A Comparison of Discriminant Analysis and Logistic Regression for the Prediction of Coliform Mastitis in Dairy Cows

M.E. Montgomery, M.E. White and S.W. Martin*

ABSTRACT

Results from discriminant analysis and logistic regression were compared using two data sets from a study on predictors of coliform mastitis in dairy cows. Both techniques selected the same set of variables as important predictors and were of nearly equal value in classifying cows as having, or not having mastitis. The logistic regression model made fewer classification errors. The magnitudes of the effects were considerably different for some variables. Given the failure to meet the underlying assumptions of discriminant analysis, the coefficients from logistic regression are preferable.

Key words: Discriminant analysis, logistic regression, coliform mastitis, dairy cattle.

RÉSUMÉ

Cette expérience visait à comparer les résultats de l'analyse discriminante avec ceux de la régression logistique, à l'aide de deux groupes de données résultant d'une étude sur les prédictions relatives à la mammite due à des bactéries coliformes, chez des vaches laitières. Les deux techniques précitées choisirent le même groupe de variantes comme facteurs importants de prédiction et affichèrent une valeur à peu près égale, en classifiant les vaches comme atteintes ou non de mammite. Le modèle de régression logistique fit moins d'erreurs de classification. La magnitude des effets fluctua considérablement, pour certaines variantes. Comme il s'avéra impossible de rencontrer les suppositions sous-jacentes de l'analyse discriminante, les coefficients de la régression logistique sont préférables.

Mots clés: analyse discriminante, régression logistique, mammite due à des bactéries coliformes, vaches laitières.

INTRODUCTION

The use of multivariate analyses is becoming more common in the veterinary literature. However, the reasons for the selection of a particular analytic method often are not presented, thus leaving readers with an incomplete understanding of what was done. Traditionally, veterinarians have tended to select an analytic procedure with which they are most familiar, although it may not be appropriate, given the underlying statistical assumptions (1).

Both discriminant analysis and logistic regression can be used to predict the probability of a specified outcome using all or a subset of available variables. A related but different use of these techniques is to elucidate the effect of one variable on the outcome (or the change in the risk of the outcome due to the variable) while controlling the effects of, and interrelationships with, other variables. Logistic regression is becoming more widely used, in relation to discriminant analysis, perhaps in large part due to its recent availability in "canned" statistical packages.

In either method, the general formula for determining the probability of an event occurring is:

$$P = \frac{1}{1 + e^{B_0 + B_1X_1 + B_2X_2 + \ldots B_nX_n}}$$

where P is the probability of the event (e.g. disease) occurring, $X_1, X_2, \ldots X_n$ are the predictor (independent) variables, $B_1, B_2, \ldots B_n$ are the coefficients representing the effects of the predictor variables, and B_0 is the intercept (the value of the equation when all of the X's are zero) (2). Discriminant analysis and logistic regression are alternative methods of estimating the intercept and the coefficients; they perform the same task by two different computational methods each of which has its own set of assumptions about the underlying data structure. Discriminant analysis programs usually provide a score, which is in fact the model, i.e. $B_0 + B_1X_1 + B_2X_2 + \ldots B_nX_n$. The ability to correctly identify (classify) those with and without the event is then evaluated at various cut-off values of the score. Most computer programs will also provide the probabilities of group membership (e.g. diseased versus nondiseased) on request. With logistic regression, the model is taken a step further by most programs and the actual probabilities of the event occurring (e.g. disease) are calculated. The classification ability is evaluated, but in this case, at various cut-off probabilities. In this regard, the alternative methods utilize similar information.

There are basic differences in the statistical assumptions which underly those two methods, however. With discriminant analysis, the assumptions are that the independent variables are normally distributed, with equal variance within each group, and that the interrelationships (correlations) among the variables within each group are the same (i.e. equal covariance matrices) (3). These assumptions are usually broken when some or all of the independent variables are attribute type

*Department of Veterinary Microbiology and Immunology, Ontario Veterinary College, University of Guelph, Guelph, Ontario N1G 2W1 (Montgomery, Martin) and Department of Clinical Studies, New York State College of Veterinary Medicine, Cornell University, Ithaca, New York 14853 (White).

