Comparison of Logistic Regression and Discriminant Analyses for Stock Identification of Anadromous Fish, with Application to Striped Bass (*Morone saxatilis*) and American Shad (*Alosa sapidissima*)

Michael H. Prager

Department of Oceanography, Old Dominion University, Norfolk, VA 23529 USA

Mary C. Fabrizio[1]

Institute for Fisheries Research, Ann Arbor, MI 48109 USA

Prager, M. H., and M. C. Fabrizio. 1990. Comparison of logistic regression and discriminant analyses for stock identification of anadromous fish, with application to striped bass (*Morone saxatilis*) and American shad (*Alosa sapidissima*). Can. J. Fish. Aquat. Sci. 47: 1570–1577.

We examined the applicability of logistic regression to stock identification studies and compared its performance on two data sets to that of linear and quadratic discriminant functions. Logistic regression can be used to model a categorical dependent variable associated with continuous or discrete independent variables, and is preferred to discriminant analyses when the explanatory variables are not multivariate normal. Our examples were American shad (*Alosa sapidissima*) from the Connecticut River and Hudson River estuaries, and striped bass (*Morone saxatilis*) from the Hudson River, Chesapeake Bay, and Roanoke River estuaries. In the examples, we used a resampling method to assess classification and allocation errors of the two methods on new data. For the shad data, the logistic model classified significantly more fish correctly, and provided a significantly better estimate of stock composition. For the striped bass data, the two methods classified about the same proportion of fish correctly, but the logistic model gave a significantly better estimate of stock composition.

On examine l'applicabilité d'une analyse de régression logistique aux études d'identification des stocks et on compare les résultats obtenus de deux séries de données à ceux générés par des fonctions discriminantes linéaire et quadratique. L'analyse de régression logistique peut servir à modeliser une variable catégorique dépendante associée à des variables indépendantes continues ou discrètes. On la préfère aux analyses discriminantes quand les variables prédictives n'ont pas une distribution normale multivariée. Comme exemples, les auteurs ont utilisé l'alose savoureuse, *Alosa sapidissima*, des estuaires des rivières Connecticut et Hudson et le bar rayé, *Morone saxatilis*, des estuaires des rivières Hudson et Roanoake et de la baie Chesapeake. Ils se sont servis d'un échantillonnage répété pour évaluer les erreurs de classification et de répartition des nouvelles données. Pour ce qui est des données sur l'alose savoureuse, l'analyse de régression logistique a correctement classé un plus grand nombre de poissons et a fourni une meilleure estimation de la composition du stock. Dans le cas des données sur le bar rayé, les deux méthodes ont correctement classé environ le même nombre de poissons mais l'analyse de régression logistique a donné une estimation nettement supérieure de la composition du stock.

Received October 19, 1989
Accepted March 12, 1990
(JA345)

Reçu le 19 octobre 1989
Accepté le 12 mars 1990

Stock identification, the assignment of fishery resources to unit stocks, is an important component of fish population dynamics (Pitcher and Hart 1982; Cushing 1975). Included in stock identification is stock composition analysis, the classification of samples of fish into a known number of possible stocks of origin. The fisheries literature includes numerous stock composition studies, most enploying some form of discriminant function analysis (e.g. Hill 1959; Berggren and Lieberman 1977; Fabrizio 1987). In this paper, we describe and demonstrate the use of logistic regression (both dichotomous and polytomous) as an alternative to discriminant analysis for classification of fish stocks. The method has desirable statistical properties, and maximum likelihood parameter estimates are readily available from standard statistical software packages, including SAS (SAS 1987) and SYSTAT (Steinberg 1985).

Stock composition analysis is frequently performed by estimating discriminant functions on fish of known origin and subsequently using the functions to assign probabilities of stock membership to fish of unknown origin. Among the assumptions of linear discriminant analysis are that the populations are multivariate normal with different mean vectors of variates, but equal dispersion (variance-covariance) matrices. Under these conditions, linear discriminant analysis is the optimal classifier (Press and Wilson 1978; Lachenbruch 1975). The use of the quadratic discriminant function relaxes the assumption of equal covariances, but not the assumption of multivariate normality (Kendall et al. 1983; Misra 1985). However, quadratic discriminant analysis is more sensitive to sampling variation (Kendall et al. 1983). Other discriminant functions, such as growth-invariant discriminant functions, have been used (Burnaby

[1] Present address: NSI-Environmental Sciences, P.O. Box 12313, Research Triangle Park, NC 27709 USA.

1966; Fabrizio 1987); these also assume multivariate normality. As discriminant analysis has been applied many times and is well accepted in fisheries work, we omit further details here, and refer the reader to Lachenbruch (1975) or a text in multivariate statistics, such as Johnson and Wichern (1988).

Regression analysis on a categorical dependent variable is a different problem from classification, but methods developed for such regressions can be used for classification. Indeed, when discriminant functions are used to predict probabilities of class membership, they can be considered such regressions. When considered apart from the classification problem, regression with a categorical dependent variable and continuous or discrete independent variates is commonly performed by either probit or logit analysis. The underlying probability model for probit analysis is the normal distribution, while that for logit model is the logistic distribution. In practice, the two are so similar as to yield almost identical results; however, the logit model is simpler computationally. While this is of little importance in the dichotomous case, it becomes quite important in the extension to the polytomous case, where the logit model is used almost exclusively (Mantel 1966; Nerlove and Press 1973; Aldrich and Nelson 1984). In this paper, we demonstrate that logistic regression, when estimated by the method of maximum likelihood (ML), provides a useful statistical model for stock composition analysis.

