

Computing the Approximate and Exact Leaving-One-Out Method in Multiple Group Logistic Discrimination with the SAS System

Mohamed H. Hussein
The University of Georgia

ABSTRACT

Despite the popularity of logistic discrimination in the social and behavioral science the author is not aware with any commercial statistical software which provides the practitioner with another estimate than resubstitution method. This paper presents SAS program to conduct an approximate leaving-one-out and exact leaving-one-out for estimating the classification hit rate in multiple-group logistic discrimination. Monte Carlo study was conducted to assess the effect of the sample size on the performance of the approximation method relative to resubstitution method and exact leaving-one-out. The usefulness of this approximation is demonstrated by means of educational example.

INTRODUCTION

In the social and behavioral sciences in general, and in education and psychology in particular, there is often a need to classify units (e.g., students) based on a collection of predictors, that are associated with a outcome variable. The reader for the predictive discriminant analysis (PDA) literature will find out that many classification rules have been suggested.

LCF is widely used because of its simplicity and its optimal properties; the performance of the LCF is optimal, in particular, when the populations are normal with known parameters.

In social sciences, deviations from normality appear more the rule than the exception.

If multivariate normality is violated, the estimated of classification error rates from LCF may be biased (Dillon, 1979).

In the 1960s, researchers began to explore alternatives to LCF when classifying observations with predictors that were not all continuous. At about the same time Cox (1966) and Day and Kerridge (1967) Proposed logistic regression as a method of classification, the later authors referring to it as

a 'general maximum likelihood discriminant'.

By the end of the 1960s, LOGDF became one of the many possible rules suggested for classification. LOGDF can be classified midway between fully distributional solutions, of which the assumption of multivariate normality, such as LCF, and the distribution-free rules, using for instance, kernel or nearest-neighbor rules. Therefore it is often called partially parametric or semi-parametric rule (Albert & Lesaffre, 1989).

Most statistical software packages now include logistic regression algorithms. The BMDP (PMDP Statistical Software, 1990) algorithm is BMDPLR. The 1990 edition of BMDP includes a routine to analyze polytomous outcomes, which is called BMPDPR. SAS (SAS Institute, Inc. 1990) includes two logistic regression procedures, CATMOD and LOGISTIC. Both SPSS (SPSS, version 9.0) and SYSTAT (version 9.0) include two logistic procedures, binary logistic and multinomial logistic.

THE MULTIPLE-GROUP LOGISTIC DISCRIMINATION

Consider classification of a unit into one of g populations ($\Pi_1, \Pi_2, \dots, \Pi_g$) on the basis of a measurement vector $\mathbf{x}=(\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_r)$ with $\mathbf{x}_0=\mathbf{1}$. The observation $\mathbf{x}$ fulfils the logistic assumption if (Anderson, 1972)

$$\ln\left(\frac{p_s(\mathbf{x})}{p_g(\mathbf{x})}\right) = \alpha_s^T \mathbf{x} \quad (s = 1, 2, \dots, g - 1) \quad (1)$$

Where $p_s(\mathbf{x})$ is the posterior probability of belonging $\mathbf{x}$ to group Π_s ; α is a vector of unknown parameters. In this model the last group g has been taken as reference (base) group; using the above results, the choice of another reference group would simply imply a reparameterization of the model that does not effect the posterior probabilities (Jobson, 1991; Albert & Lesaffre, 1986).

The objective of building a classification rule is to correctly classify as many future objects as possible. There are many methods for estimating the actual error rates. Among these are resubstitution (apparent), holdout method, L-O-O (Lachenbruch & Mickey, 1968), and bootstrap Method (Efron, 1983). The resubstitution is the simplest method; the training data are classified using the classification rule estimated from the training sample. Several studies indicated that the resubstitution estimator yields a positively biased estimate (Hand, 1983, Lesaffre & Willems, 1989; Mclachlan, 1977).

The L-O-O method makes use of all of the available data, as does the resubstitution method without serious bias in the estimated classification error rates. This method holds out one observation at a time, estimates the classification function based on n-1 observations, and classifies the held out observation. This process is repeated until all observations are classified. It has proven that this method yields a more unbiased estimate of the error rate than the resubstitution method (Lesaffre & Willems, 1989).

