DOCUMENT RESUME

ED 371 021 TM 021 652

AUTHOR Meshbane, Alice; Morris, John D.
TITLE A Method for Selecting between Linear and Quadratic Classification Models in Discriminant Analysis.
PUB DATE Apr 94
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Classification; Comparative Analysis; Correlation; *Discriminant Analysis; *Mathematical Models; *Research Methodology; *Selection; Statistical Studies

ABSTRACT

A method for comparing the cross validated classification accuracies of linear and quadratic classification rules is presented under varying data conditions for the k-group classification problem. With this method, separate-group as well as total-group proportions of correct classifications can be compared for the two rules. McNemar's test for contrasting correlated proportions is used in the statistical comparisons of the separate group and total sample proportions. The method is illustrated with some real data sets. Included are two tables. (Contains 12 references.) (Author)

A Method for Selecting Between Linear and Quadratic Classification Models in Discriminant Analysis

Alice Meshbane and John D. Morris

Florida Atlantic University

Paper presented at the annual meeting of the American Educational Research Association, April 1994, New Orleans, LA

A Method for Selecting Between Linear and Quadratic Classification Models

in Discriminant Analysis

A method for comparing the cross validated classification accuracies of linear and quadratic classification rules is presented under varying data conditions for the k-group classification problem. With this method, separate-group as well as total-group proportions of correct classifications can be compared for the two rules. McNemar's test for contrasting correlated proportions is used in the statistical comparisons of the separate group and total sample proportions. The method is illustrated with some real data sets.

A Method for Selecting Between Linear and Quadratic Classification Models in Discriminant Analysis

Alice Meshbane and John D. Morris

Florida Atlantic University

A method for comparing the cross validated classification accuracies of linear and quadratic classification rules is presented under varying data conditions for the k-group classification problem. The classification rules are based on a Bayesian conditional-probability model assuming multivariate normality within each criterion population. Defining

$$D_{ik}^2 = [(\underline{X}_i - \overline{\underline{X}}_k)' S_k^{-1} (\underline{X}_i - \overline{\underline{X}}_k)]$$

to be the square of the distance from the point in p-space representing individual i (i.e., $\underline{X}_i$) to the point representing the means of the p measures in group k (i.e., $\overline{\underline{X}}_k$), where S_k is the sample (p x p) covariance matrix for group k, the following "quadratic" classification statistic is used:

$$P_{ik} = \frac{p_k \mid S_k \mid^{-1/2} \exp(-1/2 D^2_{ik})}{\sum_{k'=1}^{K} p_{k'} \mid S_{k'} \mid^{-1/2} \exp(-1/2 D^2_{ik'})}$$

where p_k is the prior probability of membership in population k. This latter expression represents the (posterior) probability of individual i belonging to population k. An individual is classified into that population from which the sample yields the largest value of P_{ik}. In this study, equal prior probabilities are used (that

is, $p_k = 1/K$) because it is not known whether the sample group sizes represent the proportions found in the population. The linear classification rule used is based on P_{ik} values determined as above except that the S_k matrices are replaced by S, the pooled sample (p x p) covariance matrix.

Theoretically, a quadratic classification rule should lead to higher cross validation classification hit rate accuracy than a linear classification rule when group covariance structures are different (Anderson, 1984, p. 235). However, Huberty and Curry (1978) found that a linear classification rule performed nearly as well as, or superior to, a quadratic rule in seven situations (the combined conditions of equal and unequal covariance matrices, and two and three criterion groups, for three sets of real data, using n_k/N as the value for p_k). The authors point out that fewer parameters need to be estimated with a linear rule (a pooled S matrix is used instead of separate groups S_k matrices), and thus greater across-sample stability might be expected (Michaelis, 1973, p. 230). Also, "the assumption of normality seems to be more critical for quadratic rules than linear rules" (Johnson & Wichern, 1992, p. 540).

Purpose

This study extends the findings of Huberty and Curry by offering a method for determining the superior classification rule for a specific data set regardless of covariance structure. In addition, a computer program that accomplishes the method is introduced and demonstrated.