Logistic Model for Classification

The logistic model can be described relatively simply. The following material draws on the work of Nerlove and Press (1973) and Aldrich and Nelson (1984). We begin with the dichotomous model, in which a number of objects (fish) may each belong to one of two classes (stocks). The class membership of fish i is represented by Y_i, a Bernoulli variable (i.e. one which assumes only the values 0 or 1). When the fish belongs to the first stock, $Y_i \equiv 0$; when it belongs to the second, $Y_i \equiv 1$. Also associated with fish i are observations on $K - 1$ continuous or discrete variables: x_{ik}, $k \in \{2, \ldots, K\}$. These variables are characteristics of the fish, or features, thought to be indicative of stock membership. Examples of suitable features are morphometric, meristic, scale-shape, and protein-electrophoretic data. A constant term indexed by $k = 1$ is usually included in the model by defining $X_{i1} = 1$ for all i.

Under the dichotomous logistic model, the probability that fish i belongs to the second stock (i.e. the stock indicated by $Y_i = 1$) is estimated by

$$\text{(1a)} \quad \hat{P}_i \equiv \hat{P}(Y_i = 1) = \frac{\exp(Z_i)}{1 + \exp(Z_i)},$$

where

$$\text{(1b)} \quad Z_i \equiv \sum_{k=1}^{K} b_k X_{ik}.$$

The b_k are the K parameters of the model, one for each feature and one for the constant term. The values of the b_k are estimated by maximum likelihood, as described, e.g. in Nerlove and Press (1973).

Motivation for the form of the logistic model can be obtained by considering a simpler model, the linear probability model:

$$\text{(2)} \quad \hat{P}_i = Z_i.$$

This simple model can be fit by ordinary least squares, but has several undesirable properties. Most importantly, the left-hand side of equation (2) must lie between 0 and 1, while the right-hand side is unbounded. This anomaly in itself makes equation (2) a poor probability model. In addition, it causes the residuals of the model to be heteroscedastic; their variances change in a systematic way with $\hat{P}_i$, making parameter estimates quite sensitive to the distribution of the values in a particular sample.

How might we modify equation (2) to reconcile the bounded left-hand side with the unbounded right-hand side? The upper bound on the left-hand side can be eliminated by modeling the ratio of P_i to $(1 - P_i)$, rather than modeling P_i itself; the ratio will range from zero to infinity. The lower bound can be eliminated by taking the logarithm of the ratio, which results in the model.

$$\text{(3)} \quad \log \frac{\hat{P}_i}{1 - \hat{P}_i} = Z_i.$$

Each side of equation (3) is unbounded. When the equation is solved for $\hat{P}_i$, the result is the logistic model of equation (1a), in which each side is bounded by 0 and 1 (Fig. 1). Use of the logistic rather than the linear model for probabilities generally results in a better and more robust fit (Nerlove and Press 1973). The logistic model is empirically useful and mathematically tractable; it has also been justified on theoretical grounds by Truett et al. (1967).

Polytomous Logistic Regression

The logistic model can be readily extended from the dichotomous (two-stock) case to the polytomous (many-stock) case (Mantel 1966; Nerlove and Press 1973). In the extension to J stocks, the stock membership of fish i is again represented by Y_i, but here $Y_i \in$ "is a member of" $\{1, 2, \ldots, J\}$. (Note that 0 is not a possible value.) To develop the model, let $P(Y_i = j)$, the probability that fish i belongs to the j^{th} stock, be denoted P_{ij}. Taking stock J as a baseline, we can write a model, analogous to equation (3), of the ratio of the probabililty of membership in stock j, $j \in \{1, 2, 3, \ldots, J - 1\}$, to the probability of membership in stock J:

$$\text{(4a)} \quad \log (\hat{P}_{ij} / \hat{P}_{iJ}) = Z_{ij},$$

where

$$\text{(4b)} \quad Z_{ij} \equiv \sum_{k=1}^{K} b_{jk} X_{ik},$$

These are straightforward extensions of equations (3) and (1b), but here the parameters b_{jk} carry an additional subscript denoting stock membership. Solving equation (4a) for $\hat{P}_{ij}$, we get

$$\text{(5)} \quad \hat{P}_{ij} = \exp(Z_{ij}) / \hat{P}_{iJ}$$

We find the denominator, $\hat{P}_{iJ}$, by noting that $\sum^{J} P_{ij}$ must equal 1. This leads to

$$\text{(6)} \quad \hat{P}_{iJ} = \frac{1}{1 + \sum_{j=1}^{J-1} \exp(Z_{ij})}$$

The polytomous logistic model defined by equations (4b), (5), and (6) can be expressed in a simpler form, due to Mantel

