Journal of Modern Applied Statistical Methods

Volume 9 | Issue 1 Article 6

5-1-2010

The Effectiveness of Stepwise Discriminant Analysis as a Post Hoc Procedure to a Significant MANOVA

Erik L. Heiny
Utah Valley University, erik.heiny@uvu.edu

Daniel J. Mundform
University of Northern Colorado, daniel.mundfrom@unco.edu

Recommended Citation
Heiny, Erik L. and Mundform, Daniel J. (2010) "The Effectiveness of Stepwise Discriminant Analysis as a Post Hoc Procedure to a Significant MANOVA," *Journal of Modern Applied Statistical Methods*: Vol. 9 : Iss. 1 , Article 6.
DOI: 10.22237/jmasm/1272686700

Journal of Modern Applied Statistical Methods
May 2010, Vol. 9, No. 1, 43-51

1538 – 9472/10/$95.00

The Effectiveness of Stepwise Discriminant Analysis as a Post Hoc Procedure to a Significant MANOVA

Erik L. Heiny
Utah Valley University

Daniel J. Mundfrom
University of Northern Colorado

The effectiveness of SWDA as a post hoc procedure in a two-way MANOVA was examined using various numbers of dependent variables, sample sizes, effect sizes, correlation structures, and significance levels. The procedure did not work well in general except with small numbers of variables, larger samples and low correlations between variables.

Key words: Stepwise discriminant analysis, MANOVA, post hoc procedures.

Introduction

One common type of research question in multivariate analysis involves searching for differences between multiple groups on several different response variables. Considering response variables as a vector of dependent variables, a one-way MANOVA can be used to test the hypothesis that the mean vectors are the same across groups. However, if a significant MANOVA has been found, how does the researcher determine which of the response variables contribute to group differences?

Currently, most researchers use either multiple univariate F-tests, which are simply inappropriate, or descriptive discriminant analysis (DDA), which has been shown to lack power through simulation studies. Hawkins (1976) proposed the use of a stepwise MANOVA procedure, similar to stepwise regression, for selecting the best subset of variables to use in the MANOVA analysis. Hawkins further advocated for a Bonferroni adjustment to the α-level used at each step in the stepwise selection to control the overall Type I error rate, neither of these suggestions, however, seem to be in much use today. Another approach used by some researchers is stepwise discriminant analysis (SWDA). Criticisms of stepwise methods in general have been well-documented in the literature, most notably by Thompson (1995), which would appear to also apply to Hawkins' stepwise MANOVA procedure.

Essentially the criticisms center on stepwise methods being biased towards finding significance. Although this is a legitimate concern, it should be less prevalent in the context of this study; in this study real group differences exist on the dependent variables, therefore SWDA is not just fishing for differences that do not exist. Considering that some researchers are currently using SWDA in this context and that univariate F-tests and DDA are poor alternatives, empirical evidence is needed regarding the viability of SWDA as a post hoc procedure to a significant MANOVA.

The purpose of this research is to investigate the effectiveness of SWDA in distinguishing between significant and non-significant dependent variables when the MANOVA null hypothesis has been rejected. Specifically, it examines what the percentage of MANOVA dependent variables with means that differ between groups that are correctly identified as significantly different in a two-group SWDA (i.e., the power), and the percentage of MANOVA dependent variables with means that are the same in both groups that are incorrectly identified as significantly

Erik L. Heiny is an Assistant Professor in the Department of Mathematics. Email: erik.heiny@uvu.edu. Daniel J. Mundfrom is a Professor of Applied Statistics and Research Methods. Email: daniel.mundfrom@unco.edu.

different in a two-group SWDA (i.e., the Type I error). The effect of sample size, n, the number of dependent variables in the MANOVA, p, the correlation structure among the dependent variables, **ρ**, the effect size, d, and the significance level used in the stepwise selection, α, were also investigated.

Rencher and Larson (1980) performed a Monte Carlo simulation to examine the bias in Wilk's lambda in SWDA. In SWDA, an F-statistic can be used to test the significance of the reduction in Wilk's lambda when an additional variable is added to the model. The larger the reduction in Wilk's lambda due to the additional variable, the larger the F-statistic will become. Rencher and Larson note that if an arbitrary variable is considered for entry, the F-statistic follows a true F-distribution.