Despite the popularity of logistic discriminant function in the context of multiple-group situation, the current author is not aware with any commercial program that provides the practitioner with another estimate than resubstitution method. In theory, there is no problem to apply the L-O-O method; however in practice the cost in computer time is prohibitive. The application of the L-O-O method is problematic for logistic model because of the difficulty in determining $\hat{\alpha}^{(i)}$, the maximum likelihood estimate of α , based on the sample without the i^{th} case. Therefore Lesaffre et al. (1989) proposed replacing $\hat{\alpha}^{(i)}$ by its so-called one-step-approximation $\alpha^*(i)$. The "approximate L-O-O" which is based on "approximately" removing each case from the sample and calculating in this way an approximate L-O-O error rates. Following Lesaffre and Albert (1989), one can obtain

$$\alpha^*(i) = \hat{\alpha} - (X^T \hat{V} X_i^T)^{-1} X_i^T \hat{V}_i^{1/2} M_{ii}^{-1} x_i \quad (2)$$

Where $X^T = (X_1^T, \dots, X_g^T)$;

$$X_i = \begin{bmatrix} x_i^T & 0^T & \dots & 0^T & 0^T \\ 0^T & x_i^T & \dots & 0^T & 0^T \\ \vdots & \vdots & & \vdots & \vdots \\ 0^T & 0^T & \dots & 0^T & x_i^T \end{bmatrix}$$

$$\hat{V} = \text{diag} (\hat{V}_1, \dots, \hat{V}_n)$$

$$\hat{V}_i = \hat{p}_k(x_i) [\delta_{kl} - \hat{p}_i(x_i)]$$

$$M_{ii} = I - \hat{V}_i^{1/2} X_i (X^T \hat{V} X)^{-1} X_i^T \hat{V}_i^{1/2}$$

$$x_i = \hat{V}_i^{1/2} [c(1/i) - \hat{p}_1(x_i), \dots, c(g/i) - \hat{p}_g(x_i)]^T$$

Where $c(j/i) = 1$ if i^{th} unit belongs to Π_j and 0 otherwise, and δ_{kl} is Kronecker's symbol.

Lesaffre et al. (1989) show that this approximation is an attractive computational method, which results usually in error rates close to the exact-leaving-one-out error rates.

In the present paper, the author describes a SAS program for calculating exact L-O-O and one step approximation L-O-O. A Monte Carlo study was conducted to assess the influence of the sample size on the adequacy of the approximation L-O-O method as an estimate of the actual error rates. The usefulness of this approach is demonstrated by means of an educational example.

THE COMPUTER PROGRAM

This program was written in SAS language, and can be used for any mainframe or microcomputer version of SAS as long as the installation of the SAS system contains the IML module. There is no limit about the number of predictors or sample sizes, users who are familiar with SAS Programming will find that can easily customize the program and the output to suit their own needs.

INPUT

Three pieces of information are required from the user: The location of the data file, the number of predictors and the total sample size. The program assumes that the data are in ASCII format, and adjacent variables are separated by at least one blank space.

OUTPUT

The program prints the posterior probabilities, the predicted and the observed group membership for each unit, the total and individual group hit rates and the classification table.

A SIMULATION STUDY

A Monte-Carlo study was conducted to investigate the influence of sample size on the closeness of the approximation method to exact L-O-O. The sample size is determined by $n = kv$, where v is the number of the predictors, and k was taken as 3, 5, 10, and 20. For each sample size, 200 samples were generated. From each sample the resubstitution classification hit rate, the approximate and the Figure 1. Comparison of three estimates of the actual classification hit rate in relation to