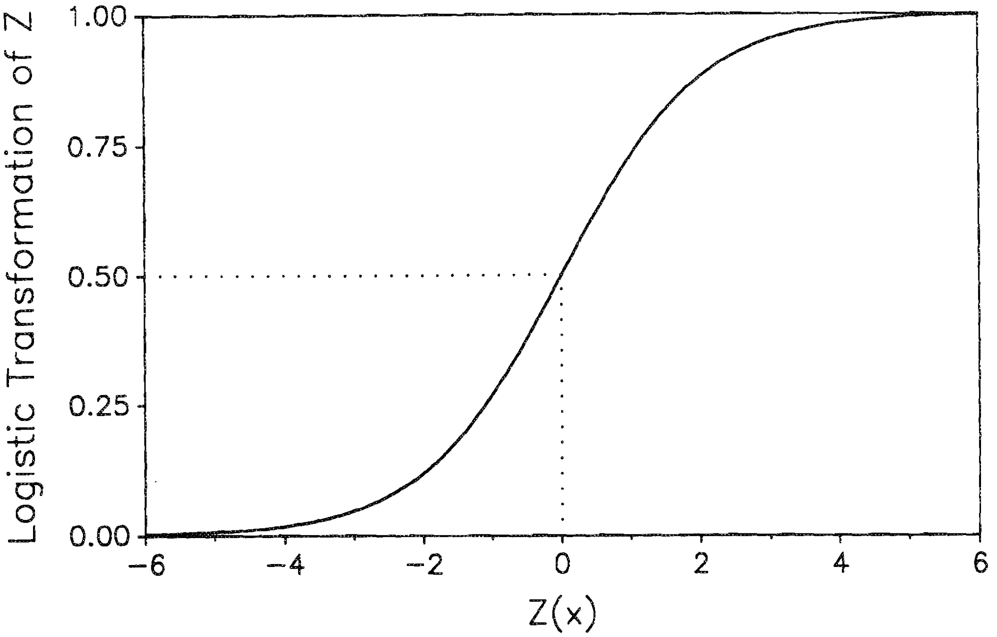

FIG. 1. Illustration of the logistic transformation, equation (1a). Although the variable Z on the horizontal axis is allowed to range from $-\infty$ to $+\infty$, the transformation (on the vertical axis) will remain between zero and one. In practice, Z is a linear combination of the feature (predictor) variables. The dotted lines illustrate the symmetry of the distribution around the point (0, 0.5).

(1966). Equation (4b) includes $K(J-1)$ parameters, where K is the number of features (including the intercept) and J is the number of stocks. (Parameter b_{jk} relates the probability of membership in stock j to the value of X_{ik}.) The simplified notation includes K additional parameters ($b_{J1}, b_{J2}, \ldots, b_{JK}$), each defined to be zero. The model can then be expressed by the single equation

$$(7)\quad \hat{P}_{ij} = \frac{\exp(Z_{ij})}{\sum_{j=1}^{J} \exp(Z_{ij})},$$

for $j \in \{1, 2, \ldots, J\}$. As in the dichotomous case, the parameters of the polytomous logistic model are readily estimated by maximum likelihood.

Comparison to Discriminant Function Analysis

In a comparison of dichotomous logistic regression to discriminant analysis, Press and Wilson (1978) enumerated arguments favoring logistic regression for classification studies. We summarize six of their arguments here: (1) When the explanatory variables for each stock are not multivariate normal, discriminant function estimators are not consistent. Thus, the-prediction accuracy of a discriminant function with such data will not necessarily improve with increasing sample sizes. The logistic model, when estimated by maximum likelihood, provides consistent estimates. (2) When the normality assumption is not met, discriminant function estimation (including stepwise discriminant procedures) can give misleading information about the significance of coefficients. Meaningless variables may be included by the discriminant estimator (even with large sample sizes), but not by the ML estimator of the logistic model. (3) Halperin et al. (1971) conducted numerical comparisons of ML logistic analysis and discriminant function estimation. They found that, under non-normal conditions, the ML logistic estimator usually gives slightly better fit than the discriminant function. (4) Maximum likelihood estimation is more likely to provide warning (by failing to converge) when the model is poorly determined by the data. (5) The maximum likelihood estimators are functions of sufficient statistics, while the discriminant estimators are not. The use of sufficient statistics is associated with lower mean squared errors. (6) Discriminant analysis may produce biased estimators in some applications, in particular where the underlying process is logistic (McFadden 1976).

Applications to Data

To compare the results of logistic regression to those of discriminant analysis, we analyzed two sets of data on fish of known stocks of origin. These analyses were not intended to be definitive characterizations of the two methods, but rather to give some indication of their comparative merit on actual fisheries data. We analyzed each data set by both classification methods, and computed statistics to describe each method's accuracy in classification and allocation. The Wilcoxon signed-ranks test (a nonparametric statistical test of paired differences; Hollander and Wolfe 1973) was used to compare the two methods. The two data sets, shad and striped bass, were analyzed by identical methods, but completely separately.

Data

The first data set used for testing was collected by Hill (1959), and contained information on 310 American shad, *Alosa sapidissima*, from the Connecticut and Hudson Rivers. We used the 300 observations without missing values (Table 1). Hill's data include counts of the posterior scutes, vertebrae, pectoral rays, anal rays, and dorsal rays of each fish. Hill chose these five characters for his own analyses because he found that these features did not vary with size. We used these data to demonstrate application of the dichotomous logistic model.

TABLE 1. Summary of data used in the study. Striped bass data and collection methods are described in Fabrizio (1987); American shad data and collection methods, in Hill (1959).

Species	Number of fish	Stock of origin	Year collected
American shad	91	Connecticut River	1945
American shad	104	Hudson River	1939
American shad	105	Hudson River	1940
Total	300		
Striped Bass	86	Chesapeake Bay	1984
Striped Bass	28	Chesapeake Bay	1985
Striped Bass	34	Hudson River	1984
Striped Bass	47	Hudson River	1985
Striped Bass	98	Hudson River	1986
Striped Bass	143	Roanoke River	1984
Striped Bass	97	Roanoke River	1985
Striped Bass	92	Roanoke River	1986
Total	625		

The second data set was compiled by Fabrizio (1987) from 644 striped bass, *Morone saxatilis*, of the Hudson River, Chesapeake Bay, and the Roanoke River stocks. We omitted data on 19 fish as being apparent outliers, leaving 625 observations (Table 1). The data collected from each fish were fork length, sex, head length, jaw length, snout length, and internostril width; in addition, six variables were derived from isoelectric focussing analysis of eye lens proteins. Fabrizio (1987) subjected these six variables to the arcsine transformation prior to further analysis; we used the transformed variables.