However, in SWDA several variables are considered for entry at each step and the maximum F-statistic from these variables is compared to the F-critical value. Because the F-statistic is maximized at each step, it does not follow an F-distribution and the procedure becomes biased towards selecting variables that do not contain discriminatory information. Rencher and Larson conclude that the bias becomes most pronounced when there are a large number of variables under consideration and a relatively small sample size. They write, "In the author's experience, such cases are fairly common. Habbema and Hermans (1977, p. 492) note that 'sample sizes of say 10-40 are not unusual, with a number of variables ranging from 10-200.'" (p. 350). The most drastic case in this study will be sample sizes of 50 with the number of variables equal to 8.

In addition, Rencher and Larson (1980) write, "we have restricted out attention to the null case of no difference between groups so as to provide some indication of the levels Wilks' lambda may reach when there is no real separation from group to group" (p. 351). In this study, SWDA was used when the null hypothesis is false, that is, real separation exists from group to group. Therefore, the bias in Wilk's lambda was not expected to be as severe in this study, but Type I errors in excess of alpha were likely and were watched closely.

Methodology

A Monte Carlo simulation was run using SAS PROC Interactive Matrix Language (IML). Two p-dimensional multivariate normal populations were created with characteristics that varied according to pre-set levels of the number of MANOVA dependent variables, p, which varied across the values, 2, 3, 4, 5, 6, 7, and 8, and a correlation structure among the p variables. In one population, the mean vector contained all zeros, whereas in the other population mean vector had half of the values set at 0 while the other half differed from 0 by an effect size, d, that varied across 0.2 (small), 0.5 (medium), and 0.8 (large). When the value of p was odd, the mean of the extra variable was set at 0; for example, with p = 5 and a small effect size, the two mean vectors were:

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \text{ and } \boldsymbol{\mu}_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ .2 \\ .2 \end{bmatrix}.$$

Both populations were generated with the same correlation matrix, **ρ**.

Six different correlation structures were examined. In each structure, variables were divided into set A, those that had the same mean in both groups, and set B, those that had means that differed between the groups. The within-set correlations, those between pairs of variables in set A (and between pairs of variables in set B), were varied across the values 0.20, 0.40 and 0.60. Initially, the across-set correlations, those between pairs of variables in which one variable came from set A and the other came from set B, was set at 0.20. For example, with p = 5, the three correlation matrices used were:

$$\boldsymbol{\rho}_1 = \begin{bmatrix} 1 & .2 & .2 & .2 & .2 \\ .2 & 1 & .2 & .2 & .2 \\ .2 & .2 & 1 & .2 & .2 \\ .2 & .2 & .2 & 1 & .2 \\ .2 & .2 & .2 & .2 & 1 \end{bmatrix},$$

$$\boldsymbol{\rho}_2 = \begin{bmatrix} 1 & .4 & .4 & .2 & .2 \\ .4 & 1 & .4 & .2 & .2 \\ .4 & .4 & 1 & .2 & .2 \\ .2 & .2 & .2 & 1 & .4 \\ .2 & .2 & .2 & .4 & 1 \end{bmatrix},$$

and

$$\boldsymbol{\rho}_3 = \begin{bmatrix} 1 & .6 & .6 & .2 & .2 \\ .6 & 1 & .6 & .2 & .2 \\ .6 & .6 & 1 & .2 & .2 \\ .2 & .2 & .2 & 1 & .6 \\ .2 & .2 & .2 & .6 & 1 \end{bmatrix}.$$

