DOCUMENT RESUME

ED 441 869		TM 031 211
AUTHOR	Tanguma, Jesus	
TITLE	Variable Selection Strategies in Discriminate Analysis.	
PUB DATE	2000-04-00	
NOTE	21p.; Paper presented at the Annual Meeting of the Southwestern Psychological Association (Dallas, TX, April 20-22, 2000).	
PUB TYPE	Reports - Descriptive (141) -- Speeches/Meeting Papers (150)	
EDRS PRICE	MF01/PC01 Plus Postage.	
DESCRIPTORS	*Discriminant Analysis; *Selection	
IDENTIFIERS	Stepwise Regression; *Variables	

ABSTRACT

This paper presents three variable selection strategies in discriminate analysis (all variables in the model, use of stepwise methods, and all possible subsets). All three methods are illustrated through examples. Although the all variables in the model and the stepwise methods are the most widely used, B. Thompson (1996) and C. Huberty (1994) strongly oppose their use. On the other hand, Huberty states, "If one is into basing predictor selection on the data on hand, the recommendation here is to use the all-possible subsets approach" (C. Huberty, 1994, p. 125). (Contains 12 tables and 6 references.) (Author/SLD)

Running head: VARIABLE SELECTION IN DA

Variable Selection Strategies in Discriminant Analysis

Jesus Tanguma

Texas A & M University

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

Abstract

The present paper presents three variable selection strategies in discriminant analysis (all variables in the model, use of stepwise methods, and all possible subsets). All three methods are illustrated by means of an example. Although the all variables in the model and the stepwise methods are the most widely used, Thompson (1996) and Huberty (1994) strongly oppose their use. On the other hand, Huberty states that "if one is into basing predictor selection on the data on hand, the recommendation here is to use the all-possible subsets approach" (p. 125).

Variable Selection Strategies in Discriminant Analysis

Researchers often gather data on predictors, which they believe to be good discriminators. This may well be the case, for example, when the researchers conduct a preliminary investigation trying to discover useful discriminating variables. Thus, they might ask themselves questions such as "(1) are all variables really necessary for effective discrimination and (2) which variables are the best discriminators?" (Johnson, 1998, p. 245). In lieu of these questions, researchers "seek a subset of the predictors (i.e., to delete some "poor" predictors) to determine a rule that will yield a high degree of classification precision as well as predictive accuracy" (Huberty, 1994, p. 117).

In regression analysis, the most frequently used variable selection methods are the so-called stepwise methods. However, Thompson (1996) has repeatedly stated that these methods are inherently flawed and should not be used for this or other purposes. In discriminant analysis, methods have been developed to assist the researcher in deciding which discriminators to select.

While the applied researcher may not desire to spend too much time and effort in figuring out the mathematics behind a discriminant analysis, the researcher may desire to get a conceptual understanding of the Statistical Package for the Social Sciences (SPSS) output. Thus, detailed discussions of pertinent output will be provided throughout the paper.

The purpose of the present paper is to discuss three variable selection strategies in discriminant analysis (DA). These three strategies are (a) all variables in the model; (b) stepwise methods; and (c) all-possible-subsets. To illustrate how to apply these three strategies, data collected by Cmajdalka and Cuellar (1998) will be analyzed using SPSS 9.0.

Three Selection Strategies

All Variables in the Model Strategy

The first variable selection strategy presented here may very well not be considered as such. This is because, as the name implies, all variables are in the model. However, the results of subjecting the data to a DA using all available variables may assist researchers who "are trying to discover useful discriminating predictors" (Klecka, 1980, p. 52).

The results of testing the equality of group means are presented in Table 1. By visually inspecting the second, third, and the sixth columns, the researcher may conclude that there are highly significant differences between the groups means for Process3 and Process4. The second column presents the Wilks' lambda values. Wilks' lambda is defined as "the ratio of the within-groups sum of squares to the total sum of squares" (SPSS Base 9.0 Applications Guide, p. 252). The values of the lambda range from zero to one. The smaller the value of the lambda is, the stronger the group differences are.

Insert Table 1 About Here

As noted by Klecka (1980), "the standardized coefficients are helpful, because we can use them to determine which variables contribute most to determining scores on the function" (p. 29). Thus, the researcher may look at the standardized canonical discriminant function coefficients when studying the usefulness of each variable in the discriminant function. To do so, the researcher first takes the absolute value of each coefficient and then compares the coefficients. The larger the coefficient, "the greater that variable's contribution" (Klecka, 1980, p. 30). The standardized canonical discriminant function coefficients are presented on

Table 2. Such values indicate that Process3 and Process4 may be considered as useful variables.