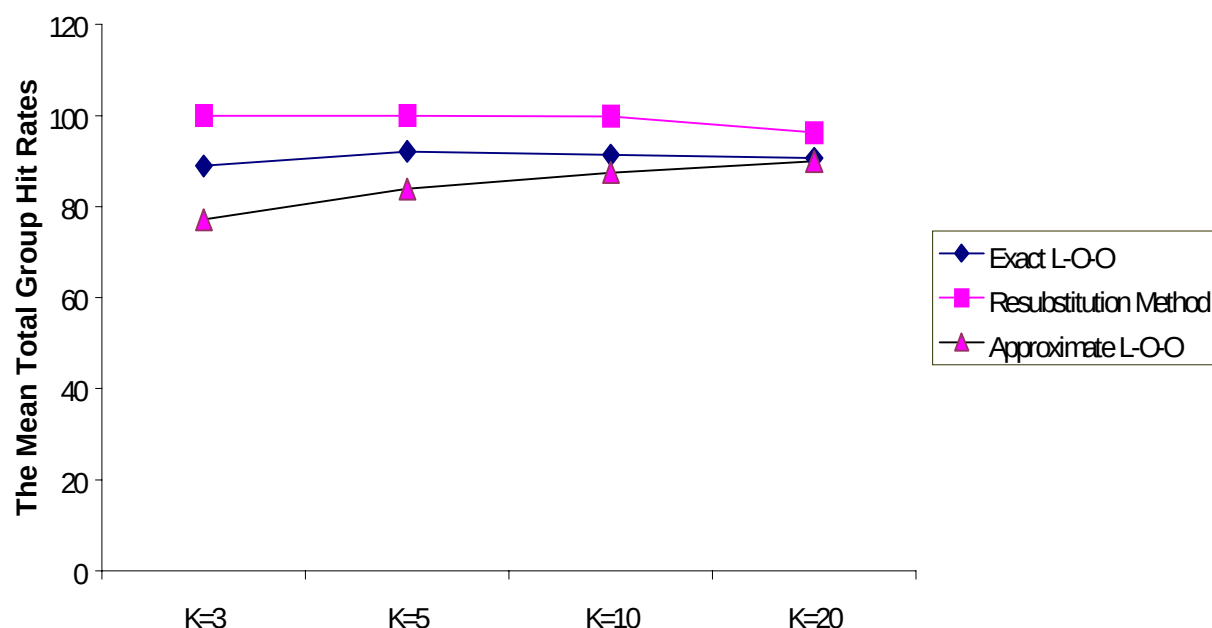


Figure 1. Comparison of three estimates of the actual classification hit rate in relation to N/R (sample size/number of predictors)

exact L-O-O classification hit rates were calculated. The location model (Krzanowski, 1976) was used to generate the data sets with 5 predictors two categorical and three continuous variables. The simulation results are presented in Figure 1. From Figure 1, we observe that The resubstitution method is usually too optimistic. The approximate L-O-O method yields very close estimate to exact L-O-O for k around 10. For practical purpose it is thus clear that the approximate L-O-O method is preferable over the two other methods.

APPLICATION

This data set was constructed by Gerald D. Bisbey in 1968. A sample of 153 student entering college was partitioned according to the college French course in they enrolled. Thirty five ($n_1=35$) Enrolled in the beginning level, $n_2=81$ in the intermediate level, and $n_3=37$ in the advanced level. Thirteen measures were obtained on each of the 153 students. The purpose of the study was to characterize differences among the three groups of students. The estimated logistic classification rule resulted in a resubstitution error rate of 10.4%. The exact L-O-O error rate was equal to

20.27%, whereas the approximation L-O-O was equal to 18.96%. The approximation was very close in this example, whereas the resubstitution was optimistic bias.

PROGRAM AVAILABILITY

Interested readers should contact the author through e-mail (hussein@arches.uga.edu) or by sending a computer disk with return postage to Mohamed Hussein, 320C Aderhold Hall, The University of Georgia, Athens, GA 30602-7143.

REFERENCE

Albert, A. & Lesaffre, E. (1986). Multiple group logistic discrimination. Computation and Mathematics with applications, 12A, 209-224.

Anderson, J. A. (1972). Separate sample logistic discrimination. Biometrika, 59, 19-35.

Cox, D. R. (1966). Some procedures connected with the logistic qualitative response curve. In F. N. David (Ed.) Festschrift for J. Neyman: Research Papers in Statistics (pp.55-71). London: Wiley.

Dillon, W. R. (1979). The performance of the linear discriminant function in nonoptimal situations and the estimation of classification error rates: A review of recent findings. Journal of Marketing Research, 16, 370-381.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. Journal of American Statistical Association, 78, 316-331.

Efron, B. (1990). More efficient bootstrap computations. Journal of the American Statistical Association, 85, 79-89.

Jobson, J. D. (1991). Applied multivariate data analysis volume II: Categorical and multivariate methods. New York: Springer.

Krzanowski, W. J. (1976). The performance of Fisher's linear discriminant function under non-optimal conditions. Technometrics, 19, 191-200.

Lesaffre, E., Willems, J. L., & Albert, A. (1989). Estimation of error rate in multiple group logistic discrimination. The approximate leaving-one-out method. Communications in Statistics: Theory and Methods, 18, 2989-3007.

McLachlan, G. J. (1976). The bias of the apparent error rate in discriminant analysis. Biometrika, 63, 239-244.