerical comparisons of MLE and discriminant function estimation for the logistic regression function model were made in Halperin, Blackwelder, and Verter (1971). They found that, under nonnormal conditions, "the maximum likelihood method usually gives slightly better fits to the model, as evaluated from observed and expected numbers of cases per decile of risk." They also found that "there is a theoretical basis for the possibility that the discriminant function will give a very poor fit, even if the [logistic regression] model holds."

4. Use of discriminant function estimators tends to mask the troublesome cases by not providing danger signals. As an illustration of the masking effect, we take an example suggested by G. Haggstrom of The Rand Corporation. Observed values of (z, y) are:

Observation	1	2	3	4	5	6	7	8
z_i	−4	−3	−2	−1	1	2	3	4
y_i	0	0	0	0	1	1	1	1

In this example, y is the dependent variable, and z the independent variable. If a logistic relationship is assumed, the MLEs of the slope and intercept terms fail to exist[1] (a warning of trouble). Moreover, this is a case in which a perfect fit to the data may be obtained. Thus, take y to be zero for $z \leq -1$, and one for $z \geq 1$. For $-1 < z < 1$, there is no unique solution and any curve is as good as any other curve in this region. Therefore, prior information should be used as a guide to understanding the relationship in this important central region. MLE signals that there is a problem here and that the researcher should exercise great care in fitting the central region. But discriminant function estimation provides no warning signal whatsoever, and quite incautiously suggests a slope coefficient estimate of $\tilde{b} = 4$. This follows since

$$\Pr\{y = 1 | z\} = [1 + \exp(-a - bz)]^{-1}$$

and $b \equiv (\theta_1 - \theta_0)/\sigma^2$, where the two hypotheses under consideration are $H_i\colon z | H_i \sim \mathrm{N}(\theta_i, \sigma^2)$, $i = 0, 1$.

5. The logistic regression model is well-known to have sufficient statistics associated with it (see, for instance, Press 1972, p. 267). The MLEs are functions of the sufficient statistics, while the discriminant function estimators are not. We know from the Rao–Blackwell theorem (Rao 1965), that we can always achieve smaller mean squared error using estimators based on sufficient statistics (when they exist, as they do here) than by using estimators not based upon sufficient statistics.

6. Maximum likelihood estimation of the logistic regression model forces the expected number of cases to equal the observed number of cases; in the notation of Halperin, Blackwelder, and Verter (1971), $\sum y_i = \sum P(x_{1i}, \ldots, x_{ki})$. This is an intuitively desirable property of any smoothing procedure, and it is one which is not enjoyed by the discriminant function approach (which sometimes generates estimated numbers much greater than actual numbers of observations).

7. There is some evidence that use of discriminant function estimators may tend to generate substantial bias in some applications. McFadden (1976, p. 521) concludes that in a Bayesian analysis, "for a typical [natural conjugate] prior distribution of the explanatory variables, multivariate normal, estimates of the selection prob-

[1] It is easy to check that one of the two equations that must be solved for unknown coefficients a and b is $\sum_1^8 z_i[1 + \exp(-a - bz_i)]^{-1} = 10$. But this is impossible since the left side must be strictly less than 10.

ability parameters [probabilities for the dependent variable] based on discriminant analysis will be substantially biased."

3. EMPIRICAL APPLICATIONS

To illustrate some of the ideas presented in the previous section, two classification problems involving empirical data were studied. In each case, both a linear discriminant analysis and a logistic regression were carried out. In both cases, the logistic regression outperformed the discriminant analysis in terms of the proportion of correct classifications (although computation time was greater for the logistic regression).

3.1 Example 1

The first example is based on data collected for certain breast cancer patients initially treated at the British Columbia Cancer Institute between 1955 and 1963. A study (Wilson 1977) was undertaken in 1976 to classify the patients by extent of nodal metastases from clinical and historical evidence. As with most medical data, the variables for the study were mixed—continuous and discrete. Many of the variables were binary. A linear discriminant analysis and a logistic regression were proposed to study the problem.