Because many of the scenarios examined with these correlation structures had large Type I error rates, the across-set correlations were reduced to 0.10 in order to see how this change would affect the results. Again for the p = 5 case, the three additional correlation matrices were:

$$\boldsymbol{\rho}_4 = \begin{bmatrix} 1 & .2 & .2 & .1 & .1 \\ .2 & 1 & .2 & .1 & .1 \\ .2 & .2 & 1 & .1 & .1 \\ .1 & .1 & .1 & 1 & .2 \\ .1 & .1 & .1 & .2 & 1 \end{bmatrix},$$

$$\boldsymbol{\rho}_5 = \begin{bmatrix} 1 & .4 & .4 & .1 & .1 \\ .4 & 1 & .4 & .1 & .1 \\ .4 & .4 & 1 & .1 & .1 \\ .1 & .1 & .1 & 1 & .4 \\ .1 & .1 & .1 & .4 & 1 \end{bmatrix},$$

and

$$\boldsymbol{\rho}_6 = \begin{bmatrix} 1 & .6 & .6 & .1 & .1 \\ .6 & 1 & .6 & .1 & .1 \\ .6 & .6 & 1 & .1 & .1 \\ .1 & .1 & .1 & 1 & .6 \\ .1 & .1 & .1 & .6 & 1 \end{bmatrix}.$$

Additionally, sample sizes were varied across 50, 100, 250, and 500 (although n = 500 was the only sample size used for the last three correlation structures), and the significance level used for variable selection in the SWDA was varied across 0.01, 0.05 and 0.10.

For each of 945 scenarios determined by the values of p, d, n, **ρ** and α, 5,000 replications were performed. Each replication consisted of selecting a random sample of size n from each population described above, which led to two sample mean vectors. A SWDA was performed on each sample using SAS PROC STEPDISC with the stepwise selection method and the F-test criterion for a chosen level of α. The percentage of correctly identified significant variables (power) and the percentage of non-significant variables incorrectly identified as significant (Type I error) were computed for each sample. Averaging these values across the 5,000 replications produced power and Type I error estimates for each scenario. Successful results were defined to be those situations for which power was maintained at 0.80 or higher and the Type I error rate did not exceed 0.10.

Results

Scenarios with Correlation Structure One, Two or Three

For correlation structures one, two and three, SWDA was only successful for certain situations when p was small, 2 or 3. As long as p was not larger than 3, varying the correlation structure between levels one, two and three had almost no effect on the results. The larger p became, however, the more the results changed for different correlation structures (see Tables 1 and 2). For p = 2 or 3 and a small sample size, n = 50, SWDA worked well for large effect sizes, d = 0.8, and α = 0.01 (Table 2). Type I errors were inflated above α but only to 0.03, and power was above 0.90.

As n increased to 100, and p was set equal to 2 or 3, SWDA was still successful for large effect sizes, but only when α was set to 0.01 (see Table 2). Power was over 0.99 and Type I error was 0.06. Additionally, for the same levels of n and p, SWDA worked well for medium effect sizes, d = 0.5, as long as α was set to 0.05 or 0.01 (see Table 1). For α = 0.01, power was around 0.82 and Type I error was near 0.025. For α = 0.05, power was 0.94 and Type I error was around 0.09.