Two difficulties arose in connection with the morphometric variables on striped bass. First, fork length was highly correlated (r>0.9) with each of the four other morphometric variables. Second, mean fork length differed strongly by stock of origin. This might have made it useful in classification, but because the relationship of fork length to stock of origin can be quite variable, we wished to consider only models not including body length effects. To resolve these issues, we regressed each morphometric variable against fork length, and performed a principal component analysis on the residuals. The first three principal components, which contained approximately 90% of the remaining variance, were used as features in further analyses. This procedure removed length effects, eliminated correlation among variables, and, by omitting the final principal component, allowed us to discard the portion of the variance least correlated among variables, a portion that we believed likely to contain much of the measurement error. We classified the striped bass data into the three component stocks to illustrate application of the polytomous logistic model.

Misclassification Statistics

We used two statistics to summarize the results of each classification method on each data set. The first statistic estimated the misclassification rate of each method. Lachenbruch and Mickey (1968) described and evaluated eight methods of estimating misclassification rates of discriminant functions. Three common methods that do not depend upon the assumption of normality, and thus are suitable for evaluating classification rules for non-normal data, are (1) the resubstitution method, which computes the error rate on the data used for fitting the model (2) the holdout method, which computes the error rate on a subsample of data withheld from model development; and (3) the jackknife method, in which each point is held out in turn and then classified by the rule developed on all other points.

Each of the three methods has its advantages and disadvantages. The resubstitution method yields a statistic known as the apparent error rate (Lachenbruch 1975) which, while simple to compute, tends to underestimate the error rate on new data, even when the populations are normal (Kendall et al. 1983). The holdout procedure, although not subject to the large optimistic bias of the resubstitution method, has three significant disadvantages. First, it requires a relatively large sample. Second, it evaluates a different model from the one that will be used in practice, because before the model is used in practice, it is usually recomputed on the entire data set. Finally, choosing the size of the holdout sample is problematic: a large holdout sample gives a good estimate of the model's performance, but that performance may be poor. A small holdout sample provides a better model, but a highly variable estimate of its performance.

Lachenbruch and Mickey (1968) found the jackknife to be the best of the methods not assuming normality. However, the jackknife is computationally intensive, a property which prohibited its use in the present study, given our use of an iterative maximum-likelihood procedure for estimation of the logistic models. Since the iterative solution was slow, computing jackknifed estimates of error rates would have been impractical.

To provide reasonably accurate and precise estimates of error rates, we developed a hybrid method combining elements of the holdout and jackknife methods. In the holdout method, the data are partitioned randomly into what might be called a training subset and a verification subset. In our hybrid procedure, we randomly partitioned the data not once, but 25 times. In making each partition, we allocated about 70% of the data to the training subset and the rest to the verification subset. Each of the 25 training subsets was then used to fit both a discriminant function and a logistic regression model. The parameters of each model were used to classify the observations in the corresponding verification subset.

After each of the 25 verification subsets had been classified both by logistic regression and by discriminant analysis, we evaluated the success of each classification. A measure of a model's misclassification rate on stock j is provided by the statistic E_j:

$$(8) \quad E_j = |N_j - \hat{N}_j| \, / \, N_j,$$

where N_j is the number of fish from stock j in the verification set and $\hat{N}_j$ is the number of those fish classified into the correct stock by the model in question. For comparative purposes, we used the total misclassification rate statistic E, a weighted mean of the E_j's:

$$(9) \quad E = \frac{\sum_{j=1}^{J} N_j E_j}{\sum_{j=1}^{J} N_j}.$$

The E statistic, then, provides an estimate of the expected misclassification rate of a model on new data.

Misallocation Statistics

The misclassification rate is not the only, and perhaps not the most appropriate, measure of error in stock composition analysis. Although the goal of such an analysis is to assign each fish to its correct stock, this is rarely achieved. In the absence of a perfect classification, one may be more interested in a different measure of error: the misallocation of the sample in relation to its actual stock composition. In many fishery contexts, the accuracy of the allocation (the proportion assigned to each stock) will be of more interest than the proportion of individual fish classified correctly. To quantify misallocation error, we propose the new statistic C, which is the mean of the allocation error statistics (C_j, defined below) on fish from the J individual stocks:

$$\text{(10a)} \quad C = \frac{1}{J}\sum_{j=1}^{J} C_j ,$$

The allocation error of the model on fish from stock j only is represented by C_j:

$$\text{(10b)} \quad C_j = \frac{|(P_j - \hat{P}_j)|^{1.5}}{d_j} .$$

Here P_j is the true proportion of the total sample that comes from stock j, $\hat{P}_j$ is the proportion of the total sample classified into stock j by the algorithm under consideration, and d_j is a normalization factor that limits C_j to values between 0 and 1:

$$\text{(10c)} \quad d_j = \begin{cases} 1, & \text{for } \mathrm{P}_j = 1 \text{ or } P_j = 0, \\ 1 - \mathrm{P}_j, & \text{for } \hat{P}_j \geqslant \mathrm{P}_j, \\ \mathrm{P}_j, & \text{for } \hat{P}_j < \mathrm{P}_j. \end{cases}$$

The numerator of equation (10b) is raised to the 1.5 power to increase the penalty at a greater-than-linear rate as the error between estimated and observed proportions becomes larger. The value of this exponent was chosen heuristically, but the resulting statistic seems to provide reasonable penalties for different values of P_j and $\hat{P}_j$ (Fig. 2). Since this is so, we believe that C is a useful statistic for comparing estimated allocations.