Insert Table 2 About Here

Although the standardized coefficients are helpful in determining the variable's contribution in calculating the discriminant score, they have a serious limitation. Namely,

> If two variables share the same discriminating information (i.e., if they are highly correlated), they must share their contribution to the score even if that joint contribution is very important. Consequently, their standardized coefficients may be smaller than when only one of the variables is used. Or, the standardized coefficients might be larger but with opposite signs, so that the contribution of one is partially cancelled by the opposite contribution of the other. This is because the standardized coefficients take into consideration the simultaneous contributions of all other variables. (Klecka, 1980, p. 33)

The structure coefficients (bivariate correlations), on the other hand, are not affected by relationships with other variables. Thus, "the structure coefficients are a better guide to the meaning of the canonical discriminant functions than the standardized coefficients are" (Klecka, 1980, p. 34). Once again, Process3 and Process4 appear to be useful variables. From Table 3, it can be readily seen that Process3 has the largest correlation with the canonical variable scores.

Insert Table 3 About Here

The results on Table 4 are indicative of the degree of success of the classification for the data on hand. SPSS obtains these results by counting the number of processes correctly classified as well as the number of processes incorrectly classified. Thus, 45 (97.8%) of the 46 who passed are correctly classified and 1 (2.2%) is incorrectly classified. Similarly, of the 28 who failed, 23 (82.1%) are correctly classified and 5 (17.9%) are incorrectly classified. However, this procedure "produces an overly optimistic estimation of the success of the classification" (SPSS Base 9.0 Applications Guide, p. 260). To alleviate this problem, SPSS provides a leave-one-out cross-validation method. According to Johnson (1998), "these estimates have been shown to be nearly unbiased estimates of the true probabilities of correct and incorrect classification" (p. 221). The results of the cross-validation are presented on Table 4. Thus, for the data on hand, using all the variables in the model 91.9% of the original grouped cases were correctly classified. Similarly, 90.5% of the cross-validated grouped cases were correctly classified.

Insert Table 4 About Here

Stepwise methods

The stepwise methods are a combination of the forward selection method and the backward selection method. When using the backward selection process, all the variables are initially included in the model. As the analysis progresses, any predictor that does not contribute to the model is deleted. The forward selection process, on the other hand, begins with no variables in the model. The model is built by entering predictors "one at a time until the increase in R^2 is no longer statistically significant or until all predictor variables have been included in the model" (Hinkle, Wiersma, & Jurs, 1994, p. 473). The basic difference between the forward selection process and the stepwise process is that the stepwise process, before entering a new predictor, checks to see if all the predictors already in the model remain significant. Thus, if a previously selected predictor is no longer useful, the procedure will drop that predictor. On the other hand, using the forward selection method, once a predictor enters the model, it remains there.

As in the case of all the variables in the model strategy, SPSS prints out a table where the equality of the group means are tested. This table is identical as the one before (Table 1) and is thus not presented again.

Stepwise procedures need a mechanism for controlling the entry or removal of predictor variables from the discriminant function. lambda is one such mechanism and is the one used in this paper. Other methods for controlling the entry or removal of predictor variables from the discriminant function are (a) Mahalanobis distance, (b) smallest F ratio, (c) Rao's V (also known as Lawley-Hotelling trace), and (d) sum of unexplained variance (SPSS Base 9.0 Applications Guide, pp. 268-269). Deciding which method to use is not always clear. However, as pointed out by Klecka (1980) "the end result will often be the same regardless of the criterion used, but it is not always the case" (p. 54).

In order for the researcher to understand how stepwise selects variables, the researcher needs to scan back and forth several tables. For example, after subjecting the data on hand to a discriminant analysis using the stepwise methods and using Wilks' lambda as the entry/removal criterion, the following tables were produced. As mentioned before, at the beginning of the analysis there are no variables in the analysis. Thus, Step 0 in Table 5 indicates that none of the five variables are in the analysis. However, at Step 1 only four

(Process1, Process2, Process 4, and Process5) variables remain not in the analysis. Thus, Process3 has already entered the analysis. But, why did Process3 get selected as the first variable to enter the analysis? Because it had the largest F (smallest Wilks' lambda) to enter. Of the four variables that remain out of the analysis, Process4 has the largest F (smallest Wilks' lambda) to enter. Thus, it is entered next into the analysis. The order in which the variables are entered/removed is presented in Table 7. After entering Proces4 into the analysis, SPSS computes, again, another test of significance. This time no F values meet the criteria for entering into the analysis, see Step 2 in Table 5. Thus, no more variables are added to the model. In other words, the stepwise procedure selected a model with only Process3 and Process4 as the variables in the model, see Table 7. Moreover, the structure coefficients also suggest that Process3 and Process4 are the variables to use in the analysis.

