Testing Interactions in Classification Problems

Bernadine Beard **Valerie Bryan**
John D. Morris **Mary G. Lieberman**
Florida Atlantic University

The purpose is to demonstrate a procedure for testing the increment to classification accuracy afforded by an interaction term using predictive discriminant analysis (PDA) and logistic regression (LR). The PSAT (Math) and ethnicity were employed to predict students at risk of passing or failing the Florida Comprehensive Assessment Test (FCAT) Math. Results favored PDA over LR for the failure group. For PDA, at an a priori alpha of .05, the moderation of ethnicity was not significant for the failure group, but was for both passing and the total sample, as were hit-rates for total sample and separate groups by Ethnicity.

Researchers employing multiple linear regression frequently use the well-known technique of hypothesis testing through contrasting the predictive variance attributable to full versus restricted models. This method's power, generality, and applicability to a very wide range of questions in science form a theoretical umbrella under which most univariate inferential statistical tests can be viewed.

In multivariate statistics, a test of the so-called additional information hypothesis (AIH) was suggested by C. R. Rao over 50 years ago (see Hand, 1981, p. 149). The context of this hypothesis is that of one-factor multivariate analysis of variance. The research question associated with this hypothesis pertains to an assessment of the difference between the intergroup distance when all response variables are analyzed and the intergroup distance for a subset of response variables. The hypothesis, then, is that the omitted set of variables adds no information (in the sense of intergroup distance) to that yielded by the subset included. The research question considered in this situation is different from that which occurs in a predictive context because the same criteria are not appropriate. For Rao's AIH, the criterion is intergroup difference, whereas, for the prediction problem, the criterion is classification accuracy (see Huberty & Wisenbaker, 1992) – total, or that obtained within each of the separate groups.

The same type of model contrast explanatory increment question can be asked, and seems to be of at least as much potential interest, in classification questions. Specifically, the question arises in studies using predictive discriminant analysis (i.e., classification), logistic regression, as well as other methods of classification. In this case, the criterion for model accuracy is some form of classification accuracy. The test concerns the difference in proportion of correct classifications (hit-rate) between full and restricted models, just as is done using the R2 in multiple regression. The appropriate test statistic is McNemar's (1947) contrast between correlated proportions, and was introduced by Morris and Huberty (1991; 1995) for the purpose of full versus restricted model testing in predictive discriminant analysis for specific planned contrasts using the total group hit-rate as the criterion.

One of the difficulties with the application of the McNemar statistic to full versus restricted model classification questions is that it requires tallying the number of subjects classified correctly and incorrectly and summarizing the results in a fourfold table corresponding to the full and restricted models. To obtain the entries for that table, one needs more than a knowledge of hit-rates for each model; one must count the number of subjects who were correctly and incorrectly classified in both the full and restricted models in turn. Thus, the total and separate-group hit-rates that are available from standard discriminant analysis and logistic regression package programs are not sufficient information to complete the comparison between full and restricted models. For each individual case, one must tally whether the subject was classified correctly or incorrectly jointly for the full and reduced models. Moreover, if one considers cross-validated classification to be the appropriate metric of model accuracy, then these classifications/misclassifications that are to be tallied must be cross-validation estimates. A computer program to accomplish this otherwise difficult task has been made available in the case of discriminant analysis (Morris & Huberty, 1991; 1995) and logistic regression (Lieberman, Morris & Huberty, 2000).

One typical use of a full vs. restricted model test in multiple regression is in the consideration of an interaction, tested in multiple regression by considering the contribution of a standard multiplicative term (often with variables centered to avoid collinearity problems). Our purpose in this paper is to extend this procedure and illustrate an example thereof in a classification model. Note that these variable importance tests are based on the increment to classification accuracy and are quite different than tests of an

interaction term β available in standard computer packages. Those tests are appropriate when considering partial influence on group separation, but not on hit-rate – the criterion of importance in a classification analysis. As this is simply an extension of the full vs. restricted model testing paradigm to the interaction question, the same aforementioned FORTRAN computer program is applicable.

One may argue, however, that because of the positive bias of estimation of hit-rate classification of the calibration sample, a cross-validated estimate of accuracy should be used. A nonparametric approach to estimating cross-validated hit-rate, which has a wide following in the discriminant analysis literature, is the leave-one-out procedure (Huberty & Olejnik, 2006; Huberty & Mourad, 1980; Lachenbruch & Mickey, 1968; Mosteller & Tukey, 1968). In this method, a subject is classified by applying the rule derived from all subjects 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. We show how interaction tests using full versus restricted model testing, parallel to that used in multiple regression, can be extended to classification studies. We illustrate the interaction test for total as well as separate-group Leave-One-Out classification accuracies for both predictive discriminant analysis (PDA) and Logistic Regression (LR).

Method

An example provided herein regards predicting a high-stakes state mandated, "pass or fail," test from a prior low-stakes test hoped to be diagnostic thus aiding in remediation. Specifically the Florida "FCAT" is predicted from the PSAT. Specifically, the question examined is whether the accuracy with which the PSAT Math can classify subjects correctly in regard to passing or failing the FCAT Math is moderated by ethnicity. In this case both the PDA and LR model were created using PSAT (centered on its mean), Ethnicity, and their product predicting FCAT success. The contrast of interest was that between the hit-rate for all three variables and that afforded by excluding the cross-product term, thus testing the increment to classification accuracy afforded by the moderator variable.

Results and Discussion

Table 1 illustrates the cross-validated (leave-one-out) hit-rates from a linear discriminant function with equal priors for the total sample (N=533), as well as by ethnicity. First, for both ethnicities prediction is most accurate for those who fail, which, if there is to be a separate group accuracy difference is in the desirable direction. Also, hit-rate is more accurate for the total sample as well as for separate groups for Ethnicity 1 than Ethnicity 2. The question of interest is whether hit-rate is significantly moderated by ethnicity.

Table 1 shows the results. Although not of primary interest in this study, the difference in cross-validated hit-rate for LR than PDA is of interest. Because of the poor predictive performance of the LR model (using both ethnicities) for the "Failed" group, a better choice would be the PDA model. As primary interest was in correct classification of the subjects who fail the high stakes FCAT, only the interaction tests for the PDA model will be discussed herein.

For the PDA results, if the researcher posited an alpha of .05, the aforementioned moderation tests would lead the researcher to conclude that the moderation of ethnicity on hit-rate was not significant for the FCAT failure group ($p > .05$), but was for both the Passing group (McNemar $z = 3.32$, $p < .01$) and Total sample (McNemar $z = 2.89$, $p < .05$). Thus one could assert that the model was significantly more accurate at predicting group membership for subjects who passed, and for the combination of subjects passing and failing, for the Majority ethnicity than for the corresponding Minority. As well, as the hit-rates are simple ratio statistics, their simple difference (e.g. 82.4% - .66.7%= 15.7%, and 83.3%-69.9= 13.4%) or a proportional increment [e.g. (82.4%/66.7% -1) = 23.5%, and (83.3%/69.9% -1) = 19.2%], depending on notions of purpose and whether group size is a good estimate of the corresponding parameter, might serve as effect size estimators.

Of course, as in multiple regression, the procedure is not limited to categorical moderators. The relevant computer program is available from the authors.

Table 1. PDA and (LR) hit-rate predicting FCAT from PSAT for all Subjects and by two Ethnicities

	Hit-Rate for: Failed FCAT		Passed FCAT		Total (of row)	
All Subjects	88.9%	(44.4%)	79.4%	(97.1%)	80.9%	(89.1%)
Majority	91.9%	(56.8%)	82.4%	(97.5%)	83.3%	(93.3%)
Minority	79.5%	(47.7%)	66.7%	(93.0%)	69.9%	(81.5%)

References

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

Hand, D. J. (1981). *Discrimination and classification*. New York: Wiley.

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

Huberty, C. J and Olejnik, S. (2006). *Applied MANOVA and discriminant analysis*. New York: Wiley.

Huberty, C. J & Wisenbaker, J. M. (1992). Discriminant analysis: Potential improvements in typical practice. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 2, pp. 169-208). Greenwich, CT: JAI Press.

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

Lieberman, M. G., Morris, J. D., & Huberty, C. J (2000, April). *Full vs. restricted model testing in logistic regression*. Paper presented at the meeting of the American Educational Research Association, New Orleans.

Looney, S. W. (1988). A statistical technique for comparing the accuracies of several classifiers. *Pattern Recognition Letters*, *8*, 5-9.

Morris, J. D., & Huberty, C. J (1991, April). *Full vs. restricted model testing in discriminant analysis*. Paper presented at the meeting of the American Educational Research Association, Chicago.

Morris, J. D., & Huberty, C. J (1995). Full versus restricted model testing in predictive discriminant analysis. *Journal of Experimental Education*, *63*, 161-165.

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

Mosteller, F. & Tukey, J. W. (1968). Data analysis, including statistics. In G. Lindzey & E. Aronson (Eds.). *Handbook of social psychology* (Vol. 2, pp. 80-203). Reading, MA: Addison-Wesley.

Send correspondence to: Mary G. Lieberman
Florida Atlantic University
Email: mlieberm@fau.edu