MULTIPLE LINEAR REGRESSION VIEWPOINTS
VOLUME 18, NUMBER 1, FALL 1991

Some Parallels Between Predictive Discriminant Analysis and Multiple Regression

John D. Morris
Florida Atlantic University

Carl J. Huberty
University of Georgia

Some Parallels Between Predictive Discriminant Analysis and Multiple Regression

The purpose of this paper is to outline some important similarities in, and differences between, predictive discriminant analysis (DA) and multiple regression (MR). The areas covered, chosen for their importance and need for clarification, are estimates of model accuracy, hypothesis testing, and non-least squares models. Some of the parallels are well known, some are less well known, and some appear to have not yet been considered at all.

It is well known that when 1) only two groups are involved, 2) the two population predictor covariance matrices are assumed equal, and 3) the two prior probabilities of group membership are taken to be equal, the popular "minimum chi-square rule" (Tatsuoka, 1971, p. 218) associated with discriminant analysis (DA) is equivalent to predicting a dichotomous criterion variable via multiple regression (MR) methods and classifying a subject into the group for which the predicted criterion is nearer the actual. An especially enlightening examination of this and some other multivariate techniques from the general perspective of MR is provided by Flury and Riedwyl (1985).

However, a precaution about the equivalence of two-group classification and multiple regression with a dichotomous criterion is appropriate. In a two-group situation, there is one linear discriminant function (LDF) and there are two linear classification functions (LCFs); an LDF and an LCF are simply linear composites of the predictors. It is true in a two-group context that the regression weights are proportional to the single set of LDF weights. When a linear regression function (LRF) or an LDF is used for classification purposes a cut-off criterion needs to be determined — with an LRF it is midway between the two values by which the dichotomous criterion is coded, with an LDF it is midway between the LDF means for the two groups. With the use of LCFs, there is no cut-off per se; rather a unit is classified into the group with which is associated the larger LCF score. It turns out that the respective LCF weight differences are proportional to the corresponding LDF and (therefore) the LRF weights.

Input scores for an LRF, an LDF, and an LCF are typically predictor variable measures. [As stated above, any of the three linear composite types may be used for a two-group classification problem.] It turns out that another, still equivalent, approach to two-group classification may be employed. Here, one uses LDF scores for each unit as input for an LCF; we thus have, in essence, a single predictor score for each unit.

When generalizing from a two-group problem to a k-group problem, it is advisable to forget the LRF and LDF approaches and focus on the LCF approach, with predictor measures as input scores.

Estimates of Model Accuracy

Estimation of the cross-validated accuracy of a prediction model offers similarities and differences between MR and DA methods. In both DA and MR the researcher must decide what type of cross-validated accuracy is of concern. For instance, is interest in simply estimating an accuracy index parameter from the associated statistic, that is, estimating the index of accuracy (R^2 or percent of "hits", respectively) that would obtain in the population from that same index in the sample, or is interest in the accuracy that would obtain on application of sample optimized weights to alternate samples from the same population? The concern in this paper will be with the latter type of accuracy.

As in an estimate of cross-validated R^2 in MR, a judgment of DA "hit-rate" based on the calibration sample is optimistically biased in reference to application to alternate samples. To estimate a cross-validated result in MR, another decision that must be made is whether interest is in relative accuracy, as manifested in the correlation of Y and $\hat{Y}$, or in absolute accuracy, as manifested in the MSE. In either case, several formula estimates are available (see Huberty & Mourad, 1980; Rozeboom, 1978). It is probable that in most predictive uses of MR in the behavioral sciences, such as in personnel selection, concern is with relative accuracy.

Unlike in MR, the concern in predictive DA is in classification accuracy; this is implicitly a concern of absolute accuracy. A formula estimate for cross-validated hit-rate in the general k-group case has largely eluded methodologists. However, a useful, although complicated, formula estimate for cross-validated hit-rate in the two-group case was derived by McLachlan (1975). According to that estimator, the hit rate, P_g for group g, where $g = 1$ or 2 is:

$$\begin{aligned}\hat{P}_g = 1 - F(-D/2) - f(-D/2)\{&(p - 1)/(Dn_g)\\ &+ D[4(4p - 1) - D^2]/(32m) + (p - 1)(p - 2)/(4Dn_g^2)\\ &+ (p - 1)[-D^3 + 8D(2p + 1) + 16/D]/(64mn_g)\\ &+ D[3D^6 - 4D^4(24p + 7) + 16D^2(48p^2 - 48p - 53)\\ &+ 192(-8p + 15)]/(12288m^2)\},\end{aligned}$$

where F is the standard normal distribution function, i.e., $F(-D/2)$ is the area to the "left" of $-D/2$, f is the standard normal density function, D is the Mahalanobis distance, p is the number of predictor variables, n_g is the number of subjects in group g, and $m = n_1+n_2-2$. While the formula looks formidable, with patience, it is *calculable* with a hand-held calculator. Moreover, as the last term in the multiplier for $f(-D/2)$ is usually very small, one may choose to ignore it, making the formula even more tractable. If the researcher with an orientation toward MR notes that $D^2 = R^2N(N-2)/(1-R)^2n_1n_2$, then the McLachlan estimator of cross-validated hit-rate can be obtained from the R^2 resulting from regressing the dichotomous criterion on the predictors.