Insert Tables 6, 7, and 8 About Here

The classification results obtained by using the stepwise procedure are presented in Table 8. Based on these results, the researcher may argue that a two-predictor (Process3 and Process4) model produces classification precision and predictive accuracy as high as those produced by the five-predictor model.

Insert Table 8 About Here

Although the stepwise procedures are widely used, Thompson (1996) and Huberty (1994) strongly oppose their use. Some of the problems with the stepwise procedures are that (a) not all variables selected may be needed, and (b) not all selected variables may actually be good discriminators. Therefore, according to Johnson (1998) " the results of any variable selection procedure must be taken with a grain of salt" (p. 248).

All-Possible Subsets Approach

The all-possible subsets approach, as the name implies, analyzes the data one-predictor at a time, two-predictors at a time, and so on. Thus, as the number of predictors increases, so does the number of analyses. In fact, "for p predictors, a total of $2^p - 1$ predictor subsets would need to be assessed" (Huberty, 1994, p. 122). For example, when there are four predictors there would be $2^4 - 1 = 15$ predictor subsets to be examined. There will be (a) four predictor subsets each containing one predictor only; (b) six predictor subsets each containing two predictors only; (c) four predictor subsets each containing three predictors only; and (d) there will be one predictor subset containing all four predictors. All-possible subsets, when the number of predictors $p = 4$, are listed in Table 9.

Insert Table 9 About Here

The data set being analyzed to illustrate the different variable selection strategies in discriminant analysis consisted of five predictors. Thus, there were $2^5 - 1 = 31$ different analyses. Each analysis was separately run and its output was carefully examined. However, due to space limitations, only selected portions of the output will be reproduced here.

When the model was run containing only one predictor, the values of the standardized canonical discriminant function coefficients were all equal to one. Similar results were found for the structure coefficients. However, as more predictors were included in the model, the values of the standardized canonical discriminant function coefficients as well as those for the structure coefficients varied across the subsets. The mean value of the standardized canonical discriminant function coefficients for each predictor, averaged over all-possible subsets, is presented in Table 10. As can be readily seen from Table 10, Process3 is contributing the most to the discriminant functions. Process4 follows closely as a good predictor. However, before making any decisions, the structure coefficients should be carefully examined. The mean value of the structure coefficients for each predictor, averaged over all-possible subsets, is presented in Table 11. As with the standardized canonical discriminant function coefficients, the structure coefficients suggest that Process3 is the predictor that contributes the most to the discriminant function. Again, Process3 is closely followed by Process4 as another useful predictor.

Insert Tables 10 and 11 About Here

To determine the degree of success of the classification for the data being analyzed, the researcher may examine the classification results table produced by SPSS. A summary of all 31 classification results tables is presented in Table 12. Such table displays the percentage of original grouped cases correctly classified as well as the percentage of cross-validated grouped cases correctly classified. As evidenced by inspecting Table 12, the best (highest percentages) results were achieved whenever Process3 and Process4 were in the model. In other words, whenever a given subset of predictors included Process3 and Process4, classification was at its highest. For example, subsets (a) Process1, Process2, Process3, Process4; (b) Process2, Process3, Process4; (c) Process1, Process3, Process4; and (d) Process3, Process4 all produced 91.9% of original grouped cases correctly classified as well as 91.9% of cross-validated grouped cases correctly classified. Thus, indicating that there is no loss in the classification precision or the predictive accuracy when going from five down to two predictors in the model. Moreover, it is easier to explain a two-predictor model than a three-, four-, or five-predictor model. Consequently, the researcher may argue that, based on the data on hand, a two-predictor (Process3 and Process4) model is the best model.