Table 1: Power and Type I Error for α = 0.01, d = 0.5, Across-Set ρ = 0.2

	Within-Set ρ = 0.2				Within-Set ρ = 0.4				Within-Set ρ = 0.6			
	n				n				n			
p	50	100	250	500	50	100	250	500	50	100	250	500
2	0.4554	0.8214	0.9988	1.0000	0.4666	0.8268	0.9980	1.0000	0.4520	0.8190	0.9990	1.0000
	0.0126	0.0266	0.0662	0.1616	0.0132	0.0282	0.0710	0.1560	0.0156	0.0246	0.0588	0.1218
3	0.4510	0.8326	0.9978	1.0000	0.4550	0.8226	0.9980	1.0000	0.4466	0.8280	0.9986	1.0000
	0.0153	0.0256	0.0665	0.1447	0.0140	0.0245	0.0653	0.1315	0.0141	0.0242	0.0588	0.1218
4	0.3615	0.6393	0.9715	0.9999	0.3236	0.5100	0.8472	0.9947	0.3012	0.4613	0.6063	0.9083
	0.0143	0.0457	0.1990	0.4206	0.0160	0.0312	0.1094	0.2720	0.0148	0.0275	0.0754	0.1804
5	0.3653	0.6388	0.9705	0.9999	0.3196	0.5088	0.8522	0.9953	0.2953	0.4601	0.6106	0.9077
	0.0203	0.0462	0.1805	0.3631	0.0157	0.0327	0.1039	0.2196	0.0124	0.0275	0.0656	0.1430
6	0.3039	0.5242	0.9134	0.9990	0.2501	0.3790	0.6610	0.9062	0.2292	0.3215	0.4527	0.6637
	0.0196	0.0664	0.2840	0.5677	0.0179	0.0373	0.1203	0.2646	0.0176	0.0282	0.0696	0.1586
7	0.3059	0.5315	0.9194	0.9996	0.2521	0.3763	0.6618	0.9135	0.2302	0.3193	0.4558	0.6677
	0.0239	0.0618	0.2589	0.5055	0.0177	0.0330	0.1070	0.2200	0.0145	0.0248	0.0615	0.1361
8	0.2622	0.4622	0.8529	0.9970	0.2055	0.3072	0.5421	0.7793	0.1839	0.2458	0.3654	0.5203
	0.0260	0.0844	0.3364	0.6519	0.0190	0.0374	0.1184	0.2383	0.0155	0.0275	0.0641	0.1388

Table 2: Power and Type I Error for α = 0.01, d = 0.8, Across-Set ρ = 0.2

	Within-Set ρ = 0.2				Within-Set ρ = 0.4				Within-Set ρ = 0.6			
	n				n				n			
p	50	100	250	500	50	100	250	500	50	100	250	500
2	0.9146	0.9994	1.0000	1.0000	0.9158	0.9990	1.0000	1.0000	0.9074	0.9992	1.0000	1.0000
	0.0328	0.0662	0.1836	0.4286	0.0294	0.0602	0.1906	0.4310	0.0312	0.0600	0.1900	0.4278
3	0.9180	0.9984	1.0000	1.0000	0.9160	0.9992	1.0000	1.0000	0.9162	0.9986	1.0000	1.0000
	0.0321	0.0622	0.1656	0.3547	0.0304	0.0587	0.1555	0.3119	0.0262	0.0542	0.1455	0.2880
4	0.6908	0.9575	1.0000	1.0000	0.5509	0.8083	0.9985	1.0000	0.4894	0.5855	0.9495	1.0000
	0.0529	0.1698	0.4612	0.7833	0.0361	0.0979	0.3072	0.5063	0.0312	0.0645	0.2189	0.3919
5	0.6909	0.9612	1.0000	1.0000	0.5444	0.8162	0.9984	1.0000	0.4872	0.5874	0.9511	1.0000
	0.0565	0.1530	0.4060	0.6827	0.0371	0.0867	0.2470	0.3790	0.0281	0.0563	0.1693	0.2840
6	0.5708	0.8580	0.9997	1.0000	0.4020	0.6227	0.9427	0.9995	0.3351	0.4275	0.6946	0.9328
	0.0753	0.2309	0.5992	0.8669	0.0397	0.1097	0.2889	0.4639	0.0296	0.0602	0.1801	0.3013
7	0.5711	0.8704	0.9998	1.0000	0.4089	0.6258	0.9447	0.9996	0.3347	0.4313	0.6957	0.9379
	0.0713	0.2092	0.5294	0.8123	0.0410	0.0972	0.2391	0.3765	0.0267	0.0562	0.1533	0.2334
8	0.4906	0.7726	0.9978	1.0000	0.3318	0.5055	0.8111	0.9866	0.2545	0.3454	0.5533	0.7617
	0.0850	0.2610	0.6603	0.9231	0.0430	0.1033	0.2565	0.4191	0.0299	0.0595	0.1593	0.2369

As n increased to 250 while p remained equal to 2 or 3, the procedure was successful for medium effect sizes and $\alpha = 0.01$ (see Table 1). Power was over 0.99 and Type I error was less than 0.07. The procedure became too aggressive for large effect sizes, with observed Type I error going as high as 0.50 in some situations.

When n increased to 500 while p was still limited to 2 or 3, SWDA was only successful for small effect sizes, $d = 0.20$, and $\alpha = 0.05$. Power was approximately 0.89 and Type I error was near 0.09. When α was lowered to 0.01, Type I error dropped to 0.02 but power went down to 0.72. When α was increased to 0.10, power increased to 0.94 but Type I error was high, 0.15. Due to the aggressive nature of SWDA, the procedure did not work well for medium or large effect sizes when $n = 500$. The power was very high, but Type I error increased well above 0.10.

Scenarios with Correlation Structure Four, Five or Six

As noted, correlation structures four, five and six were simulated with $n = 500$ to investigate the increase in Type I error which accompanied any increase in effect size. Recall that in correlation structures one, two and three, the across-set correlations were kept constant at 0.20. In correlation structures four, five and six, these correlations were reduced to 0.10. In comparison, SWDA was much more successful under correlation structures four, five and six. The procedure worked well for many scenarios when p was equal to 2 or 3, and it also worked well under certain conditions for p as high as 7 (see Table 3). When p was 2 or 3, alternating between correlation structures four, five and six produced almost identical results (see Table 3).

When p was equal to 2 or 3, SWDA worked well for small effect sizes, $d = 0.20$ and $\alpha = .05$. Power was equal to 0.89 and Type I error was 0.06. For $\alpha = 0.01$ power decreased to 0.71, and for $\alpha = 0.10$ Type I error increased to 0.11. For medium and large effect sizes, $d = 0.50$ and $d = 0.80$ respectively, SWDA worked well if $\alpha = 0.01$. Power was equal to 1.00 in both cases, and Type I error was 0.04 and 0.08 respectively (see Table 3).

Table 3: Power and Type I Error for $\alpha = 0.01$, $n = 500$, Across-Set $\rho = 0.1$

	Within-Set ρ = 0.2			Within-Set ρ = 0.4			Within-Set ρ = 0.6		
	d			d			d		
p	0.2	0.5	0.8	0.2	0.5	0.8	0.2	0.5	0.8
2	0.7158	1.0000	1.0000	0.7234	1.0000	1.0000	0.7198	1.0000	1.0000
	0.0116	0.0354	0.0804	0.0122	0.0354	0.0824	0.0122	0.0392	0.0786
3	0.7128	1.0000	1.0000	0.7292	1.0000	1.0000	0.7222	1.0000	1.0000
	0.0143	0.0353	0.0737	0.0123	0.0331	0.0702	0.0110	0.0308	0.0624
4	0.5581	0.9999	1.0000	0.4551	0.9921	1.0000	0.4197	0.8966	0.9995
	0.0159	0.0857	0.2103	0.0126	0.0662	0.1443	0.0108	0.0444	0.0972
5	0.5621	0.9999	1.0000	0.4563	0.9916	1.0000	0.4519	0.8933	0.9995
	0.0153	0.0796	0.1884	0.0138	0.0578	0.1194	0.0124	0.0380	0.0840
6	0.4601	0.9963	1.0000	0.3390	0.8811	0.9991	0.3002	0.6499	0.9062
	0.0179	0.1391	0.2933	0.0135	0.0689	0.1622	0.0138	0.0402	0.0986
7	0.4641	0.9961	1.0000	0.3424	0.8791	0.9987	0.2991	0.6527	0.9120
	0.0188	0.1229	0.2564	0.0134	0.0654	0.1413	0.0118	0.0365	0.0813
8	0.4025	0.9746	0.9999	0.2757	0.7325	0.9664	0.2305	0.5070	0.7395
	0.0222	0.1646	0.3248	0.0146	0.0683	0.1553	0.0128	0.0377	0.0869

For values of p greater than 3, alternating between correlation structures four, five and six begins to make a difference. For correlation structure four, the within-set correlations were set equal to 0.20. For correlation structures five and six, these correlations were increased to 0.40 and 0.60 respectively. SWDA worked well for p = 4 or 5 when d = 0.50 and α = 0.01 (see Table 3). Power and Type I error values were very similar for both p = 4 or 5, but were different for different correlation structures. Under correlation structures four, five and six, power was equal to 0.9999, 0.9920, and .8950 respectively, and Type I error was equal to 0.08, 0.06, and 0.04 respectively.