Reprint requests to Dr. M.E. White or Dr. S.W. Martin.

Submitted January 21, 1986.

variables (i.e. present or absent). This is particularly important when the objective is to estimate the magnitude of the effects (coefficients) of the predictor variables. Using discriminant analysis, the sign of the coefficients will always be correct, B's equal to zero will be estimated accurately, but estimates of nonzero B's may be biased (2). When the objective is overall prediction or classification as in this study, these assumptions are less constraining, and both methods should provide essentially the same model. The assumptions underlying the discriminant model need not be met when using logistic regression models. Experience also indicates that the coefficients of the logistic model are usually slightly closer to the null value of one, than those derived from discriminant analysis (4).

Several authors have formally compared the two techniques. For example, Halperin *et al* (2) obtained results with from none to several attribute type predictor variables, and noted only small differences in the classification ability between the two analytic procedures. Kleinbaum *et al* (4) compared the classification ability of logistic regression and discriminant analysis using a data set which met the assumptions of discriminant analysis and noted that the logistic model was slightly superior. Press and Wilson (5) concluded that each analytic technique served a unique function: discriminant analysis was useful for classification of observations into one of two or more populations, whereas logistic regression was useful for relating a qualitative (binary) dependent variable to one or more independent variables by a logistic distribution functional form of P (as noted above). The general conclusions are that if the assumptions of normality and equal variance/covariance matrices are met, then discriminant analysis estimators are preferred for the task of classification. If, however, the assumptions are not met, then the logistic regression estimates are preferred for either application (5).

The objective of this paper was to compare the two methods of analysis for classifying subjects into one of two populations. In an earlier study, discriminant analysis was used to select useful variables and build a model for the prediction of coliform mastitis in dairy cattle (6). The resulting model was then validated using a separate data set (7). In this study we compared the results of discriminant analysis to those of logistic regression using these data sets.

MATERIALS AND METHODS

A description of the two data sets used in the creation of the discriminant model and its validation are available elsewhere (6,7). A list of the independent variables available for selection are listed in Table I. Stepwise discriminant analysis was performed using the program DISCRIMINANT from the Statistical Package for the Social Sciences (SPSS^x) (8). The selection criterion for entry of each independent variable was the maximization of the Mahalanobis distance (D squared) between groups, with a minimum F-to-enter of 1.00 and a maximum F-to-remove of 1.00 (6). Equality of the covariance matrices was checked using the test for homogeneity option in the program DISCRIM from the Statistical Analysis System (SAS) (9).

Stepwise logistic regression was performed on the original data set, using the program PLR from the BMD Biomedical Computer Programs (BMDP) (10). Significance levels for entry and removal of variables were set at 0.25 and 0.30 respectively; these p values were selected to approximate the F values used in the discriminant analysis. The coefficients of the variables were estimated by the maximum likelihood technique, after the asymptotic covariance approach was used to determine whether or not terms should remain in the model. (This is a recommended approach to save computer costs since the variable selection is performed consistently well by the asymptotic, noniterative approach.) The overall fit of the model to the data was assessed using the Hosmer and Brown tests (10).

Comparisons between the results of the two analytic techniques were made on the following: the variables selected, the order of selection, and the sign and magnitude of coefficients. Also, for each cow's record, the probability of disease was calculated using the coefficients from each analytic technique. Sensitivity, specificity, and overall accuracy (total correct classification %) at several probability cut-off points were calculated and compared between the two techniques. The latter analyses were also conducted on the validation data set.