Method

The data

Thirty three classification data sets varying in number of subjects, predictor variables, groups (two or three), and heterogeneity of covariance structure were employed to illustrate the method. To bolster validity, all data sets were taken from real classification studies. The sources were journal articles, paper presentations and research texts. No pathological distributional problems are known in any of the data sets; it is expected that they are much as one would find in typical classification studies.

Procedure

In comparing the predictive accuracy of the linear rule to that of the quadratic rule, "external" rather than "internal" results were considered. Results of an internal classification analysis are those obtained when measures for the individuals on whom the statistics were based are resubstituted to obtain the P_{ik} values. In an external classification analysis statistics based on one set of individuals are used in classifying "new" individuals. An external analysis is appropriate for making inferences about the discriminatory power of the predictors for a new set of data (Huberty, 1984).

External, or cross validated, hit-rate accuracy was estimated using the "leave-one-out" procedure. 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. This procedure has a relatively wide following in the discriminant

analysis literature (see, for example, Huberty, 1984; Huberty & Mourad, 1980; Lachenbruch, 1967; Mosteller & Tukey, 1968).

Separate group as well as total group proportions of correct classifications were compared for the linear and quadratic rules. McNemar's (1947) test for contrasting correlated proportions was used in the statistical comparisons between linear and quadratic models for the separate group and total sample proportions. This method was previously suggested for comparing full and reduced classification methods (Morris & Huberty, 1991), but is equally applicable in comparing linear and quadratic models. [See Looney (1988) for a method of comparing classification results of more than two models.] As the calculation of the McNemar correlated proportion statistic requires the joint distribution of "hits" and "misses" for both the linear and quadratic classification rule, no statistical package will accomplish the method. Therefore, a FORTRAN computer program was written to provide this information.

The Box test was used for testing the assumption of homogeneity of covariance structures. Not withstanding concerns over this test, one could argue that, theoretically, a quadratic classification rule is appropriate when the Box test indicates that the covariance structures are unequal.

Results and Discussion

For each of the data sets, Table 1 gives a short description, the number of subjects (N), the number of predictor variables (p), results of the Box test for homogeneity of covariance structures, the appropriate classification rule (quadratic

when the Box test suggests the assumption of equal covariance matrices is untenable), and a comparison of the performance of the linear and quadratic rules for each group separately and for the total sample. Performance of the two classification rules, displayed as the hit rate percent obtained by the p predictor variables, was compared via McNemar's test for contrasting correlated proportions.

As can be seen in Table 1, differences between the linear and quadratic rules in classifying the total sample were not statistically significant ($z < 2.58$, $p > .01$, two-tailed test), with the exception of data set 2. Here, the linear rule was judged appropriate by the Box test and yielded a significantly higher total hit rate. Differences between the two classification rules in separate group hit rates were statistically significant in nine of the 33 data sets: the linear rule outperformed the quadratic rule in data sets 3 and 4, where the linear rule was judged appropriate by the Box test, and in data sets 5, 7, 11, 13, 14, and 17, where the quadratic rule was judged appropriate by the Box test; the quadratic rule outperformed the linear rule in five situations where the Box test indicated the quadratic rule was appropriate (data sets 5, 11, 14, 16, and 17), but in no situation where the Box test indicated the linear rule was appropriate. These results are summarized in Table 2.

Although some researchers have urged caution in using anything but equal prior probabilities of group membership for classification (e.g., Lindeman, Merenda, & Gold, 1980, pp. 211-212), data sets with unequal numbers of subjects per group (data sets 1-4, 6, and 22-33) were tested using prior probabilities of n_k/N for the purpose of replicating Huberty & Curry's (1978) study. The results were identical to

the findings reported above for equal priors, with the following trivial exceptions: (1) in data sets 2, 3, and 4, the difference between the two models was no longer statistically significant; (2) in data set 3, the separate group hit rate for the quadratic rule was significantly higher than for the linear rule.

These results extend the findings of Huberty & Curry (1978) to a broader range of data sets. More important, however, is that a method is now available for comparing the performance of the two rules. The method will be helpful in determining when to use a quadratic rather than a linear classification rule to maximize classification accuracy for a specific data set in a predictive discriminant analysis.