The individuals who were observed were 173 of the female breast cancer patients for whom no data were missing and whose nodal status had been determined by a surgical procedure. The patients were randomly divided into two groups. The first group of 115 patients was used as the training set for the classification procedures. The remaining group of 58 patients was used to cross-validate the classification functions estimated from the first group. These data are available on request from the authors.

The binary grouping variable was defined to be 0 if the lymph nodes were not involved with metastatic carcinoma, and 1 if the nodes were involved. The attribute (independent) variables were number of births, a history of hysterectomy (0–1), a history of benign breast disease during lactation (0–1), presence of nipple changes as the first disease symptom (0–1), and duration of symptoms in months. Thus, there were three binary independent variables, one ordered categorical independent variable, and one continuous independent variable.

The discriminant analysis was performed using the computer program BMD Stepwise Discriminant Analysis. The logistic regression was performed with the program listed in Nerlove and Press (1973, pp. 101–130), as implemented at the University of British Columbia. All computations were done on an IBM 370-168. After the appropriate functions were calculated, the individuals in both the training sets were classified from the estimated functions (the functions estimated from the training sets). The equations used for classification are given below.

The logistic regression classified 82 (65 + 17) of the 115 patients in the training set correctly, for a 71.30 percent classification rate (see Table 1). In the validation set, 36 (25 + 11) of the 58 patients were correctly classified, for a 62.07 percent correct classification rate. The discriminant analysis correctly classified 77 of the 115 patients in the training set, for a 66.96 percent correct classification rate. The prior probabilities used were .66 of having no metastases, and .34 of having metastases, the approximate proportions of actual cases in our data. In the validation set only 34 of the 58 patients were correctly classified, for a 58.62 percent correct classification rate. In Table 1, 0 and 1 indicate the absence and presence of nodal metastases, respectively.

Of particular interest in this example is the pattern of errors. When we looked at the cases that were misclassified in the validation set by each procedure, we found some overlap. Sixteen cases were misclassified the same by both procedures. All sixteen were positives that were classified as negatives. In addition, logistic regression misclassified six negatives as positives that discriminant analysis classified properly. Discriminant analysis classified eight positives as negatives that logistic regression correctly classified. Thus, there is a clear difference in the types of cases misclassified by the two procedures. The discriminant functions consistently misclassify many more patients into the 0 group than does the logistic function.

Computing time for logistic regression was found to be 1.38 times longer than that for discriminant analysis, but this may primarily reflect the particular computational algorithms that were used.

The associations of the variables with metastases provide some new areas of interest to the medical researcher. These associations may be studied by inspection of the

1. Summary of Classifications of Breast Cancer Patients by Logistic Regression and Discriminant Function Methods

		Discriminant classification			*Logistic regression classification*		
Case	*Actual group*	*0*	*1*	*Classification rate (%)*	*0*	*1*	*Classification rate (%)*
Training set	0	71	5	67	65	11	71
	1	33	6		22	17	
	Total	104	11		87	28	
Validation set	0	31	0	59	25	6	62
	1	24	3		16	11	
	Total	55	3		41	17	

2. *Summary of Classifications of 50 States by Logistic Regression and Discriminant Function Methods*

			Discriminant analysis			*Logistic regression*		
			Nodal status			*Nodal status*		
Case	*Set*	*Group*	*0*	*1*	*Classification rate (%)*	*0*	*1*	*Classification rate (%)*
1	Training set	0	13	8		17	4	
		1	9	10	57.50	5	14	77.50
	Validation set	0	3	1		4	0	
		1	1	5	80.00	1	5	90.00
2	Training set	0	12	6		14	4	
		1	6	16	70.00	4	18	80.00
	Validation set	0	5	2		5	2	
		1	2	1	60.00	3	0	50.00
3	Training set	0	16	5		17	4	
		1	9	10	65.00	6	13	75.00
	Validation set	0	2	2		3	1	
		1	0	6	80.00	0	6	90.00
4	Training set	0	17	4		18	3	
		1	6	13	75.00	2	17	87.50
	Validation set	0	2	2		2	2	
		1	3	3	50.00	3	3	50.00
5	Training set	0	14	5		15	4	
		1	6	15	72.50	4	17	80.00
	Validation set	0	4	2		4	2	
		1	1	3	70.00	0	4	80.00

equations estimated. The estimated functions are for logistic regression:

$$Y(\mathbf{X}) = .058 - .233X_1 - 1.096X_2 + .713X_3 - .028X_4 + .995X_5 ,$$

and for discriminant analysis:

$$U(\mathbf{X}) = .362 - .251X_1 - 1.245X_2 + 1.104X_3 - .036X_4 + 2.114X_5 ,$$

where X_1 is number of births, X_2 is hysterectomy, X_3 is benign breast disease, X_4 is duration of symptoms, and X_5 is nipple change symptom. As we might have expected from the earlier analyses, the two functions are quite similar. We observe that absence of nodal metastases was associated with a larger number of births than presence of metastases. Absence of metastases was also associated with a history of hysterectomy. A longer duration of symptoms was slightly correlated with absence of metastases. The presence of nodal metastases was associated with a history of benign breast disease during lactation and nipple changes as the first disease symptom.

3.2 Example 2

Population change data were collected for the 50 states of the U.S. Associated demographic data were collected for each state in an effort to explain population changes. The data were obtained from census records (Delury 1973). The percent change in population from the 1960 census to the 1970 census for each state was coded 0 or 1, according as the change was below or above the median change for all states. This became the binary "grouping" or dependent variable for the analyses. The median was chosen to divide the groups so that the prior probabilities are .5 for each group. The attribute (independent) variables were (for the year 1970) per capita income (in $1,000), birth rate (percent), death rate (percent), urbanization of population (0 or 1 as population is less than or greater than 70 percent urban), and absence or presence of coastline (0 or 1). Thus, there are three continuous independent variables and two binary independent variables. None of the continuous variables were normally distributed. Births were skewed, deaths were peaked (not normal at 5 percent of significance), and income was quite flat with a suggestion of trimodality. No other assumptions about the distributions of the variables were made.

The same computational procedures used in the previous example were used in this example. Because the data set contained only 50 cases, a different method of cross-validation was necessary. The 50 states were randomly assigned to five groups of ten states each. Estimation procedures were performed on 40 states (four groups) and then validated on the remaining ten states (fifth group). This was done five times with a different group as the validation set each time. The results of the five analyses appear in Table 2. The new data and the sets of excluded states are presented in Table 3. For the training sets the mean correct classification rate was 68.00 percent for discriminant analysis and 80.00 percent for logistic regression. For the validation sets the respective mean correct classification rates were 68.00 and 72.00 percent. The mean execution times were 1.50 seconds for discriminant analysis and 2.37 seconds for logistic regression.

As might be expected, the states with increasing population had higher birth rates and lower death rates than the states with decreasing (relative to median increase) population. Increasing population was also associated with higher per capita income, a nonurban (less than 70 percent urban) environment, and the presence of a coastline. People were choosing the good life—higher wages and nice surroundings.

Again computation time was longer for logistic regression, but it provided better discrimination for both the training set and the validation set. The coefficients estimated for the two types of analyses are presented in Table 4.

