DOCUMENT RESUME

ED 441 029 TM 030 827

AUTHOR	McGee, Jennifer
TITLE	Using Linear versus Quadratic Rules in Predictive and Descriptive Discriminant Analysis.
PUB DATE	2000-04-21
NOTE	13p.; Paper presented at the Annual Meeting of the Southwestern Psychological Association (Dallas, TX, April 21, 2000).
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Discriminant Analysis; Multivariate Analysis
IDENTIFIERS	Covariance Matrices

ABSTRACT

Both predictive discriminant analysis (PDA) and descriptive discriminant analysis (DDA) require a decision to pool group covariance matrices, or alternatively, to retain separate group covariance matrices when the group covariance matrices are too dissimilar to pool. Pooling the group covariance matrices involves the so-called "linear" rule, generally preferred in predictive and descriptive analysis. Retaining separate group covariance matrices invokes the "quadratic" rule, resulting in a higher hit rate in PDA and a lower lambda in DDA. However, the quadratic rule is influenced by unique sampling error variance, making the generalizability of quadratic results suspect. (Contains 12 references.) (Author/SLD)

Using Linear Versus Quadratic Rules in Predictive and Descriptive Discriminant Analysis

Jennifer McGee

Texas A&M University

Paper presented at the annual meeting of the Southwestern Psychological Association, Dallas, April 21, 2000.

Abstract

Both predictive discriminant analysis (PDA) and descriptive discriminant analysis (DDA) require a decision to pool group covariance matrices, or alternatively to retain separate group covariance matrices when the group covariance matrices are too dissimilar to pool together. Pooling the group covariance matrices invokes the so-called "linear" rule, generally preferred in predictive and descriptive analysis. Retaining separate group covariance matrices invokes the "quadratic" rule, resulting in a higher hit rate in PDA and a lower lambda in DDA. However, the quadratic rule is influenced by unique sampling error variance, therefore the generalizability of quadratic results is suspect.

Multivariate statistics provide adept researchers with methods that (a) control experimentwise error rate, and (b) honor a complex reality with multiple causes and multiple effects (Fish, 1988; Thompson, 1997). According to empirical studies (Emmons, Stallings & Layne, 1990), "In the last 20 years, the use of multivariate statistics has become commonplace" (Grimm, & Yarnold, 1995, p. vii). However, with discriminant analysis, common use does not ensure responsible use. Responsible use of discriminant analysis depends on distinguishing between predictive (PDA) and descriptive (DDA) discriminant analysis (Huberty, 1994; Huberty & Barton, 1989; Huberty & Wisenbaker, 1992).

Broadly speaking, discriminant analysis either predicts <u>or</u> describes group membership. As Huberty and Lowman (1997) have noted: "Simply put, we have different analyses (PDA and DDA) for different questions; one is for prediction of group membership [PDA] and one is for description of grouping variable effects [DDA]" (p. 759). Even popular statistical software packages such as SAS and SPSS muddle the PDA/DDA distinction, providing misleading or incorrect information (Huberty & Lowman, 1997).

In order to use either PDA or DDA, a researcher must decide whether to pool group covariance matrices, or retain separate group covariance matrices when the group covariance matrices are too dissimilar to pool together. Pooling group covariance

matrices invokes the so-called "linear" rule, generally preferred in predictive and descriptive analysis. Retaining separate group covariance matrices invokes the "quadratic" rule, resulting in a higher hit rate in PDA and a lower lambda in DDA. However, the quadratic rule is influenced by unique sampling error variance, making the generalizability of quadratic rules results suspect (Huberty, 1994). While using separate group covariance matrices improves PDA and DDA results for an individual study, the results are unlikely to replicate in future studies.

Pooling Covariance Matrices

Pooling group covariance matrices invokes the linear rule. While linear rule results may be less exciting and more conservative (i.e., have lower DDA lambdas or lower PDA hit rates) than quadratic rule results, linear rule results are not as susceptible to unique sampling error variance. However, pooling variance "... **is legitimate if, and only if, the variabilities of the scores in each group are roughly the same**" (Haase & Thompson, 1996, p. 6). Assessing whether group covariance matrices are "roughly the same" is usually accomplished with statistical significance testing (Huberty & Lowman, 1997).