One slightly "unnerving" aspect of the McLachlan estimator is that it can yield estimated hit-rates that are *larger* than those that are estimated from the known positively biased process of reclassifying the calibration sample (Morris & Huberty, 1986; 1987). This is unlike the case in MR where the "shrunken" multiple correlation is necessarily less than the value of the multiple correlation derived from the calibration sample. The explanation for this apparent paradox between methods is that estimators of the cross-validated multiple correlation are functions of the corresponding calibration sample multiple correlation, and are therefore guaranteed to yield smaller values than the sample value. In this sense, the McLachlan hit-rate estimator is not parallel to the MR formula estimators. While it is an estimator of cross-validation hit-rate, it is not a function of the calibration sample generated hit-rate. Rather, it is a function of the Mahalanobis distance between groups, as well as other variables. That is, it does not simply estimate a parameter from a function of the corresponding statistic as do MR formula estimators.

An alternate nonparametric approach to estimating cross-validated hit-rate, which has a wide following in the DA literature, is the "leave-one-out" procedure (Huberty, 1984; Huberty & Mourad, 1980; Lachenbruch & Mickey, 1968; Mosteller & Tukey, 1968). In this method, a subject is classified by applying the rule derived from all Ss except the one being classified. This process is repeated "round-robin" for each subject with a count of the overall classification accuracy used to estimate the cross-validated accuracy.

Clearly the same "round-robin" procedure can be used to estimate either relative or absolute accuracy in the use of MR, and has appeared in that context, with perhaps the earliest reference due to Gollob (1967). In a system intended to select optimal MR predictor variable subsets, Allen (1971) coined the procedure "PRESS," and he appears to be the source most often cited in the MR literature.

The apparent computational difficulties due to the inversion of N matrices can be avoided in both MR and DA by using a matrix identity due to Bartlett (1951). This identity is cited and used explicitly in introducing the technique in the DA context by Lachenbruch and Mickey (1968), but was not mentioned by Allen in the first introduction of PRESS (1971) nor in its presentation in a later text (Allen & Cady, 1982, p. 254), although the same identity was implicitly used. Moreover, Allen doesn't cite the DA literature and the parallel application of the PRESS procedure. It appears that this resampling process was "invented" independently in the MR and DA literatures.

Full vs Restricted Model Hypothesis Testing

A technique that is well known and widely used by MR researchers is that of hypothesis testing through contrasting full and restricted prediction models. The power of this method, its generality, and its applicability to a very wide arena of theoretical questions in science is no doubt part of the reason for the establishment of the MLRSIG within AERA.

The same types of model contrast "explanatory increment" questions can be asked and seem to be at least as much potential interest when the criterion is classification accuracy. However, we know of no examples of this technique being used in the literature. There seems to be no reason not to test the difference in proportion of correct classifications (hit-rate) between full and restricted models to examine meaningful hypotheses, just as is done using the R^2 in MR. The appropriate test statistic is McNemar's (1947) contrast between correlated proportions. Moreover, as the index, "I", of increase in classification accuracy over chance (see Huberty, 1984, p. 168) is distributed similarly, it becomes apparent that such a test would also be applicable to that statistic.

An example of such a test from a study in which the subsequent high school dropout of a sample of 76 children was predicted from data available in fifth grade will now be presented. The six predictor variables were gender, race (two levels), number of elementary schools in which the child had been a student, the number of grades the child had repeated, the family structure (living with at least one natural parent and no other adult, or not), and the child's total number of fifth grade absences. As we have evidence of the relationship between both gender and race and the criterion of high school dropout, the hypothesis to be tested concerned the significance of the increment to classification accuracy afforded by adding the four "non-organismic" variables (number of elementary schools, number of grades repeated, family structure, and the total number of fifth grade absences) to the prediction model containing only gender and race.

Classifying the calibration sample, the proportion of correct classifications for the total model was 75% and for the model including only gender and race it was 63%. A 2x2 table illustrating the number of hits and misses for both models is:

All Predictors

		MISS	HIT
Gender and Race	HIT	9	39
	MISS	10	18

The test statistic, z = 1.73, would typically be considered non-significant (P = .08) and therefore offers no evidence that the additional variables add to the classifi [illegible] accuracies for these two three predictor variable models (number of elementary schools, number of repeats, and family structure, 79%; number of elementary schools, number of repeats, and number of absences, 79%) were each greater than for the total six variable predictor model. Thus, unlike the multiple correlation coefficient in MR, even with non-cross-validated "internal" estimates of classification hit-rate, accuracy does not necessarily monotonically increase as one adds predictor variables. A different perspective concerning contrasting reduced and full model predictor variable subsets may therefore be necessary for DA applications.

One may argue, however, that the cross-validated estimate of accuracy should be used in any case. An illustration of the impact that using a cross-validated estimator might have is that the leave-one-out estimator for the hit rates involved in the hypothesis tested above were 64% for the full six-variable model, and 49% for the three variable model, with a resulting test statistic of z = 2.45, which is of course significant at the .02 level. Therefore, the researcher would most likely come to a different conclusion concerning the significance of the increment due to these additional variables using cross-validated estimates.

Nonleast Squares Methods

Non least-squares prediction strategies, particularly ridge regression, have received a great deal of attention in the MR literature (e.g., Darlington, 1978; Morris, 1982, 1986; Pagel & Lunneborg, 1985; Rozeboom, 1979), and some attention in DA (Campbell, 1980; DiPillo, 1976; 1977; 1979; Morris & Huberty, 1987). As the benefit to predictive accuracy of such methods is a function of whether the context is relative or absolute accuracy, the results for DA tend to be a subset of those for MR. They appear to be largely parallel to the case of absolute accuracy in the MR case (Morris & Huberty, 1987); enhanced predictive accuracy is available under certain limited circumstances, however, decrements in accuracy are just as likely to occur without an informed decision about when to use the technique. Ridge methods are far from the panacea that they have been purported to be for either the MR or DA case. A suggested method for choosing between alternate predictor weighting algorithms, including ridge and least squares, has been advanced for the DA case by Morris and Huberty (1987), and for the MR case by Morris (1986). Computer programs for both analysis types are available.

References

Allen, D. A. (1971). The prediction sum of squares as a criterion for selecting predictor variables (Tech. Rep. No. 23). University of Kentucky, Department of Statistics.

Allen, D. A., & Cady, F. B. (1982). Analyzing experimental data by regression. Belmont, CA: Wadsworth.

Bartlett, M. S. (1951). An inverse matrix adjustment arising in discriminant analysis. Annals of Mathematical Statistics, 22, 107.

Campbell, N. A. (1980). Shrunken estimates in discriminant and canonical variate analysis. Journal of the Royal Statistical Society, 29, 5-14.

Darlington, R. B. (1978). Reduced variance regression. Psychological Bulletin, 85, 1238-1255.

DiPillo, P. J. (1976). The application of bias to discriminant analysis. Communications in Statistics, A5, 843-854.

DiPillo, P. J. (1977). Further applications of bias to discriminant analysis. Communications in Statistics, A6, 933-943.

DiPillo, P. J. (1979). Biased discriminant analysis: Evaluation of the optimum probability of misclassification. Communications in Statistics, A8, 1447-1457.

Flury, B., & Riedwyl H. (1985). T^2 tests, the linear two-group discriminant function and their computation by linear regression. The American Statistician, 39, 20-25.

Gollob, H. F. (1967, September). *Cross-validation using samples of size one.* Paper presented at the meeting of the American Psychological Association, Washington D.C.

Huberty, C. J. (1984). Issues in the use and interpretation of discriminant analysis. *Psychological Bulletin, 95,* 156-171.

Huberty, C. J, & Mourad, S. A. (1980). Estimation in multiple correlation/prediction. *Educational and Psychological Measurement, 40,* 101-112.

Lachenbruch, P. A., & Mickey, M. R. (1968). Estimation of error rates in discriminant analysis. *Technometrics, 10,* 1-11.

McLachlan, G. J. (1975). Confidence intervals for the conditional probabilities of misallocation in discriminant analysis. *Biometrics, 31,* 161-167.

McNemar, Q. (1947). Note on the sampling error of the differences between correlated proportions or percentages. *Psychometrika, 12,* 153-157.

Morris, J. D. (1982). Ridge regression and some alternate weighting techniques: A comment on Darlington. *Psychological Bulletin, 91,* 203-210.

Morris, J. D. (1983). Stepwise ridge regression: A computational clarification. *Psychological Bulletin, 91,* 363-366.

Morris, J. D. (1986). Microcomputer selection of a predictor weighting algorithm. *Multiple Linear Regression Viewpoints, 15,* 53-68.

Morris, J. D., & Huberty, C. J. (1986, April). *A comparison of three methods of classification hit-rate estimation.* Paper presented at the meeting of the American Educational Research Association, San Francisco.

Morris, J. D., & Huberty, C. J. (1987). Selecting a two-group classification weighting algorithm. *Multivariate Behavioral Research*, In press.

Mosteller, F., & Tukey, J. W. (1968). Data analysis, including statistics. In G. Lindzey, & E. Aronson (Eds.), *Handbook of social psychology: Vol 2.* Reading Mass.: Addison-Wesley.

Pagel, M. D., & Lunneborg, C. E. (1985). Empirical evaluation of ridge regression. *Psychological Bulletin*, *97*, 342-355.

Rozeboom, W. W. (1978). Estimation of cross-validated multiple correlation: A clarification. *Psychological Bulletin*, *85*, 1348-1351.

Rozeboom, W. W. (1979). Ridge regression: Bonanza or beguilement? *Psychological Bulletin*, *86*, 242-249.

Tatsuoka, M. M. (1971). *Multivariate analysis.* New York: John Wiley.