Insert Table 12 About Here

A problem when using the all-possible subsets approach is that "in those occasional cases when the pool of potential X variables contains 40 to 60 or even more variables, use of a "best" subsets algorithm may not be feasible" (Neter, Kutner, Nachtsheim, &Wasserman, 1996, p. 347). In such situations, one of the automatic selection procedures may need to be employed. A second problem with the all-possible subsets approach is that such an "analysis may be criticized as one that milks the data" (Huberty, 1994, p. 126). This follows from the fact that the researcher actually gets to see the results of combining all the variables in all possible ways.

Conclusion

Three variable selection strategies commonly used in discriminant analysis (DA) have been discussed. All three strategies were illustrated by means of analyzing a data set collected by Cmajdalka and Cuellar (1998). Although only the stepwise methods explicitly select the predictor variables to be used, the other two strategies implicitly suggest which predictor variables to use. In other words, by running a DA using all the variables in the model, the researcher may argue, based on their high structure coefficient values, that Process3 and Process4 are the most contributing predictors. When the data was subjected to a DA using the stepwise procedure, the algorithm selected, based on their Wilks' lambda values, Process3 and Process4 as the most contributing predictors. When the all-possible subsets approach was used to analyze the data, four different subsets produced the highest percentages of original and cross-validated grouped cases correctly classified. Of the four subsets, one consisted of three predictors, two consisted of three predictors, and one consisted of two predictors. Since it is easier to explain a two-predictor model than a three- or a four-predictor model, the two-predictor model was chosen as the best model for the data on hand. Moreover, the two predictors (Process3 and Process4) selected were the same ones that the all variables in the model and the stepwise method selected. Thus, for the data on hand, all three strategies suggested the same predictors, Process3 and Process4.

References

Hinkle, D.E., Wiersma, W., & Jurs, S.G. (1994). Applied statistics for the behavioral sciences (3rd ed.). Boston: Houghton Mifflin Company.

Huberty, C.J. (1994). Applied discriminant analysis. New York: John Wiley & Sons, INC.

Johnson, D.E. (1998). Applied multivariate methods for data analysts. Pacific Grove: Duxbury Press.

Klecka, W.R. (1980). Discriminant analysis. Beverly Hills, CA: Sage Publications.

Thompson, B. (1996). Canonical correlation analysis. In L. Grimm & P. Yarnold (Eds), Reading and understanding multivariate statistics, 2. Washington, DC: American Psychological Association.

Neter, J., Kutner, M.H., Nachtsheim, C.J., & Wasserman, W. (1996). Applied linear statistical models (4th ed.). Chicago, IL: Irwin.

Table1. Tests of Equality of Group Means

	Lambda	F	df1	df2	Sig.
PROCESS1	.890	8.910	1	72	.004
PROCESS2	.971	2.142	1	72	.148
PROCESS3	.432	94.547	1	72	.000
PROCESS4	.706	30.040	1	72	.000
PROCESS5	.932	5.273	1	72	.025

Table 2. Standardized Canonical Discriminant Function Coefficients

	Function
	1
PROCESS1	.306
PROCESS2	.157
PROCESS3	.855
PROCESS4	.458
PROCESS5	-.130

Table 3. Structure Matrix

	Function
	1
Process3	.834
Process4	.470
Process1	.256
Process5	.197
Process2	.125

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions Variables ordered by absolute size of correlation within function.

Table 4. Classification Results

			Predicted Group Membership		Total
		SUCCESS	pass	fail	
Original	Count	pass	45	1	46
		fail	5	23	28
	%	pass	97.8	2.2	100.0
		fail	17.9	82.1	100.0
Cross-validated	Count	pass	45	1	46
		fail	6	22	28
	%	pass	97.8	2.2	100.0
		fail	21.4	78.6	100.0

a Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.
b 91.9% of original grouped cases correctly classified.
c 90.5% of cross-validated grouped cases correctly classified.

Table 5. Variables Not in the Analysis

Step		Tolerance	Min Tolerance	F to Enter	Wilks' Lambda
0	Process1	1.000	1.000	8.910	.890
	Process2	1.000	1.000	2.142	.971
	Process3	1.000	1.000	94.547	.432
	Process4	1.000	1.000	30.040	.706
	Process5	1.000	1.000	5.273	.932
1	Process1	.994	.994	5.902	.399
	Process2	.995	.995	1.944	.421
	Process4	.999	.999	11.323	.373
	Process5	.999	.999	1.819	.422
2	Process1	.983	.983	3.671	.354
	Process2	.986	.986	.996	.368
	Process5	.761	.760	.098	.372

Table 6. Variables Entered/Removed

Variables Entered/Removed

Step	Entered	Lambda Statistic	df1	df2	df3	Exact F Statistic	df1	df2	Sig.
1	PROCESS3	.432	1	1	72.000	94.547	1	72.000	.000
2	PROCESS4	.373	2	1	72.000	59.713	2	71.000	.000

At each step, the variable that minimizes the overall Lambda is entered.
a Maximum number of steps is 10.
b Minimum partial F to enter is 3.84.
c Maximum partial F to remove is 2.71.
d F level, tolerance, or VIN insufficient for further computation.