In the context of evaluating homogeneity of variance, Huberty (1994) noted three major problems with statistical

testing procedures: (a) the methods are extremely sensitive "to relatively small matrix discrepancies;" (b)" the degrees of freedom used in the chi square and F-tests are quite large even for modest-sized data sets;" and (c) the statistical tests such as Box's M are extremely sensitive "to lack of multivariate normality of the outcome variables vectors" (p. 70). Essentially, tests for homogeneity of variance are statistically powerful, have many, many degrees of freedom, and are influenced by lack of multivariate normality.

Therefore, as in the univariate world, "common sense may be the best guide to evaluating whether the homogeneity of variance assumption has been met" (Haase & Thompson, 1996, p. 11). A thoughtful researcher may use log determinant values that "provide an indication of which groups' covariance matrices differ most" (SPSS, Inc., 1998, p. 263). Furthermore, both SPSS and SAS outputs provide log determinants. Also, box plots and within group scatterplots may be useful in assessing homogeneity of group covariance matrices (SPSS, Inc., 1998).

Comparing Distances

Homogeneity of group covariance is critical, as Huberty (1994) noted:

> The basic requirement in comparing distances involving measures on two (or more) variables is that the same metric is used in computing the distances. One way

> this is assured is, of course, if all standard deviations or variances are equal…. If this is not the case, the unequal variances must be 'taken into consideration.' This is accomplished by dividing the measures by the corresponding standard deviation. (p. 42)

Comparing distances that are not in the same metric is as nonsensical as comparing different currencies.

For example, three types of distances are investigated in classic one-way analysis of variance. One, the distance of individual scores from the grand mean is the total sum of squares. Two, the distance of group means from the grand mean is the sum of squares between groups. And, three, the distance of individual scores from the group means pooled together is the sum of squares within groups.

Again, as in predictive and descriptive discriminant analysis, pooling variance for the sum of squares within groups is "**legitimate if, and only if, the variabilities of the scores in each group are roughly the same**" (Haase & Thompson, 1996, p. 6). However, while in ANOVA pooling variance is not optional, in PDA and DDA, the linear and quadratic rules empower the researcher to determine whether or not group covariance matrices are similar enough to justify implementing the preferred linear rule.

In multivariate methods such as predictive and descriptive discriminant analyses, the Mahalanobis distances among individual points, distances among group centroids, and distances among points and centroids are explored and compared (Huberty, 1994). In addition to answering different research questions, another distinction between PDA and DDA is the type of distance used in the respective analyses.

Distances Between Points

Huberty (1994) explained that the Mahalanobis distance between points is determined by:

$$\Delta^2_{ug} = (X_A - X_B)'\Sigma^{-1}(X_A - X_B)$$

In this equation, Δ^2_{AB} is a "squared generalized index between Pont A (defined by the column vector X_A) and Point B (defined by the column vector X_B)" (p. 43). The influence of unequal variance is "taken into consideration by using the inverse of the population covariance matrix Σ^{-1}" (Huberty, 1994, p. 43).

Distances Among/Between Centroids

The distance among or between group centroids is the distance of interest in descriptive discriminant analysis (Huberty, 1994). Group separation is determined by:

$$\Delta^2_{12} = [(\mu_1 - \mu_2)'\Sigma_g^{-1}(\mu_1 - \mu_2)]^{1/2}$$

In the formula for distances among group centroids, "Σ is the covariance matrix common to the two populations; that is, the

covariance matrices are assumed to be equal" (Huberty, 1994, p. 44). Using the covariance matrix "common to the two populations" mitigates influence the of unique sample variance. In descriptive discriminant analysis, "the grouping variable plays the role of a predictor variable, and the p response variables are outcome variables" (Huberty & Lowman, 1997, p. 759). From the formula for distance among/between group centroids, it is evident that the focus of DDA is the distance between/among group centroids.

Distances Among Points and Centroids

The distances among points and centroids is the type of distance on which "emphasis is given in predictive discriminant analysis" (Huberty, 1994, p. 44). Huberty (1994) noted that the distance index among individual points and group centroids is calculated by:

$$\Delta^2_{ug} = [(X_u - \mu_g)' \Sigma_g^{-1} (X_u - \mu_g)]^{½}$$

"where Σ_g is the covariance matrix for population g" (Huberty, 1994, p. 44). In PDA, group membership is the outcome variable, and the response variables are the predictors (Huberty & Lowman, 1997). From the formula for distance among points and group centroids, it is evident that the focus of PDA is the distance between points and group means.