How does the misallocation rate C compare to the misclassification rate E? When all fish are classified correctly, E and C are each zero. When some fish are misclassified, C conveys different information from E. To illustrate this point, consider a hypothetical case in which 200 fish with a true stock composition of 3:1 (150:50) are classified. A value of $E = 0.25$, reflecting an overall misclassification rate of 25% (50 fish), might be the best obtainable in a particular case. If the model misclassified exactly 25 fish from each stock (which would give $E = 0.25$), the estimate of stock composition would nonetheless be exactly correct (which would give $C = 0$). However, a value of $E = 0.25$ would also occur if all 50 fish of the smaller stock were misclassified. This would be evidence of a poor model, and the value of $C = 0.5$ would reflect that.

The values of $\hat{P}_j$ for computing C can be obtained by at least two methods. The methods differ in the computation of $\hat{N}_j$, the number of fish assigned to each stock. In the first method, $\hat{N}_j$ is the sum of the predicted probabilities that each fish belongs to stock j. In the second method, each fish is assigned to the stock for which its predicted probability of membership is highest. This method must be used in computing the misclassification statistic E, as for that purpose each fish must be classified into one stock or the other. In computing the misallocation statistic C, however, one can use the first method, which distinguishes a very certain classification (where one of the predicted probabilities approaches unity) from an uncertain one (where the predicted probabilities are nearly equal). In this paper we used the first method to compute C.

We used standard software for most of the analyses. The parameters of the logistic model for each of the 25 partitions of each species were estimated with the CATMOD procedure of SAS (SAS 1987). To classify the corresponding verification set for each partition, we used a short computer program (written in FORTRAN) to implement equation (7) and provide summary statistics. The discriminant functions were computed by

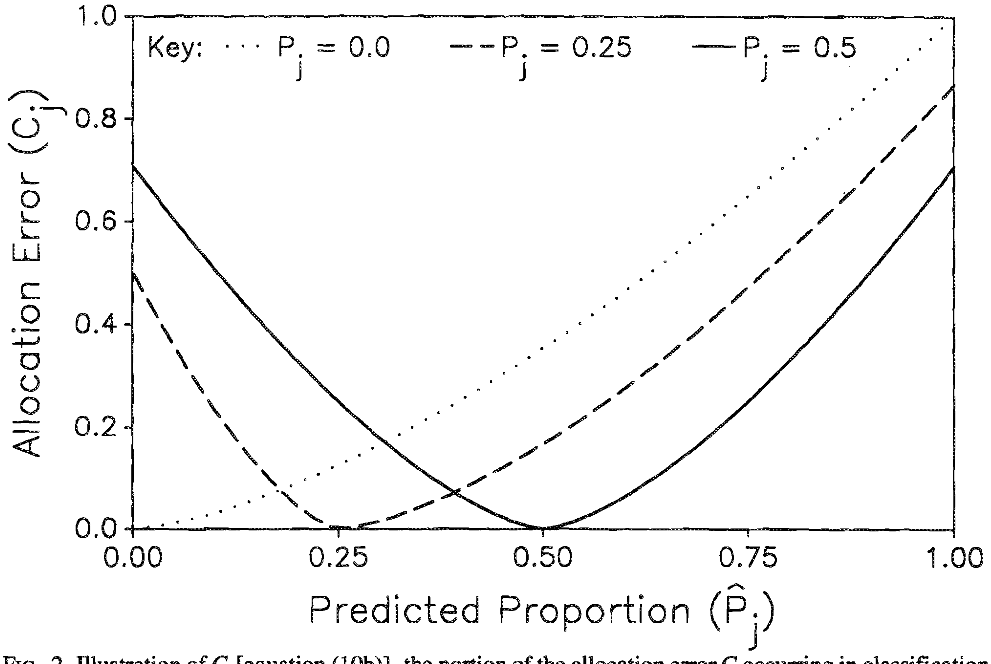

FIG. 2. Illustration of C_j [equation (10b)], the portion of the allocation error C occurring in classification of fish from stock j. The three curves represent values of C_j over all $\hat{P}_j$ for the three true values $P_j = 0.0$, $P_j = 0.25$, and $P_j = 0.5$. The curve for any $P_j = y$ ($0 \leqslant y \leqslant 1$) is a mirror image of the curve for $P_j = 1 - y$.

TABLE 2. Comparison of discriminant analysis and logistic regression results in classifying American shad and striped bass. An asterisk (*) indicates statistically significant differences in misallocation (C statistic) or misclassification (E statistic) at $\alpha = 0.05$ by Wilcoxon's signed-rank test.

Error statistic	Mean for logistic regression ($N = 25$)	Mean for discriminant analysis ($N = 25$)	Number of cases in which LR classifier was superior	Probability that methods are equally good classifiers
American shad				
C	0.0213	0.395	19	0.015*
E	0.1496	0.1704	19[a]	0.003*
Striped bass				
C	0.0106	0.0218	20	< 0.002*
E	0.2500	0.2401	8[b]	0.121

[a] In addition, two pairs were ties.
[b] In addition, three pairs were ties.

the DISCRIM procedure of SAS (SAS 1987). The American shad data gave no indication of unequal dispersion matrices between populations, and therefore a linear discriminant function was used. For the striped bass, the data in each partition were tested for homogeneity of the dispersion matrices. In each case, the null hypothesis (of homogeneity) was rejected, and therefore a quadratic discriminant function was used.

We compared the results of the two classification methods by forming paired differences of the E and C statistics for each partition. Within each pair, both values were of either E or C; one described the logistic analysis; the other, the discriminant analysis of the same data. The sample size of 25 for each test reflected the 25 partitions of each data set. We tested at $\alpha = 0.05$ the null hypothesis that the paired differences were not different from zero. To avoid the assumptions of parametric testing, we used the Wilcoxon's signed-rank test (Hollander and Wolfe 1973).