Table 7. Variables in the Analysis

Step		Tolerance	F to Remove	Lambda
1	PROCESS3	1.000	94.547	
2	PROCESS3	.999	63.366	.706
	PROCESS4	.999	11.323	.432

Table 8. Classification Results

		Success	Predicted Group Membership pass	fail	Total
Original	Count	pass	46	0	46
		fail	6	22	28
	%	pass	100.0	.0	100.0
		fail	21.4	78.6	100.0
Cross-validated	Count	pass	46	0	46
		fail	6	22	28
	%	pass	100.0	.0	100.0
		fail	21.4	78.6	100.0

a Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.
b 91.9% of original grouped cases correctly classified.
c 91.9% of cross-validated grouped cases correctly classified.

Table 9. All-possible subsets listing

Subset size	Predictors
1	X1
	X2
	X3
	X4
2	X1, X2
	X1, X3
	X1, X4
	X2, X3
	X2, X4
	X3, X4
3	X1, X2, X3
	X1, X2, X4
	X1, X3, X4
	X2, X3, X4
4	X1, X2, X3, X4

Table 10. Mean Values of Standardized Canonical Discriminant Function Coefficients

	Function
	1
Process1	0.47531
Process2	0.27944
Process3	0.91456
Process4	0.70338
Process5	0.15231

Table 11. Mean Values of Structure Coefficients

	Function
	1
Process1	0.48163
Process2	0.27931
Process3	0.90488
Process4	0.71544
Process5	0.40288

Table 12. Summary of Classification Results

Predictors	% of original grouped cases correctly classified	% of cross-validated grouped cases correctly classified
Process1	67.6	67.6
Process2	58.1	58.1
Process3	86.5	86.5
Process4	81.1	81.1
Process5	54.1	54.1
Process1, Process2	63.5	62.2
Process1, Process3	89.2	89.2
Process1, Process4	83.8	83.8
Process1, Process5	67.6	67.6
Process2, Process3	90.5	90.5
Process2, Process4	81.1	81.1
Process2, Process5	68.9	54.1
Process3, Process4	91.9	91.9
Process3, Process5	89.2	89.2
Process4, Process5	81.1	81.1
Process1, Process2, Process3	89.2	89.2
Process1, Process2, Process4	81.1	77.0
Process1, Process2, Process5	67.6	64.9

Process1, Process3, Process4	91.9	91.9
Process1, Process3, Process5	89.2	89.2
Process1, Process4, Process5	79.7	79.7
Process2, Process3, Process4	91.9	91.9
Process2, Process3, Process5	89.2	89.2
Process2, Process4, Process5	81.1	81.1
Process3. Process4, Process5	91.9	90.5
Process1, Process2, Process3, Process4	91.9	91.9
Process1, Process2, Process3 Process5	90.5	89.2
Process1, Process3, Process4, Process5	90.5	90.5
Process1, Process2, Process4, Process5	79.7	75.7
Process2, Process3, Process4, Process5	91.9	90.5
Process1, Process2, Process3, Process4, Process5	91.9	90.5

U.S. Department of Education
Office of Educational Research and Improvement (OERI)

National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM031211

Reproduction Release

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Variable Selection Strategies in Discriminant Analysis	
Author(s): Jesus Tanguma	
Corporate Source: Texas A&M University	Publication Date:

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 in the indicated space following.

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 t Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)	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)	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED SAMPLE TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
Level 1	**Level 2A**	**Level 2B**
☒	☐	☐
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 o

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.

Signature: [illegible]	Printed Name/Position/Title: Jesus Tagume Grad Student	
Organization/Address: Texas A&M University Educational Psychology Dept College Station, TX 77843	Telephone: 409-862-8089	Fax:
	E-mail Address: jot1609@lbss.tamu.edu	Date: 1/6/00

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: