Mallow's C_p for Selecting Best Performing Logistic Regression Subsets

Mary G. Lieberman **John D. Morris**
Florida Atlantic University

Mallow's C_p is used herein to select maximally accurate subsets of predictor variables in a logistic regression. Across a wide variety of data sets, an examination of the cross-validated prediction accuracy, posited as the ultimate criterion for model performance, contrasts the leave-one-out performance of Mallow's C_p selections with the accuracy afforded by optimal subsets. Losses in accuracies ranged from no loss in several data sets up to a maximum of 10%. The performance of C_p selected subsets can be viewed as promising. It is posited that one should also consider parsimony and the richness of multiple optimal models.

This study investigates the proposition by Hosmer and Lemeshow (2000) that Mallow's C_p be used to select subsets of maximally accurate predictor variables in a logistic regression. As accurate cross-validated prediction accuracy is considered the ultimate criterion for prediction model performance, an examination, across a wide variety of data sets, of the leave-one-out performance of Mallow's C_p selected subsets (in respect to the accuracy of the optimal subset) is examined.

Multiple regression is so thoroughly entrenched in statistical methods that it hardly needs an introduction herein, and is, thus, an obvious modeling technique used to examine the predictive accuracy of subsets of variables. Among the techniques used for solving classification problems, logistic regression (LR) and predictive discriminant analysis (PDA) are two of the most popular (Yarnold, Hart & Soltysik, 1994). Unlike PDA, LR captures the probabilistic distribution embedded in a categorical outcome variable, avoids violations to the assumption of homogeneity of variance, and does not require strict multivariate normality. Therefore, when PDA assumptions are violated, we might expect greater cross-validated classification accuracy with LR than PDA.

Although several studies have compared the classification accuracy of LR and PDA, the results have been inconsistent. For example, some studies (Baron, 1991; Bayne, Beauchamp, Kane, & McCabe, 1983; Crawley, 1979) suggest that LR is more accurate than PDA for nonnormal data. However, several researchers (e.g., Cleary & Angel, 1984; Knoke, 1982; Krzanowski, 1975; Lieberman & Morris, 2003; Meshbane & Morris, 1996; Press & Wilson, 1978) found little or no difference in the accuracy of the two techniques with PDA often performing better than LR. Part of the reason these results are in dispute is that one may consider accuracy for all groups or separate-groups. As well, one may consider a cross-validated index of accuracy or the accuracy of reclassifying the calibration sample; these studies are not consistent in respect to the criterion of accuracy used. Specifically, examination of cross-validation accuracy in LR studies is uncommon, and when done is usually of the most basic (and unstable) sort (hold-out sample). No commercial computer packages support more appropriate resampling cross-validation methods (variously called PRESS, Lachenbruch U, leave-one-out, jackknife and the bootstrap).

Whichever method (LR or PDA) is selected, one may consider subsets of all possible variables for purposes or parsimony, or to *increase* cross-validation accuracy of the model (Morris & Meshbane, 1995). The most usual method is to consider accuracy in classification of the sample upon which the model is created (internal) with the objective of parsimony. That is, realizing that some accuracy will be lost in reducing the number of predictor variables in classifying the calibration sample, but compromising that loss with the gain in parsimony by the reduction in size of the prediction model. However, as in multiple regression, an increase in cross-validated prediction accuracy (the most appropriate criterion) is almost always available using a model composed of fewer than all available variables. Thus one may gain both parsimony and some degree of explanatory power for the model. In addition, although traditional methods considering the piecemeal change in performance of models in respect to prediction within the calibration sample have often been used (forward, backward, stepwise, or variants thereof), they are neither optimal, nor unique, and are now generally in disfavor.

In the case of PDA an examination of the cross-validation accuracy of all 2^p-1 (where p is the number of predictor variables) subsets of variables has been recommended and utilized (Huberty, 1994; Huberty & Olejnik, 2006; Morris & Meshbane, 1995). In this case the method of cross-validation is the leave-one-out method. In the leave-one-out procedure (Huberty, 1994, p. 88; Lachenbruch & Mickey, 1968; Mosteller & Tukey, 1968) 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. Clearly the same round-robin procedure can be used to estimate either relative or absolute accuracy in the use of multiple regression and has appeared in that context, with perhaps the earliest reference due to Gollob (1967). In a system intended to select optimal multiple regression predictor variable subsets, Allen (1971) coined the procedure *PRESS*, and he appears to be the source most often cited in the multiple regression literature.

In the case of PDA (and regression) a matrix identity due to Bartlett (1951) allows the task of N-1 discriminant analyses to be accomplished with far less computational labor that would otherwise be necessary. However, this mathematical tool is irrelevant to the iterative method of LR optimization, thus N-1 LR optimizations must be completed for each of 2^p-1 subsets of predictor variables.