TABLE I. Independent Variables Available for Entry

Variable	Description
1) **Continuous Variables**	
AGE	age of cow (months)
PULSE	pulse rate per minute
RESPRATE	respiratory rate per minute
TEMP	temperature °F
2) **Attribute Variables (0 = no; 1 = yes)**	
ABSCESS	palpable abscess in udder
APPET	depressed appetite
CLEAR	milk colour clear
CLOTS	clots in milk
COLDEAR	ear temperature cold
DEHYD	cow dehydrated
DEPRESS	cow depressed
DURMAST	time from mastitis to examination more than one day
FIRMNESS	udder firmness present
HOTSKIN	skin temperature hot
OTHCOL	milk colour other than white or clear
PREVMASQ	previous mastitis in quarter
PREVMAST	previous mastitis in cow
RUMENMOT	rumen motility 2-4 contractions per minute
SWELLING	swelling in quarter
WATERY	watery consistency of milk
WEAK	cow weak

Response operating characteristics (ROC) curves were plotted for each model. An ROC curve graphically displays sensitivity and 100% minus specificity (false positive rate) at several cut-off points, and provides a quick visual assessment of the implications of changing the cut-off value (11). By plotting the ROC curves for two tests on the same axes, one is able to determine which test is better for classification, namely, that test whose curve encloses the larger area beneath it.

RESULTS

The discriminant analysis and logistic regression models are presented in Table II. Results of the test of homogeneity indicated that the covariance matrices were not equal ($p<0.00$), thus this assumption for discriminant analysis was violated. The same variables were selected using both models however, the direction of the relationships were the same, and the order of entry of variables into the models were similar (Table III). There were some extreme differences in magnitude of the coefficients, however (e.g. ABSCESS) (Table II).

TABLE II. Variables and Coefficients for the Discriminant Analysis Model and the Logistic Regression Model

Variable	Discriminant Coefficients	Logistic Coefficients
PREVMASQ	1.241617	2.4915
WEAK	1.638367	1.7480
OTHCOL	- 0.8895808	- 1.0682
SWELLING	0.7703097	1.1384
WATERY	0.7399289	0.95506
PREVMAST	- 0.6466322	- 1.5634
ABSCESS	- 1.387280	-23.270
TEMP	0.1664481	0.22856
Intercept	-17.97046	-25.412

Table IV presents values of sensitivity, specificity, and accuracy for several probability cut-off points for the discriminant and logistic models using the original data. The maximum overall accuracy (i.e. correct classification rate) was increased slightly with the logistic model (82.2% versus 77.5%). At this maximum accuracy, sensitivity and specificity were 47.4% and 96.7% respectively for the logistic model. For the discriminant model, there were two combinations of sensitivity and specificity at the maximum accuracy level: 44.7%/91.2% and 31.6%/96.7%.

TABLE III. Summary of Steps in Analyses

Step	Discriminant Analysis: Variable Added	Discriminant Analysis: Variable Deleted	Logistic Regression: Variable Added	Logistic Regression: Variable Deleted
1	CLEAR		CLEAR	
2	SWELLING		SWELLING	
3	WEAK		WEAK	
4	TEMP		ABSCESS	
5	PREVMASQ		TEMP	
6	ABSCESS		PREVMASQ	
7	OTHCOL		PREVMAST	
8	WATERY		WATERY	
9		CLEAR	OTHCOL	
10	PREVMAST			CLEAR

TABLE IV. Sensitivity, Specificity, and Accuracy of the Discriminant Analysis Model and the Logistic Regression Model at Several Probability Cut-off Values: Original Data Set

Cut-off Value[a]	Discriminant Model: Sens	Spec	Accuracy	Logistic Model: Sens	Spec	Accuracy
0.05	100.0	0.0	29.5	97.4	28.6	48.8
0.10	100.0	4.4	32.6	92.1	39.6	55.0
0.15	97.4	9.9	35.7	92.1	47.3	60.5
0.20	97.4	13.2	38.0	89.5	56.0	65.9
0.25	94.7	20.9	42.6	81.6	67.0	71.3
0.30	94.7	26.4	46.5	76.3	73.6	74.4
0.35	92.1	35.2	51.9	68.4	75.8	73.6
0.40	89.5	47.3	59.7	57.9	83.5	76.0
0.45	84.2	60.4	67.4	52.6	90.1	79.1
0.50	76.3	65.9	69.0	50.0	93.4	80.6
0.55	73.7	72.5	72.9	47.4	96.7	82.2
0.60	68.4	76.9	74.4	44.7	96.7	81.4
0.65	57.9	83.5	76.0	39.5	97.8	80.6
0.70	55.3	85.7	76.7	34.2	98.9	79.8
0.75	44.7	91.2	77.5	21.1	98.9	76.0
0.80	31.6	96.7	77.5	21.1	98.9	76.0
0.85	21.1	98.9	76.0	7.9	100.0	72.9
0.90	5.3	100.0	72.1	2.6	100.0	71.3
0.95	0.0	100.0	70.5	0.0	100.0	70.5

[a] P (disease); values less than or equal to the cut-off value are test negative; those greater than the cut-off value are test positive

Table V presents values of sensitivity, specificity, and accuracy for several probability cut-off values for each model using the validation data set. Note that the ROC curve (Fig. 1) for the logistic regression model remains above that for the discriminant model, indicating that the former model is slightly superior in its classification ability, as was indicated for the original data set.

DISCUSSION

In general, results from the logistic model agreed with those from the discriminant analysis. The overall accuracy of classification was good for both, and either would be useful for the prediction of coliform mastitis in the field. Previously, it was noted that clinicians had a higher sensitivity but lower specificity than the discriminant models (7). The advantage of using a formal mathematical model is that one can manipulate the cut-off point given various costs of incorrect diagnoses to change the sensitivity, specificity, and accuracy to obtain the optimal results for that setting.

Although the assumption of equal covariance was not met with these data sets, both methods selected the same subset of variables. However, if the task was to quantify the effects of the predictor variables, the coefficients of the logistic model would be preferable to those of the discriminant model.

In conclusion, for this particular problem and in agreement with theory, the logistic regression technique resulted in essentially the same model as did discriminant analysis. However, given the more robust nature of logistic

TABLE V. Sensitivity, Specificity, and Accuracy of the Discriminant Analysis Model and the Logistic Regression Model at Various Probability Cut-off Points: Validation Data Set

Cut-off Value[a]	Discriminant Model			Logistic Model		
	Sens	Spec	Accuracy	Sens	Spec	Accuracy
0.01	100.0	0.0	31.6	100.0	0.0	31.6
0.05	100.0	1.3	32.5	100.0	12.8	40.4
0.10	100.0	1.3	32.5	97.2	21.8	45.6
0.15	100.0	2.6	33.3	94.4	29.5	50.0
0.20	100.0	9.0	37.7	88.9	35.9	52.6
0.25	94.4	13.2	40.4	86.1	43.6	57.0
0.30	94.4	24.4	46.5	72.2	55.1	60.5
0.35	88.9	30.8	49.1	66.7	60.3	62.3
0.40	86.1	34.6	50.9	63.9	62.8	63.2
0.45	77.8	51.3	59.6	55.6	64.1	61.4
0.50	75.0	59.0	64.0	47.2	74.4	65.8
0.55	69.4	59.0	62.3	41.7	79.5	67.5
0.60	55.6	69.2	64.9	41.7	84.6	71.1
0.65	41.7	80.8	68.4	38.9	85.9	71.1
0.70	41.7	84.6	71.1	33.3	91.0	72.8
0.75	38.9	91.0	74.6	22.2	92.3	70.2
0.80	27.8	93.6	72.8	11.1	96.2	69.3
0.85	16.7	97.4	71.9	8.3	98.7	70.2
0.90	8.3	98.7	70.2	5.6	100.0	70.2
0.95	2.8	100.0	69.3	2.8	100.0	69.3
0.99	0.0	100.0	68.4	0.0	100.0	68.4

[a]P (disease); values less than or equal to the cut-off value are test negative; those greater than the cut-off value are test positive