For values of p greater than 5, SWDA was effective in a couple of scenarios: for p = 6 or 7, the procedure worked well for medium and large effect sizes when α = .01. Power was around 0.88 and 0.90 respectively, and Type I error was around 0.07 and 0.09 respectively. Lowering the across-set correlations from 0.20 to 0.10 appeared to improve the effectiveness of SWDA, specifically with respect to Type I error. However, even with the across-set correlations reduced, SWDA still appeared to enjoy limited success when values of p increased above 3.

Effect of Independent Variables on Power and Type I Error

p – The Number of MANOVA Dependent Variables

SWDA appeared to become less effective as the number of MANOVA dependent variables increased. Generally, as p increased, the power decreased and Type I error increased. Power and Type I error tended to be very similar when results are grouped by p = 2 or 3, then by p = 4 or 5, by p = 6 or 7 and finally by p = 8. It should be noted that for each of these groupings, the number of variables with means that differed between the two groups is the same. Satisfactory results were usually obtained for only p = 2 or 3, this may be largely due to having only one variable whose mean is different between the two groups. Satisfactory results might still be obtained for values of p greater than 3, as long as only one of the variables has a mean that differs between the groups.

When the sample size was large, especially if the within-set correlation was low, SWDA became too aggressive resulting in Type I errors that were too high. This problem was exacerbated as p increased. In some cases, Type I error increased from 0.30 to 0.80 as p increased from 2 to 8. These results support the claim by Thompson (1995) that stepwise methods tend to increase the likelihood of Type I errors, especially for larger values of p. Thompson suggests that because several variables are considered for entry at each step, more degrees of freedom should be charged to the numerator from the denominator of the F-statistic. This technique will produce a smaller value for the F-statistic, making Type I errors less likely. However, Thompson mentions as a caveat that this outcome is less likely to be an issue when the number of dependent variables is small.

Less favorable results regarding power were also observed when p increased. The F-statistic used in SWDA is described by Klecka (1980) as “the F-to-enter is a partial multivariate F-statistic which tests the additional discrimination introduced by the variable being considered after taking into account the discrimination achieved by the other variables already entered (Dixon, 1973, p. 241)” (p. 57). For certain variables, when only the additional contribution to discrimination is considered, problems can arise if these variables share information with other variables that are already in the model. “…two or more of the variables may share the same discriminating information even though individually they are good discriminators. When some of these are employed in the analysis, the remainder are redundant” (Klecka, p. 52).

For this study, if multiple variables differ between the two groups the power can be reduced if SWDA considers one or more of these variables as redundant, thus, when p increased, the number of variables that differed between the two groups also increased. With respect to power in SWDA, it could be that increasing p by itself does not reduce power, but increasing the number of variables whose means differ between the two groups does reduce power because SWDA may consider some of these to be redundant. This effect can be observed in Tables 1, 2 and 3.

n – Sample Size

Results based on sample size were as expected: as n increased, both power and Type I error increased as well. Unfortunately, SWDA appears to be too aggressive when the sample size gets large. Under correlation structures one, two or three, when n was 250 or 500, Type I error was too high except under certain conditions. High enough power was not an issue when n got large, but in order to keep Type I error below 0.10 the effect size needed to be small (d = 0.20) and α = 0.01. When the across-set correlation was reduced from 0.20 to 0.10, the Type I error rate was controlled much better (see Table 3). For correlation structures four, five and six, there were situations for medium and large effect sizes, as well as small effect sizes, where the Type I error stayed below 0.10. Lower levels of across-set correlation enables SWDA to perform more efficiently but caution should be used by the researcher when using SWDA with large sample sizes; at the very least, small levels of α should be used in this situation.

d – Effect Size

As expected, when effect size increased, power increased. This pattern was observed regardless of sample size, but was more apparent with smaller values of n. When the sample size became large the power of SWDA was high even for small effect sizes. Discrepancy in power for different effect sizes can be observed in Table 3.

Surprisingly, Type I error increased as well as power when effect size increased. It was believed that with higher effect sizes it would be easier for SWDA to distinguish between variables with means that differed between the groups and variables with means that were the same in both groups. However, this outcome was not the case and the pattern became even more apparent as n and p increased. This pattern is shown when comparing Tables 1 and 2. In some cases, for large n, large p and large d, Type I errors in excess of 0.90 were observed - SWDA becomes more aggressive as effect size increases.

The only connection between variables whose means are different in the two groups and variables whose means are the same in the two groups is the across-set correlation. Increasing the effect size did nothing to change the across-set correlation, but when the effect size became larger, a variable with the same mean in both groups was now correlated with a variable whose mean had an even larger difference between the two groups. This relationship appeared to increase the likelihood of the variable with the same mean in both groups, being incorrectly identified by SWDA.

To examine this relationship further, additional simulations were run at n = 500 and with the across-set correlation reduced to 0.10. Results for these scenarios (see Table 3) show that the same pattern was still observed. As effect size increased, the likelihood of Type I error increased as well. However, the Type I error rate was reduced significantly under correlation structures four, five and six. With the across-set correlation reduced from 0.20 to 0.10, a variable with the same mean in both groups now had a smaller correlation with a variable whose mean differed between the two groups. When the effect size was increased, therefore, the variable with the same mean in both groups was less likely to be incorrectly identified by SWDA.

It is difficult to explain why this happens in SWDA, but it appears that the across-set correlation is the key. Apparently, when a variable with the same mean in both groups is correlated to a degree with a variable with a high level of discriminatory power, SWDA has a tendency to select both variables. There appears to be a guilty-by-association factor present. The likelihood of incorrectly selecting the variable with the same mean in both groups increases as the correlation between the two variables increases.

ρ – Correlation

Within-set correlations varied among levels 0.20 (correlation structures one and four), 0.40 (correlation structures two and five) and 0.60 (correlation structures three and six). With all other independent variables held constant, as the within-set correlations increased, power and Type I error both decreased. This result indicates that SWDA becomes more conservative as correlations among MANOVA dependent variables increases; this pattern became more apparent as p increased. When one variable with means that differed between the groups had been correctly selected by SWDA, the likelihood of selecting another variable with means that differed between the groups went down as the correlation between these two variables increased. The higher the correlation between these two variables, the less unique discriminatory information was offered by the second variable.

The same pattern was observed among variables with the same mean in both groups. Once one of these variables had been incorrectly selected by SWDA, the likelihood of incorrectly selecting a second variable went down as the correlation between the two variables increased. Again with higher correlation between these two variables, any imagined discriminatory information detected by SWDA, appeared to be redundant for the second variable.

Across-set correlations varied among levels 0.20 (correlation structures one, two and three) and 0.10 (correlation structures four, five and six). As across-set correlations increased, the likelihood of Type I error also increased. For a variable with the same mean in both groups, any correlation it shared with a variable with means that differed between the groups, made it more likely to be incorrectly selected by SWDA (this outcome is the same guilty-by-association factor previously mentioned).

A final observation was made on the effect of correlations among MANOVA dependent variables on SWDA due to a programming error early in the simulation process. The error in the simulations produced correlation matrices that were identity matrices so that all MANOVA dependent variables were statistically independent. The results for power and Type I error were very good using SWDA in this context. It should be noted that complete statistical independence between all dependent variables is not a realistic correlation structure, but it gives a little more insight into the effectiveness of SWDA as a post hoc procedure to MANOVA. For situations in which there is little correlation among the MANOVA dependent variables, SWDA may be an effective post hoc procedure to a significant MANOVA. The sample correlation matrix can help researchers estimate the level of correlations among the dependent variables.

α – Level of Significance

As expected, when α increased, power and Type I error increased as well. For small n and small d, observed values of Type I errors were very close to the set level α. This relationship was consistent regardless of p or the level of correlation among the MANOVA dependent variables. However, as n and/or d increased, the observed value of Type I error tended to increase to well above the set level of α. In some extreme cases the observed Type I error exceeded 0.90 and Type I error values in the 0.40 to 0.50 range were commonplace for large values of n or d.

Inflated Type I error levels were expected in this study but the actual inflation in the Type I error rates were much larger than expected. Rencher and Larson (1980) observed that the F-statistic used in SWDA is biased towards including variables that should not be selected. However, Rencher and Larson only considered the case where the MANOVA null hypothesis was true. In this study, the MANOVA null hypothesis was false, therefore it was expected that Type I errors would not be drastically inflated since SWDA wasn't fishing for significant results. Inflated Type I errors were observed, however, suggesting that researchers using SWDA should set α to lower than desired values of Type I error, especially for larger sample sizes (n = 250 or 500).

Conclusion

Although SWDA appears to be a very powerful procedure, it seems to be too aggressive in general. The biggest issue in this study was inflated Type I error; researchers who are using SWDA need to be aware of this problem. However, researchers may be able to use the procedure quite successfully under certain conditions. First, researchers should keep the number of dependent variables small, probably no more than three or four according to this study. Secondly, SWDA will be most successful when the correlations among the dependent variables are small. This condition is very important and researchers should check the sample correlation matrix before using SWDA. Finally, researchers may be able to interpret the order in which variables are selected, albeit with some caution. Although there is no empirical evidence offered in this study, it was observed that when SWDA became too aggressive and selected too many variables, the variables with means that differed between the groups were generally selected first. If researchers are aware of this pattern, they can compare sample mean vectors on the variables that were selected later by SWDA, and make some tentative conclusions on the discriminatory power of these variables.

Inflation of Type I error was a serious issue in this study when sample size increased. Because the order in which the variables were selected was generally correct, future researchers should look for ways to make SWDA stop in time, especially for larger sample sizes. One possible solution would be to use the squared partial correlation criterion, rather than the F-test criterion used in this study. The squared partial correlation criterion and the F-test criterion select variables in the same order, but the F-test criterion tends to select more variables as the sample size increases (SAS Institute Inc., 2004). Future researchers can also conduct simulations using Thompson's (1995) adjustment for degrees of freedom to determine how well this method controls Type I error.

Another possibility could be to make a Bonferroni-type adjustment to the α-level that is used to select the significant variables, similar to what Hawkins (1976) advocated with his stepwise MANOVA procedure. When SWDA is used in this context, it can be reasonably viewed as a multiple comparison-type procedure, similar to how the Scheffe' and Bonferroni procedures are used as a follow-up to a significant ANOVA. In that context, it is common practice to adjust the significance level for each of the multiple follow-up tests to control the family-wise error rate.

Because SWDA is also performing multiple tests on several variables at each step of the selection process, using some type of adjustment for each test at each step would seem like a reasonable step to take. This study did not address the utility of making a Bonferroni-type adjustment, so further research would be needed in order to determine the effectiveness of doing so, as well as how much of an adjustment to the α-level for each test would be needed to control the overall Type I error rate at the nominal level.

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* (*2nd Ed.*) Hillsdale, NJ: L. Erlbaum Associates.

Dixon, W. J. (1973). *BMD: Biomedical programs*. Berkeley, CA: University of California Press.

Habbema, J. D. F., & Hermans, J. (1977). Selection of variables in discriminant analysis by F-statistic and error rate. *Technometrics*, *19*, 487-493.

Hawkins, D. M. (1976). The subset problem in multivariate analysis of variance. *Journal of the Royal Statistical Society*, *B38*, 132-139.

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

Rencher, A. C., & Larson, S. F. (1980). Bias in Wilk's Lambda in stepwise discriminant analysis. *Technometrics*, 22(*3*), 349-356.

SAS Institute Inc. (2004). *SAS/STAT user's guide, version 9.1, SAS OnlineDoc® 9.1.3*. Cary, NC: SAS Institute Inc.

Thompson, B. (1995). *Inappropriate statistical practices in counseling research: Three pointers for readers of research literature*. (ERIC Document Reproduction Service No. ED391990.)