Unlike most LR studies that consider calibration sample statistics as the criterion for model fit (e.g., the Cox & Snell, or Nagelkerke R^2), the criterion for model accuracy is construed in this study, as is typically done in PDA, as classification accuracy - that is, the proportion of correct leave-one-out cross-validated classifications (hit-rate) for the total sample and each separate group. Thus for a two-group problem, we order the accuracy of our 2^p-1 candidate LR equations according to three different (total sample and each group) cross-validated classification accuracy criteria.

An alternative logistic regression variable selection strategy has been proposed by Hosmer and Lemeshow (2000) using a technique due to C. L. Mallows (1973). Although Mallows' technique was intended for OLS regression variable subset selection, with attendant consideration of its merit in that context (e.g., Schumacker, 1994), the direct suggestion of Hosmer and Lemeshow of its use in variable subset selection in logistic regression is directly examined herein.

Methods

Analyses from 19 two-group classification problems from Morris and Huberty (1987) were used in this comparison. Although not purported to represent all potential data structures, these data sets have been used in several classification studies as representing a wide variety of number of predictor variables, group separation, and covariance structures.

For a variety of data sets the leave-one-out cross-validated classification accuracies for the Mallows C_p selected variable subset was compared to that derived from the subset manifesting maximum classification accuracy. The difference between the maximum hit-rate and number of predictors for the best subset and that selected by Mallows C_p was compared. The criterion of model accuracy in this study is the proportion of correct leave-one-out cross validated classifications (hit rate) for the total sample and each separate group.

Results & Discussion

Table 1 shows the data source, number of predictors for the full model, hit-rate for the full model, number of predictors in the best subset (s), and maximum hit-rate in the first five columns from left to right. For Mallows C_p, the final three columns show the number of predictors in the C_p selected subset, the hit-rate for that subset, and the percentage loss in hit-rate from the best subset chosen from the maximum hit-rate.

In all cases selection of the best performing subset (of the 2^p-1 possibilities) offers a reduction in the number of predictor variables, often by more than half, thus parsimony is well served. In the first five data sets there is no loss in hit-rate accuracy and equal parsimony using Mallows C_p as with respect to all possible subsets. In data sets numbered seven and fifteen there is no loss in hit-rate accuracy, although the most parsimonious subset is not selected by C_p. In the remaining data sets, losses in accuracy incurred by use of the C_p strategy ranged from .97% – 10.60%.

In several cases one can have enhanced parsimony, hit-rate accuracy close to maximum, and reduced computational intensity using Mallows C_p as the predictor variable selection procedure. The performance of Mallows C_p could be viewed as promising.

Another use of the consideration of the accuracy of all possible subsets involves the treatment of missing data. Table 2 demonstrates the potential use of several alternative "best" models. These data represent the top twenty best subsets of variables in an 8th grade dropout profiling study including

Table 1. Data set, # variables (p), Hit rate for all, Maximum, and C_p selected and % Loss.

#	Data Set Source	p	Hit-rate for p Predictors	# Predictors in Best Subset(s)	Maximum Hit-rate	C_p # Predictors	C_p Hit-Rate	% Loss
1	Rulon Grps 1 & 2	4	0.803	3	.815	3	.815	**0.00**[a]
2	Rulon Grps 1 & 3	4	0.914	3	.934	3	.934	**0.00**
3	Rulon Gps 2 & 3	4	0.824	3	.830	3	.830	**0.00**
4	Block - Grps 1 & 2	4	0.692	1,2	.718	1	.718	**0.00**
5	Block - Grps 1 & 3	4	0.620	3,4	.620	3	.620	**0.00**
6	Block - Grps 1 & 4	4	0.577	1,2	.628	2	.615	0.02
7	Block - Grps 2 & 3	4	0.566	1,2	.605	2	.605	0.00
8	Block - Grps 2 & 4	4	0.587	2	.627	1	.587	6.37
9	Block - Grps 3 & 4	4	0.684	3	.697	1	.632	9.32
10	Demographics	8	0.591	4	.620	3	.609	1.77
11	Dropout from 4th	10	0.660	4	.787	4	.681	10.60
12	Dropout from 8th	11	0.725	3	.782	4	.746	4.60
13	Fitness	10	0.591	4	.620	4	.588	5.16
14	Warncke-Grps 1 & 3	10	0.600	4	.667	3	.619	7.19
15	Bisbey 1& 2	13	0.879	6,7,8,9,10	.914	9	.914	0.00
16	Bisbey 2& 3	13	0.856	5,6,7	.924	3	.915	.97
17	Talent - Grps 1 & 3	14	0.621	5	.733	2	.707	3.54
18	Talent - Grps 3 & 5	14	0.787	6,7,8,9	.858	7	.811	5.47
19	Talent - Grps 1 & 5	14	0.740	5	.797	7	.751	5.77

[a] Bold denotes equal performance and parsimony.