If you would like a copy of the FORTRAN program that accomplishes the method, just send a returnable 5 1/4" or 3 1/2" diskette and diskette mailer to:

John D. Morris

Florida Atlantic University

Department of Educational Foundations and Technology

College of Education

P.O. Box 3091

Boca Raton, FL 33431-0991

References

Anderson, T. W. (1984). An introduction to multivariate statistical analysis (2nd ed.). New York: Wiley.

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

Huberty, C. J, & Curry, A. R. (1978). Linear versus quadratic multivariate classification. Multivariate Behavioral Research, 13, 237-245.

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

Johnson, R. A., & Wichern, D. W. (1992). Applied multivariate statistical analysis (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Lachenbruch, P. A. (1967). An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. Biometrics, 23, 639-645.

Lindeman, R. H., Merenda, P. F., & Gold, R. Z. (1980). Introduction to bivariate and multivariate analysis. Glencoe, IL: Scott, Foresman.

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

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

Michaelis, J. (1973). Simulation experiments with multiple group linear and quadratic discriminant analysis. In T. Cacoullos (Ed.), Discriminant analysis and applications. New York: Academic Press.

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

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.

Table 1

Data Set Description, Results of Box M Test for Equality of Covariance Matrices, and Comparison of Hit Rate Percents for Linear and Quadratic Models

#	Data Set Description	N	p	Results of Box M Test for Equal Covariance Matrices; (Appropriate Rule)	Rule Used	Hit Rate Percents GR 1	GR 2	GR 3	Total
1	Bisbey Data - Groups 1 & 2	116	13	$\chi^2_{91} = 1.0021, p = .4777$ (Linear)	Linear	89	89		89
					Quadratic	74	89		85
					McNemar's z	1.89	.00		1.51
2	Bisbey Data - Groups 1 & 3	72	13	$\chi^2_{91} = .9939, p = .5013$ (Linear)	Linear	94	100		97
					Quadratic	91	84		88
					McNemar's z	1.00	2.45		2.65
3	Bisbey Data - Groups 2 & 3	118	13	$\chi^2_{91} = 1.2131, p = .0929$ (Linear)	Linear	83	87		84
					Quadratic	91	68		84
					McNemar's z	-1.94	2.65		.00
4	Bisbey Data - Groups 1, 2 & 3	153	13	$\chi^2_{91} = 1.1421, p = .1134$ (Linear)	Linear	86	68	87	77
					Quadratic	74	80	68	76
					McNemar's z	1.41	-2.13	2.65	.16
5	Demographic # 1 - Body Char	279	8	$\chi^2_{36} = 5.4808, p < .0001$ (Quadratic)	Linear	61	55		58
					Quadratic	82	24		52
					McNemar's z	-4.13	5.26		1.49
6	Demographic # 2 - Body Char	279	8	$\chi^2_{36} = 6.6870, p < .0001$ (Quadratic)	Linear	83	82		82
					Quadratic	79	83		81
					McNemar's z	1.60	-.30		1.00

Table 1, cont.

Data Set Description, Results of Box M Test for Equality of Covariance Matrices, and Comparison of Hit Rate Percents for Linear and Quadratic Models

#	Data Set Description	N	p	Results of Box M Test for Equal Covariance Matrices; (Appropriate Rule)	Rule Used	Hit Rate Percents GR 1	GR 2	GR 3	Total
7	Demographic # 3 - Body Char	279	8	$\chi^2_{36} = 5.2724$, $p < .0001$ (Quadratic)	Linear	70	76		73
					Quadratic	55	82		68
					McNemar's z	4.31	-2.00		2.16
8	Block Data - Groups 1 & 2	78	4	$\chi^2_{10} = 2.2028$, $p = .0157$ (Quadratic)	Linear	60	76		68
					Quadratic	50	76		63
					McNemar's z	1.41	.00		1.15
9	Block Data - Groups 1 & 3	79	4	$\chi^2_{10} = 5.3857$, $p < .0001$ (Quadratic)	Linear	58	72		65
					Quadratic	50	87		68
					McNemar's z	1.73	-2.45		-1.00
10	Block Data - Groups 1 & 4	78	4	$\chi^2_{10} = 1.5500$, $p = .1163$ (Linear)	Linear	58	63		60
					Quadratic	50	66		58
					McNemar's z	1.13	-.30		.47
11	Block Data - Groups 2 & 3	76	4	$\chi^2_{10} = 4.5542$, $p < .0001$ (Quadratic)	Linear	57	54		55
					Quadratic	19	85		53
					McNemar's z	3.74	-3.46		.39
12	Block Data - Groups 2 & 4	75	4	$\chi^2_{10} = 1.2033$, $p = .2838$ (Linear)	Linear	62	55		59
					Quadratic	65	50		57
					McNemar's z	-.58	.63		.28

Table 1, cont.

Data Set Description, Results of Box M Test for Equality of Covariance Matrices, and Comparison of Hit Rate Percents for Linear and Quadratic Models

#	Data Set Description	N	p	Results of Box M Test for Equal Covariance Matrices; (Appropriate Rule)	Rule Used	Hit Rate Percents GR 1	GR 2	GR 3	Total
13	Block Data - Groups 3 & 4	76	4	$\chi^2_{10} = 4.3098, p < .0001$ (Quadratic)	Linear	74	63		68
					Quadratic	82	40		61
					McNemar's z	-1.73	2.71		1.60
14	Block Data - Groups 1, 2, & 3	116	4	$\chi^2_{20} = 3.7144, p < .0001$ (Quadratic)	Linear	48	50	34	44
					Quadratic	38	11	82	43
					McNemar's z	1.63	3.87	-4.24	.16
15	Block Data - Groups 1, 2, & 4	116	4	$\chi^2_{20} = 1.7091, p = .0265$ (Quadratic)	Linear	53	53	37	47
					Quadratic	40	58	42	47
					McNemar's z	1.89	-1.00	-.82	.24
16	Block Data - Groups 1, 3, & 4	116	4	$\chi^2_{20} = 3.4370, p < .0001$ (Quadratic)	Linear	40	55	47	47
					Quadratic	33	79	24	45
					McNemar's z	1.34	-3.00	2.32	.56
17	Block Data - Groups 2, 3, & 4	114	4	$\chi^2_{20} = 2.9833, p = .0001$ (Quadratic)	Linear	40	50	40	43
					Quadratic	08	79	21	36
					McNemar's z	3.00	-3.05	2.11	1.26
18	Fisher Data - Groups 1 & 2	100	4	$\chi^2_{10} = 5.0455, p < .0001$ (Quadratic)	Linear	100	100		100
					Quadratic	100	100		100
					McNemar's z	.00	.00		.00

Table 1, cont.

Data Set Description, Results of Box M Test for Equality of Covariance Matrices, and Comparison of Hit Rate Percents for Linear and Quadratic Models

#	Data Set Description	N	p	Results of Box M Test for Equal Covariance Matrices; (Appropriate Rule)	Rule Used	Hit Rate Percents GR 1	GR 2	GR 3	Total
19	Fisher Data - Groups 1 & 3	100	4	$\chi^2_{10} = 6.9057, p < .0001$ (Quadratic)	Linear	100	100		100
					Quadratic	100	100		100
					McNemar's z	.00	.00		.00
20	Fisher Data - Groups 2 & 3	100	4	$\chi^2_{10} = .7148, p = .7125$ (Linear)	Linear	92	94		93
					Quadratic	92	92		92
					McNemar's z	.00	1.00		1.00
21	Fisher Data - Groups 1, 2, & 3	150	4	$\chi^2_{20} = 3.7663, p < .0001$ (Quadratic)	Linear	100	94	94	96
					Quadratic	100	92	92	95
					McNemar's z	.00	1.00	1.00	1.41
22	Rulon Data - Groups 1 & 2	178	4	$\chi^2_{10} = 4.9003, p < .0001$ (Quadratic)	Linear	84	79		81
					Quadratic	82	80		81
					McNemar's z	.38	-.38		.00
23	Rulon Data - Groups 1 & 3	151	4	$\chi^2_{10} = 3.4973, p = .0003$ (Quadratic)	Linear	94	91		93
					Quadratic	93	94		93
					McNemar's z	1.00	-1.41		-.58
24	Rulon Data - Groups 2 & 3	159	4	$\chi^2_{10} = 3.4962, p = .0003$ (Quadratic)	Linear	85	80		83
					Quadratic	83	77		81
					McNemar's z	.82	1.00		1.26

Table 1, cont.