Results

The results of the empirical study indicated that, for these data, logistic regression was a better classifier than discriminant analysis (Table 2). Logistic regression provided both lower allocation errors (C statistic; $P = 0.015$) and lower classification errors (E statistic; $P = 0.003$) on the American shad data. On the striped bass data, logistic regression provided significantly lower allocation errors ($P = 0.002$), while the two methods were not significantly different in classification error ($P = 0.12$).

The superiority of the logistic regression classifiers is also visible in the distribution of errors of each type. Figures 3 and 4 illustrate the distribution of allocation errors and classification errors on the American shad data, while Fig. 5 and 6 illustrate the same information on the striped bass data. Most of the largest errors were from discriminant analyses, and the logistic analyses had consistently lower errors.

Discussion

In recent years, several new laboratory techniques have been described for fish stock identification, but discriminant analysis has remained the standard statistical tool for stock composition analysis. However, discriminant analysis assumes normality and is not robust to violations of that assumption (Eisenbeis

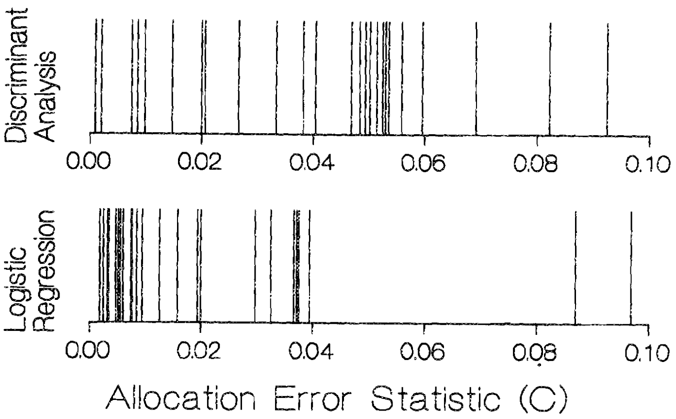

FIG. 3. Values of the allocation error statistic C observed when classifying American shad by two statistical methods. Upper stripe plot: values of C from 25 discriminant function analyses; lower plot: values of C from 25 logistic regression analyses. Each vertical line in the stripe plots represents one (or more, if equal) observed value of C. The vertical scale carries no information. Stripe plots provide the same information as histograms, but are used here because they illustrate the range of the data more distinctly.

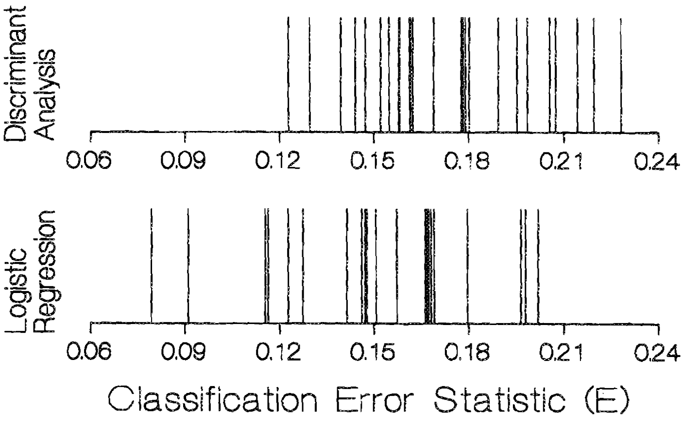

FIG. 4. Distribution of the classification error statistic E when classifying American shad by two statistical methods. Upper stripe plot: values of E from 25 discriminant function analyses; lower plot: values of E from 25 logistic regression analyses. (The caption to Fig. 3 contains more information on stripe plots.)

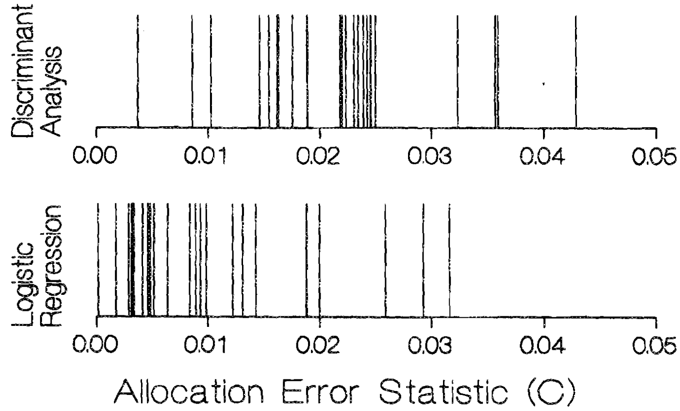

FIG. 5. Distribution of the allocation error statistic C when classifying striped bass by two statistical methods. Upper stripe plot: values of C from 25 logistic regression analyses. (The caption to Fig. 3 contains more information on stripe plots.)

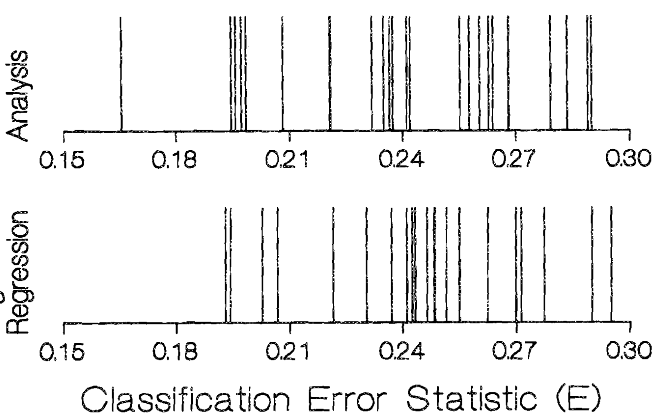

FIG. 6. Distribution of the classification error statistic E when classifying striped bass by two statistical methods. Upper stripe plot: values of E from 25 discriminant function analyses; lower plot: values of E from 25 logistic regression analyses. (The caption to Fig. 3 contains more information on stripe plots.)

1977; Krzanowski 1977). It would seem that evaluation of other methods is necessary, or at least desirable.

The logistic model described here is strictly appropriate (maximum-likelihood) for the following kinds of data (Kendall et al. 1983): (1) normal variables with equal covariances;(2) independent binary variables; (3) binary variables following the linear logistic model with second- and higher-order effects equal; and (4) a combination of (1) and (3). However, the empirical results of our study and the observations of Press and Wilson (1978) and others (e.g. Krzanowski 1975, 1977; Halperin et al. 1971) suggest that logistic regression is relatively robust to the distribution of the underlying data. Thus we would expect it to be at least as useful as, if not more useful than, discriminant analysis in many cases.

Efron (1975) demonstrated that with true multivariate normal data and large separation between populations, the asymptotic efficiency of the logistic model is less than that of discriminant analysis. Some biological variates are approximately normal; others approximate normality after logarithmic transformation. However, many data used in stock composition studies do not appear to follow any distribution strictly, and certainly many such data, such as counts and proportions, are not multivariate normal. Moreover, there is no conclusive test for multivariate normality (Johnson and Wichern 1988). In this context, our conclusion from Efron's (1975) finding is that discriminant analysis should be included among the methods of analysis when the data appear to be multivariate normal.

Beyond the method discussed here, we are aware of two recent papers that have described novel statistical methods for stock composition analysis. The first method was developed by Fournier et al. (1984), who called it FMP-LS (finite mixture problem with learning sample). That procedure uses conditional maximum likelihoods and requires assuming a specific probability density function for each feature variable. The second method, a neutral network variant called GMDH (group method of data handling), was presented by Prager (1988) as an adaptation of methods developed by Ivakhnenko et al. (1979). While either or both of these methods may prove valuable, they appear to share two drawbacks at present. First, each requires relatively complex computations for which computer programs are not readily available. Second, the statistical properties of these two methods have not been well studied. In particular, the performance of the empirical GMDH procedure of Prager (1988) will require further Monte Carlo studies to establish.

In comparison to these two methods, logistic regression appears relatively attractive. The computations for logistic regression can be performed by several widely used statistical software packages. Only the assignment of new individuals to stocks may require programming, and that computation is simple enough to be done on a hand calculator if necessary. Furthermore, since logistic regression has been the subject of many studies (e.g. Anderson 1972; Gordon 1974, Efron 1975), the statistical properties of the method are relatively well known.

While the empirical results reported here do not prove that logistic regression will be superior to discriminant analysis in any given case, they indicate that it has that potential. At this stage of our knowledge, it would be premature to suggest that fishery biologists abandon the well-known tool of discriminant analysis. We do suggest, however, that those performing stock composition analyses should evaluate more than one statistical method, and that the logistic regression model described here is worthy of inclusion.

Acknowledgments

This research was funded by an Old Dominion University Faculty Research Fellowship to M. H. Prager. Final revisions were made while a visiting fellow at the Cooperative Institute for Marine and Atmospheric Studies, University of Miami, under NOAA Cooperative Agreement NA85-WCH-06134; and were also supported in part by a grant from the Virginia Sea Grant program. We thank Pierre Jolicoeur and an anonymous reviewer for their helpful remarks. John Hoenig provided valuable comments on an earlier draft of the manuscript.

References

ALDRICH, J. H., AND F. D. NELSON. 1984. Linear probability, logit, and probit models. Sage Univ. Paper series on Quantitative Applications in the Social Sciences, 07–045. Sage Publications, Beverly Hills, CA.

ANDERSON, J. A. 1972. Separate sample logistic discrimination. Biometrika 59: 19–35.

BERGGREN, T. J., AND J. T. LIEBERMAN. 1977. Relative contribution of Hudson, Chesapeake, and Roanoke striped bass, *Morone saxatilis*, stocks to the Atlantic coast fishery. U.S. Fish. Bull. 76: 335–345.

BURNABY, T. P. 1966. Growth-invariant discriminant functions and generalized distances. Biometrics 22: 96–110.

CUSHING, D. C. 1975. Marine ecology and fisheries. Cambridge University Press, Cambridge, UK. 278 p.

EFRON, B. 1975. The efficiency of logistic regression compared to normal discriminant analysis. J. Am. Stat. Assoc. 70: 892–898.

EISENBEIS, R. A. 1977. Pitfalls in the application of discriminant analysis in business, finance, and economics. J. Finance 32: 875–900.

FABRIZIO, M. C. 1987. Growth-invariant discrimination and classification of striped bass stocks by morphometric and electrophoretic methods. Trans. Am. Fish. Soc. 116: 728–736.

FOURNIER, D. A., T. D. BEACHAM, B. E. RIDDELL, AND C. A. BUSACK. 1984. Estimating stock composition in mixed stock fisheries using morphometric, meristic, and electrophoretic characteristics. Can. J. Fish. Aquat. Sci. 41: 400–408.

GORDON, T. 1974. Hazards in the use of the logistic function with special reference to data from prospective cardiovascular studies. J. Chron. Dis. 27: 97–102.

HALPERIN, M., W. C. BLACKWELDER, AND J. I. VERTER. 1971. Estimation of the multivariate logistic risk function: a comparison of the discriminant function and maximum likelihood approaches. J. Chronic Dis. 24: 125–158.

HILL, D. R. 1959. Some uses of statistical analysis in classifying races of American shad (*Alosa sapidissima*). U.S. Fish. Bull. 59: 269–286.