The pattern of misclassifications in this example was somewhat different from that in Example 1. Combining all cases we found twelve cases incorrectly classified by

3. Raw Data for Example 2

State	*Population change*	*Income*	*Births*	*Coast*	*Urban*	*Deaths*
a. Set I						
Arkansas	0	2.878	1.8	0	0	1.1
Colorado	1	3.855	1.9	0	1	.8
Delaware	1	4.524	1.9	1	1	.9
Georgia	1	3.354	2.1	1	0	.9
Idaho	0	3.290	1.9	0	0	.8
Iowa	0	3.751	1.7	0	0	1.0
Mississippi	0	2.626	2.2	1	0	1.0
New Jersey	1	4.701	1.6	1	1	.9
Vermont	1	3.468	1.8	0	0	1.0
Washington	1	4.053	1.8	1	1	.9
b. Set II						
Kentucky	0	3.112	1.9	0	0	1.0
Louisiana	1	3.090	2.7	1	0	1.3
Minnesota	1	3.859	1.8	0	0	.9
New Hampshire	1	3.737	1.7	1	0	1.0
North Dakota	0	3.086	1.9	0	0	.9
Ohio	0	4.020	1.9	0	1	1.0
Oklahoma	0	3.387	1.7	0	0	1.0
Rhode Island	0	3.959	1.7	1	1	1.0
South Carolina	0	2.990	2.0	1	0	.9
West Virginia	0	3.061	1.7	0	0	1.2
c. Set III						
Connecticut	1	4.917	1.6	1	1	.8
Maine	0	3.302	1.8	1	0	1.1
Maryland	1	4.309	1.5	1	1	.8
Massachusetts	0	4.340	1.7	1	1	1.0
Michigan	1	4.180	1.9	0	1	.9
Missouri	0	3.781	1.8	0	1	1.1
Oregon	1	3.719	1.7	1	0	.9
Pennsylvania	0	3.971	1.6	1	1	1.1
Texas	1	3.606	2.0	1	1	.8
Utah	1	3.227	2.6	0	1	.7
d. Set IV						
Alabama	0	2.948	2.0	1	0	1.0
Alaska	1	4.644	2.5	1	0	1.0
Arizona	1	3.665	2.1	0	1	.9
California	1	4.493	1.8	1	1	.8
Florida	1	3.738	1.7	1	1	1.1
Nevada	1	4.563	1.8	0	1	.8
New York	0	4.712	1.7	1	1	1.0
South Dakota	0	3.123	1.7	0	0	2.4
Wisconsin	1	3.812	1.7	0	0	.9
Wyoming	0	3.815	1.9	0	0	.9
e. Set V						
Hawaii	1	4.623	2.2	1	1	.5
Illinois	0	4.507	1.8	0	1	1.0
Indiana	1	3.772	1.9	0	0	.9
Kansas	0	3.853	1.6	0	0	1.0
Montana	0	3.500	1.8	0	0	.9
Nebraska	0	3.789	1.8	0	0	1.1
New Mexico	0	3.077	2.2	0	0	.7
North Carolina	1	3.252	1.9	1	0	.9
Tennessee	0	3.119	1.9	0	0	1.0
Virginia	1	3.712	1.8	1	0	.8

4. Coefficients for Classification Equations of Example 2

Case	Classification Method	Constant	Income	Births	Coast	Urban	Deaths
1	Discriminant analysis	−10.500	+1.610	+3.060	+1.118	−0.360	−1.830
	Logistic regression	−6.238	+1.388	+2.484	+0.874	−0.579	−4.046
2	Discriminant analysis	−7.000	+1.110	+2.240	+0.972	+0.710	−2.240
	Logistic regression	+1.918	+0.580	+0.560	+0.706	+0.249	−5.910
3	Discriminant analysis	−10.700	+1.960	+2.100	+1.482	−0.250	−1.020
	Logistic regression	−6.655	+1.399	+1.894	+0.841	−0.436	−2.428
4	Discriminant analysis	−17.400	+3.550	+5.100	+1.966	−1.890	−5.800
	Logistic regression	−15.162	+3.432	+4.378	+1.391	−1.900	−6.037
5	Discriminant analysis	−13.600	+2.300	+3.610	+0.284	−0.013	−1.870
	Logistic regression	−6.854	+1.542	+2.728	+0.437	−0.452	−4.120
Mean of 5 cases	Discriminant analysis	−11.840	+2.106	+3.222	+1.164	−0.361	−2.552
	Logistic regression	−6.598	+1.668	+2.409	+0.850	−0.634	−4.508

both procedures. Six of 12 cases were positives (increasing population) and six were negatives. Discriminant analysis misclassified four cases that logistic regression classified correctly—two negatives and two positives. Logistic regression misclassified two cases that discriminant analysis had correct—one negative and one positive. Thus, there did not appear to be a bias to negative cases for discriminant analysis for this example.