Data Set Description, Results of Box M Test for Equality of Covariance Matrices, and Comparison of Hit Rate Percents for Linear and Quadratic Models

#	Data Set Description	N	p	Results of Box M Test for Equal Covariance Matrices; (Appropriate Rule)	Rule Used	Hit Rate Percents GR 1	GR 2	GR 3	Total
25	Rulon Data - Groups 1, 2, & 3	244	4	$\chi^2_{20} = 3.8489, p < .0001$ (Quadratic)	Linear	81	67	77	75
					Quadratic	79	66	74	73
					McNemar's z	.71	.33	1.00	1.09
26	Talent Data - Groups 1 & 3	116	14	$\chi^2_{105} = .9401, p = .6493$ (Linear)	Linear	64	42		58
					Quadratic	68	33		58
					McNemar's z	-.58	.77		.00
27	Talent Data - Groups 1 & 5	177	14	$\chi^2_{105} = 1.5086, p = .0014$ (Quadratic)	Linear	74	76		75
					Quadratic	76	70		73
					McNemar's z	-.43	1.39		.51
28	Talent Data - Groups 3 & 5	127	14	$\chi^2_{105} = 1.1086, p = .2238$ (Linear)	Linear	79	77		77
					Quadratic	49	83		74
					McNemar's z	2.50	-1.60		.73
29	Talent Data - Groups 1, 3, & 5	210	14	$\chi^2_{210} = 1.2804, p = .0086$ (Quadratic)	Linear	47	42	73	58
					Quadratic	57	30	66	57
					McNemar's z	-1.57	1.15	1.61	.40
30	Warncke Data - Groups 1 & 2	112	10	$\chi^2_{55} = 1.0593, p = .3611$ (Linear)	Linear	51	45		48
					Quadratic	55	38		48
					McNemar's z	-.73	.90		.00

Table 1, cont.

Data Set Description, Results of Box M Test for Equality of Covariance Matrices, and Comparison of Hit Rate Percents for Linear and Quadratic Models

#	Data Set Description	N	p	Results of Box M Test for Equal Covariance Matrices; (Appropriate Rule)	Rule Used	Hit Rate Percents GR 1	GR 2	GR 3	Total
31	Warncke Data - Groups 1 & 3	105	10	$\chi^2_{55} = 1.5335, p = .0086$ (Quadratic)	Linear	62	50		57
					Quadratic	63	38		53
					McNemar's z	-.21	1.67		.71
32	Warncke Data - Groups 2 & 3	87	10	$\chi^2_{55} = 1.2556, p = .1039$ (Linear)	Linear	45	35		40
					Quadratic	60	43		52
					McNemar's z	-1.70	-.65		-1.62
33	Warncke Data - Groups 1, 2, & 3	152	10	$\chi^2_{110} = 1.2649, p = .0404$ (Quadratic)	Linear	43	19	25	31
					Quadratic	39	26	18	29
					McNemar's z	.60	-.90	.90	.44

Table 2

Summary of Linear vs. Quadratic Classification Model Superiority by Condition (Equal or Unequal Covariance Matrices)

Equality of Covariance Matrices Based on Box M Test	Appropriate Rule Based on Box M Test	# of Data Sets in which Linear Model Hit Rate was Superior*		# of Data Sets in which Quadratic Model Hit Rate was Superior*	
		Separate Group	Total	Separate Group	Total
Equal	Linear (11 data sets)	2	1	0	0
Unequal	Quadratic (22 data sets)	6	0	5	0

* $z > 2.58$, $p < .01$