Table 2. Ranked 20 best (of 255) performing subsets, and total model.

HIT-RATE	SCHOOLS8	REPEATS8	READING8	MATH8	LANG8	SCIENCE8	SOCST8	DSFS8
0.753	√			√		√		√
0.747	√			√				√
0.747	√							√
0.747	√				√	√	√	√
0.747	√			√	√	√	√	√
0.741	√		√		√		√	√
0.741	√			√		√	√	√
0.735	√		√		√	√		√
0.735	√			√			√	√
0.735	√	√	√				√	
0.735	√			√	√		√	√
0.735	√					√	√	√
0.735	√		√	√	√			√
0.735	√	√						√
0.735	√	√	√					
0.728	√	√				√	√	
0.728	√					√		√
0.728	√	√				√		√
0.728	√	√	√			√		√
0.728	√	√				√		
Total Model								
0.679	√	√	√	√	√	√	√	√

number of schools attended by the 8th grade, standardized test scores, and the number of D's and F's obtained during the 8th grade year. Depending on which variables are missing for a subject, with knowledge of the best performing subsets, it may be possible to select a superior subset appropriate for data that a subject has available. An advantage to looking at all possible subsets is the allowance for the elimination of variables for which numbers of subjects are missing data.

The table shows a check mark if a variable appears in each of the top twenty models (out of two hundred and fifty five). Considering column-wise entries, a frequent notion of variable importance seems appropriate. When parsimony and accuracy are considerations for model fit, it is clear from these data that, for example, schools attended by 8th grade is a 'don't leave home without it' variable, as it appears in all of the top twenty models. Similarly, Number of D's and F's obtained by eighth grade appears in most models as does number of science courses taken by eighth grade. The other variables, although desirable, may demonstrate little adequacy, in an additive sense, for inclusion in a prediction model. Therefore, this view of variable importance is such that since some variables appear in all or most models, one might suggest this as a defensible measure of variable importance.

In this particular case, since current emphases on standardized testing, and other indices of achievement, tend to focus on predicting success and profiling students at risk, while lessening the drain on time consumption and fiscal resources, such a measure of variable importance may be considered a vital aspect of any prediction formula.

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.

Baron, A. E. (1991). Misclassification among methods used for multiple group discrimination – The effects of distributional properties. *Statistics in Medicine*, *10*, 757-766.

Bartlett, M. S. (1951). An inverse matrix adjustment arising in discriminant analysis. *Annals of Mathematical Statistics*, *22*, 107-111.

Bayne, C. K., Beauchamp, J. J., Kane, V. E., and McCabe, G. P. (1983). Assessment of Fisher and logistic linear and quadratic discrimination models. *Computational Statistics and Data Analysis*, *1*, 257-273.

Cleary, P. D. & Angel, R. (1984). The analysis of relationships involving dichotomous dependent variables. *Journal of Health and Social Behavior*, *25*, 334-348.

Crawley, D. R. (1979). Logistic discriminant analysis as an alternative to Fisher's linear discriminant function. *New Zealand Statistics*, *14(2)*, 21-25.

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

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression*. New York: Wiley.

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

Huberty, C. J (1994). *Applied discriminant analysis*. New York: Wiley.

Knoke, J. D. (1982). Discriminant analysis with discrete and continuous variables. *Biometrics*, *38*, 191-200.

Krzanowski, W. J. (1975). Discrimination and classification using both binary and continuous variables. *Journal of the American Statistical Association*, *70*, 782-790.

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

Lieberman, M. G, & Morris, J. D. (2003, April). *Comparing classification accuracies between predictive discriminant analysis and logistic regression in specific data sets*. Paper presented at the meeting of the American Educational Research Association, Chicago.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, *15*, 661-675.

Meshbane, A., & Morris, J. D. (1996, April). *Predictive discriminant analysis versus logistic regression in two-group classification problems*. Paper presented at he meeting of the American Educational Research Association, New York.

Morris, J. D., & Huberty, C. J (1987). Selecting a two-group classification weighting algorithm. *Multivariate Behavioral Research*, *22*, 211-232.

Morris, J. D., & Meshbane, A. (1995). Selecting predictor variables in two-group classification problems. *Educational and Psychological Measurement, 55*, 438-441.

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.

Press, S. J., & Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, *73*, 699-705.

Schumacker, R. E. (1994). A comparison of the Mallows C_p and principal component regression criteria for best model selection in multiple regression. *Multiple Linear Regression Viewpoints*, 21, 12-22.

Yarnold, P. R., Hart, L. A. & Soltysik, R. C. (1994). Optimizing the classification performance of logistic regression and Fisher's discriminant analysis. *Educational and Psychological Measurement*, *54*, 73-78.

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