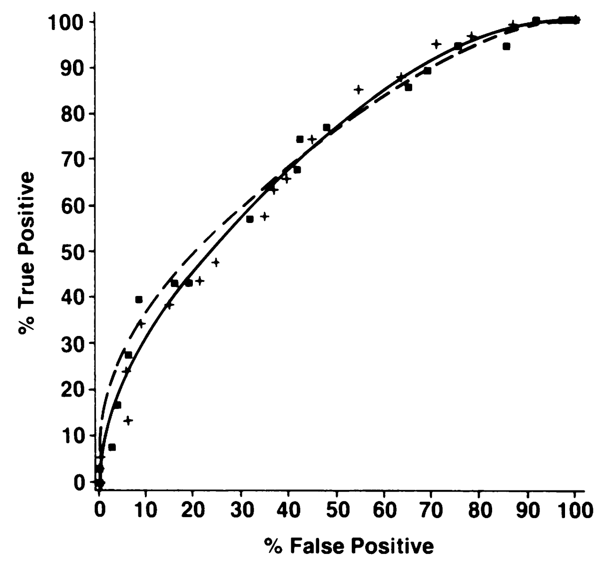

Fig. 1. Response operating characteristics (ROC) curves for the discriminant analysis model and the logistic regression model using the validation data set. Legend: ■————■ discriminant analysis model; + — — — — — — — — + logistic regression model. Probability cut-off values from the top righthand corner to the bottom lefthand corner are: 0.01, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90 and 0.95.

regression, and its slightly better performance, it is preferable to discriminant analysis particularly when the assumptions of normality and equal variance are not met. Nonetheless, previous research in which discriminant analysis was used for purposes of classification (and when the assumptions were invalid or simply not considered) should not be misleading in terms of the variables selected or the signs of the coefficients. As mentioned, the magnitudes of the coefficients may be biased.

In conclusion, a more thorough understanding of analytic techniques and their underlying statistical assumptions will improve the validity of veterinary medical research. For those wishing more information on discriminant analysis, see Morrison (3). Kleinbaum *et al* (4), Engleman (10) and Breslow and Day (12) are useful references on logistic regression.

REFERENCES

1. **GODFREY K**. Statistics in practice: Comparing the means of several groups. New Engl J Med 1985; 313:1450-1456.
2. **HALPERIN M, BLACKWELDER WE, VERTER JI**. Estimation of the multivariate logistic risk function: A comparison of the discriminant function and maximum likelihood approaches. J Chron Dis 1971; 24:125-158.
3. **MORRISON DG**. On the interpretation of discriminant analysis. J Marketing Res 1969; 6:156-163.
4. **KLEINBAUM DG, KUPPER LL, MORGENSTERN H**. Epidemiologic research: principles and quantitive methods. Toronto: Lifetime Learning, 1982: 461-470.
5. **PRESS SJ, WILSON S**. Choosing between logistic regression and discriminant analysis. J Am Stat Assoc 1978; 73:699-705.
6. **WHITE ME, GLICKMAN LT, BARNES-PALLESON FD, PEARSON EG, MONTGOMERY ME, ARMSTRONG D, WICKENDED RP, HICKEY G**. Discriminant analysis of the clinical indicants for bovine coliform mastitis. Cornell Vet 1986; 76: 335-341.
7. **WHITE ME, GLICKMAN LT, BARNES-PALLESEN FD, STEM ES III, DINSMORE P, POWERS MS, POWERS P, SMITH MC, MONTGOMERY ME, JASKO D**. Accuracy of a discriminant analysis model for prediction of coliform mastitis in dairy cows and a comparison with clinical prediction. Cornell Vet 1986; 76:342-347.
8. **SPSS INC**. SPSS[x] user's guide. New York: McGraw-Hill, 1983: 623-645.
9. **SAS INSTITUTE INC**. SAS's user's guide: statistics. 1982 edition. Cary, North Carolina: SAS Institute Inc., 1982: 381-396.
10. **ENGLEMAN L**. PLR — stepwise logistic regression. In: BMDP statistical software. Los Angeles: University of California, 1983: 330-344.
11. **SACKETT DL, HAYNES RB, TUGWELL P**. Clinical epidemiology: A basic science for clinical medicine. Toronto: Little, Brown and Co., 1985: 106-107.
12. **BRESLOW NE, DAY NE**. Statistical methods in cancer research. Vol. I. The analysis of case-control studies. ISBN 92 832 11324. Lyon: Int Ag Res Cancer, 1980: 192-246.