4. SUMMARY AND CONCLUSIONS

We have presented theoretical arguments for using logistic regression with maximum likelihood estimation compared to using linear discriminant analysis, in both the classification problem and the problem of relating qualitative to explanatory variables. We carried out two empirical studies of nonnormal classification problems, compared the two methods, and found logistic regression with MLE outperforming classical linear discriminant analysis in both cases (supporting the results of Halperin, Blackwelder, and Verter 1971), but not by a large amount. It is unlikely that the two methods will give markedly different results, or yield substantially different linear functions unless there is a large proportion of observations whose x-values lie in regions of the factor space with linear logistic response probabilities near zero or one.

Truett, Cornfield, and Kannel (1967) emphasize that the assumption of multivariate normality is unlikely to be satisfied in applications, even approximately. We thus agree with the conclusion of Halperin, Blackwelder, and Verter (1971) that "use of the maximum likelihood method would be preferable, whenever practical, in situations where the normality assumptions are violated, especially when many of the independent variables are qualitative."

[*Received February 1977. Revised May 1978.*]

REFERENCES

Anderson, J.A. (1972), "Separate Sample Logistic Discrimination," *Biometrika*, 59, 19–35.

Cox, D.R. (1970), *The Analysis of Binary Data*, London: Methuen & Co.

D'Agostino, Ralph B., Pozin, Michael W., Mitchell, Janet, Teebagy, Nicholas C., Guglielmino, Joyce T., Bielawski, Lesley I., and Hood, William B., Jr. (1978), "Comparison of Logistic Regression and Discriminant Analysis as Emergency Room Decision Models for the Diagnosis of Acute Coronary Disease," Boston University Research Report #2-78.

Delury, G.E. (ed.) (1973), *The 1973 World Almanac and Book of Facts*, New York: Newspaper Enterprise Association.

Efron, Bradley (1975), "The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis," *Journal of the American Statistical Association*, 70, 892–898.

Gordon, Tavia (1974), "Hazards in the Use of the Logistic Function with Special Reference to Data from Prospective Cardiovascular Studies," *Journal of Chronic Diseases*, 27, 97–102.

Halperin, Max, Blackwelder, William C., and Verter, Joel I. (1971), "Estimation of the Multivariate Logistic Risk Function: A Comparison of the Discriminant Function and Maximum Likelihood Approaches," *Journal of Chronic Diseases*, 24, 125–158.

McFadden, Daniel (1976), "A Comment on Discriminant Analysis 'Versus' Logit Analysis," *Annals of Economic and Social Measurement*, 5, 511–523.

Nerlove, M., and Press, S. James (1973), "Univariate and Multivariate Log-Linear and Logistic Models," R-1306, Santa Monica, Calif.: The Rand Corporation.

Press, S. James (1972), *Applied Multivariate Analysis*, New York: Holt, Rinehart & Winston.

Rao, C. Radhakrishna (1965), *Linear Statistical Inference and Its Applications*, New York: John Wiley & Sons.

Truett, Jeanne, Cornfield, Jerome, and Kannel, William (1967), "A Multivariate Analysis of the Risk of Coronary Heart Disease in Framingham," *Journal of Chronic Diseases*, 20, 511–524.

Wilson, S.L. (1977), "A Statistical Classification of Breast Cancer Patients by Degree of Nodal Metastases," unpublished M.Sc. thesis, Department of Mathematics, University of British Columbia.