HOLLANDER, M., AND D. A. WOLFE. 1973. Nonparametric statistical methods. John Wiley and Sons, New York, NY. 503 p.

IVAKHNENKO, A. G., G. I. KROTOV, AND V. N. VISOTSKY. 1979. Identification of the mathematical model of a complex system by the self-organization method, p. 324–352. *In* E. F. Halfon [ed.] Theoretical systems ecology. Academic Press, New York, NY.

Johnson, R. A., and D. W. Wichern. 1988. Applied multivariate statistical analysis, 2nd ed. Prentice Hall, Englewood Cliffs, NJ. 607 p.

Kendall, M., A. Stuart, and J. K. Ord. 1983. The advanced theory of statistics, vol. 3, 4th ed. Charles Griffin, London, UK. 780 p.

Krzanowski, W. J. 1975. Discrimination and classification using both binary and continuous variables. J. Am. Stat. Assoc. 70: 782–790.

1977. The performance of Fisher's linear discriminant function under non-optimal conditions. Technometrics 19: 191–200.

Lachenbruch, P. A. 1975. Discriminant analysis. Hafner Press, New York, NY. 128 p.

Lachenbruch, P. A., and M. R. Mickey. 1968. Estimation of error rates in discriminant analysis. Technometrics 10: 1–11.

Mantel, N. 1966. Models for complex contingency tables and polychotomous dosage response curves. Biometrics 22: 83–95.

McFadden, D. 1976. A comment on discriminant analysis "versus" logit analysis. Ann. Econ. Social Meas. 5: 511–523.

Misra, R. K. 1985. Quadratic discriminant analysis with covariance for stock delineation and population differentiation: a study of beaked redfishes (*Sebastes mentella* and *S. fasciatus*). Can. J. Fish. Aquat. Sci. 42: 1672–1676.

Nerlove, M., and S. J. Press. 1973. Univariate and multivariate log-linear and logistic models. Report R-1306-EDA/NIH. Rand Corporation, Santa Monica, CA. 134 p.

Pitcher, T. J., and P. J. B. Hart. 1982. Fisheries ecology. Croom Helm, London, UK. 414 p.

Prager, M. H. 1988. Group method of data handling: a new method for stock identification. Trans. Am. Fish. Soc. 117: 290–296.

Press, S. J., and S. Wilson. 1978. Choosing between logistic regression and discriminant analysis. J. Am. Stat. Assoc. 73: 699–705.

SAS Institute Inc. 1987. SAS/STAT guide for personal computers, vers. 6 ed. SAS Institute, Cary, NC. 1028 p.

Steinberg, D. 1985. Logit: a supplementary module for SYSTAT. Supplied with software purchase from SYSTAT, Inc., Evanston, IL. 60201.

Truett, J., J. Cornfield, and W. Kannel. 1967. A multivariate analysis of the risk of coronary heart disease in Framingham. J. Chron. Dis. 20: 511–524.

This article has been cited by:

1. Peter D. Chase. Meristics 171-184. [Crossref]
2. RamsayAlice L., MilnerNigel J., HughesRoger N., McCarthyIan D.. 2011. Comparison of the performance of scale and otolith microchemistry as fisheries research tools in a small upland catchment. *Canadian Journal of Fisheries and Aquatic Sciences* **68**:5, 823-833. [Abstract] [Full Text] [PDF] [PDF Plus] [Supplemental Material]
3. Michael H. Prager, Kyle W. Shertzer. An Introduction to Statistical Algorithms Useful in Stock Composition Analysis 499-516. [Crossref]
4. Mary C. Fabrizio. Experimental Design and Sampling Strategies for Mixed-Stock Analysis 467-498. [Crossref]
5. Jerome Pella, Michele Masuda. Classical Discriminant Analysis, Classification of Individuals, and Source Population Composition of Mixtures 517-552. [Crossref]
6. John R. Waldman. Meristics 197-207. [Crossref]
7. Silvia Salas, Ussif Rashid Sumaila, Tony Pitcher. 2004. Short-term decisions of small-scale fishers selecting alternative target species: a choice model. *Canadian Journal of Fisheries and Aquatic Sciences* **61**:3, 374-383. [Abstract] [PDF] [PDF Plus]
8. A. P. Wheeler, M. S. Allen. 2002. Comparison of Three Statistical Procedures for Classifying the Presence-Absence of an Aquatic Macrophyte from Microhaibitat Observations. *Journal of Freshwater Ecology* **17**:4, 601-608. [Crossref]
9. Shijie Zhou. 2002. Size-Dependent Recovery of Chinook Salmon in Carcass Surveys. *Transactions of the American Fisheries Society* **131**:6, 1194-1202. [Crossref]
10. John R. Waldman. 1999. The importance of comparative studies in stock analysis. *Fisheries Research* **43**:1-3, 237-246. [Crossref]
11. BRENDA L. NORCROSS, ARNY BLANCHARD, BRENDA A. HOLLADAY. 1999. Comparison of models for defining nearshore flatfish nursery areas in Alaskan waters. *Fisheries Oceanography* **8**:1, 50-67. [Crossref]
12. Eric B. Taylor, Terry D. Beacham, Masahide Kaeriyama. 1994. Population Structure and Identification of North Pacific Ocean Chum Salmon (Oncorhynchus keta) Revealed by an Analysis of Minisateliite DNA Variation. *Canadian Journal of Fisheries and Aquatic Sciences* **51**:6, 1430-1442. [Crossref]
13. Martyn Norman Futter. 1994. Pelagic food-web structure influences probability of mercury contamination in lake trout (Salvelinus namaycush). *Science of The Total Environment* **145**:1-2, 7-12. [Crossref]