Discussion

The distinction between PDA and DDA is fundamental for responsible use of discriminant analysis. Typically, descriptive discriminant analysis is used after group membership is determined, as a post hoc method for describing how predictor variables reflect group membership (Huberty & Lowman, 1997). Because the response variables are the focus of DDA, standardized discriminant functions (multiplicative weights), structure coefficients, should be reported.

For meaningful interpretation of descriptive discriminant analysis SAS or SPSS output, the researcher should consult MANOVA results and lambda, an r^2 type-effect size. However, r^2 type-effect sizes are "uncorrected for the positive bias in all variance-accounted-for effect sizes (due to **ALL** analyses being correlational and capitalizing on sampling error variance with the sampled data's total variance that does not exist anywhere except in this particular sample)" (Thompson, 1997, p. 1). Therefore, using a quadratic rule (separate group covariance matrices) in DDA, then consulting lambda, compounds the influence of unique sample variance.

In order to interpret predictive discriminant analysis results, the adept researcher should consult the hit rate/classification rate and linear classification functions (LCF). However, when reporting PDA results, only the hit rate

should be reported, because the focus of PDA is the accuracy of group classification, not how well the predictors explain group membership. Thus, weights and structure coefficients are not relevant. In PDA, the most important response variable is the variable that most hurts hit rate when that variable is not used in the analysis (Thompson, 1998).

Both PDA and DDA involve a decision to retain separate group covariance matrices, the quadratic rule, or to pool group covariance matrices, the linear rule. This decision is analogous to assessing homogeneity of variance in analysis of variance. The quadratic rule typically produces a higher hit rate in predictive discriminant analysis and a lower lambda in descriptive discriminant analysis. However, the quadratic rule is influenced by unique sample variance, therefore the generalizability of quadratic rule results is suspect. Therefore, invoking the linear rule results is generally preferred.

REFERENCES

Emmons, N.J., Stallings, W.m., & Layne, B.H. (1990, April). Statistical methods used in *American Educational Research Journal, Journal of Educational Psychology,* and *Sociology of Education* from1972 through 1987. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA. (ERIC Document Reproduction Service No. ED 319 797)

Fish, L.J. (1988). Why multivariate methods are usually vital. Measurement and Evaluation in Counseling and Development, 21, 130-137.

Grimm, L.G., & Yarnold, P.R. (Eds.). (1995). Reading and understanding multivariate statistics. Washington, DC: American Psychological Association.

Haase, T., & Thompson, B. (1992, January). The homogeneity of variance assumption in ANOVA: what it is and why it is required. Paper presented at the annual meeting of the Southwest Educational Research Association, Houston, TX.

Huberty, C. (1994). Applied discriminant analysis (2nd ed.). Hillsdale, NJ: Erlbaum.

Huberty, C. J., & Barton, R.M. (1989). An introduction to discriminant analysis. Measurement and Evaluation in Counseling and Development, 22, 158-168.

Huberty, C.J., & Lowman, L.L. (1997). Discriminant analysis via statistical packages. Educational and Psychological Measurement, 57, 759-784.

Huberty, C., & Wisenbaker, J. (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.

SPSS, Inc. (1998). SPSS base 8.0 applications guide, Chicago, IL: SPSS, Inc.

Thompson, B. (1994, February). Why multivariate methods are usually vital in research: Some basic concepts. Paper presented as a Featured Speaker at the biennial meeting of the Southwestern Society for Research in Human Development, Austin, TX. (ERIC Document Reproduction Service No. ED 367 687)

Thompson, B. (1997). The two-stage hierarchical, contingency approach to interpreting results from GLM Analyses. http://acs.tamu.edu/~bbt6147/

Thompson, B. (1998, April). Five methodology errors in educational research: The pantheon of statistical significance and other faux pas. Invited address presented at the annual meeting of the American Educational Research Association, San Diego, (ERIC Document Reproduction Service, No. ED419023).

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: USING LINEAR VERSUS QUADRATIC RULES IN PREDICTIVE AND DESCRIPTIVE DISCRIMINANT ANALYSIS

Author(s): JENNIFER McGEE

Corporate Source:

Publication Date: 4/21/00

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1	Level 2A	Level 2B
XXXX		
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please

Signature: Jennifer McGee

Printed Name/Position/Title: Jennifer McGee

Organization/Address: TAMU DEPT EDUC PSYC
COLLEGE STATION, TX 77843-4225

Telephone: 979/845-1335

FAX:

E-Mail Address:

Date: 4/19/00

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com