HBAT Multiple Discriminant Analysis (August 2019)

Jason Huggy, Student, Lewis University

I. INTRODUCTION

The purpose of this study was to utilize multiple discriminate analysis (MDA) to analyze the HBAT dataset found in *Multivariate Data Analysis* [1]. The data set includes 99 lines of unlabeled data, and 23 variables. MDA analyzes the relationship between multiple independent variables and a non-metric dependent variable. In MDA, the primary purpose is to build discriminating functions that classify each observation in the data set. In this study, SAS Enterprise Guide (EG) is used to estimate all discriminant functions and to analyze all variables.

II. METHODOLOGY

To conduct this study, SAS Enterprise Guide is used to produce all data tables. To begin, the HBAT data is imported into EG and verified to be free of missing values or errors. A quick examination of the data shows that the first five columns of data, as well as the last, are nonmetric features. The rest of the features are all metric. For this study only three independent variables are utilized: 11, 13, and 17. The dependent variable is a classification variable labeled as either 0 or 1.

A. Splitting the Data

In order to conduct this study, the HBAT data was split into a training set and a testing set. The training set contains 60 percent of the data, while the testing set contains the other 40 percent. This allows the initial model to be built with the training set, which will determine the discriminating functions. The testing set is then applied to validate the effectiveness of the model.

B. Estimating the Discriminate Functions

To build the MDA model in EG, the *Discriminant Analysis* function is used. Variables 11, 13, and 17 are selected as the independent variables, and the column representing the non-metric, categorical data is selected as the classification variable. EG does not contain a method to use MDA in a stepwise manner, so the options are left at the default setting. The training data, which contains 60 percent of the observations, is used to train the model and determine the discriminant functions. The testing set is selected as the data for the program to predict once the model is built.

III. RESULTS AND DISCUSSION

After running the MDA function in EG, the results are analyzed. All results can be seen in the appendix. The first thing to notice is that the training data contained data with mostly zeros as their classification. 50 observations were zeros while only 10 were ones. This is a problem because it doesn’t allow the model to get enough exposure to cases marked as ones. Hair et al. recommends at least 20 observations per category, but there are only ten for the ones [1]. Problems like this make it harder for the model to distinguish a difference between the two categories because there is mostly data to support the zeros but not the ones. The impact of this issue is seen throughout the rest of the study.

When it comes to the impact each independent variable has on the discriminant functions, variable 11 has the greatest contribution. At 7.9 for zeros and 8.3 for ones, it carries the largest load out of the independent variables. Therefore, variable 11 can do the most to tell us about what each observation should be classified as. Variable 13 wasn’t too far behind 11, and 17 had the least impact on the discriminant functions.

When discriminant analysis is run, we can see that the model labeled half of the observations as ones and the other half as zeros. This in an implication that has to do with the sample sizes and may also be a problem with EG. The data was essentially split in half with the cutting score, but the cutting score does not consider the number of observations classified as ones. Since the ones only consisted of 16 percent of the data, it should have held less of a weight when assigning values. Since there is more data to support the zeros, it would be safer to increase the area that labels points as zeros. This would allow for the model to correctly predict values with less error. Even before getting to the test set, this MDA process miscategorized 31 observations. That is one over half of the observations. When looking at the data for the points that were misclassified, one can see that the probabilities for ones and zeros are very close. Meaning the model did not allow for a great enough discriminant to tell observations apart. This is most likely due to the fact that there were too few ones in the training data. With a better proportion of ones, the model may have been able to better separate between the two classes.

Because of this error the prediction onto the testing set was inadequate. The model incorrectly classified all ones in the testing set. Therefore, the reliability of this model is low. This study would need to be revamped in a way to provide more data and better discriminatory power amongst the independent variables.

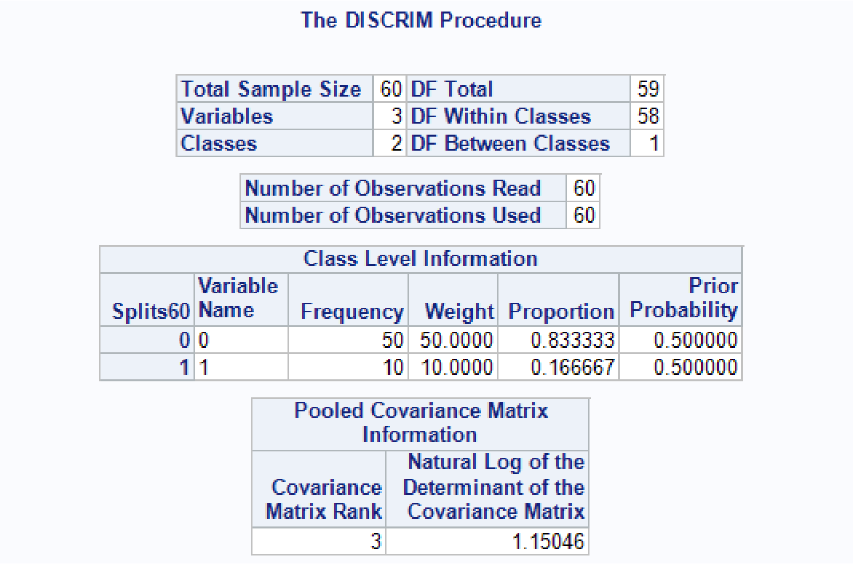
IV. CONCLUSION

Overall, the MDA process was not successful in this instance, but it did present some major talking points. In this study it was seen how important it is to have a large enough sample size. If there are not enough observations within a certain category, then it can cause major problems when it comes time for the model to discriminate between values. This model had too much of a crowded middle zone between categories, and it showed when it came down to validating the predictions. In the end, it is up to the researcher to try to do everything possible to create great discriminatory functions. The idea is to get as much separation as possible between each category; otherwise, it is probably better to use another form of multivariate analysis.

REFERENCES

[1] J. Hair, W. Black, B. Babin, R. Anderson, *Multivariate Data Analysis: Pearson New International Edition*. 2

Appendix



The DISCRIM Procedure

Total Sample Size	60	DF Total	59
Variables	3	DF Within Classes	58
Classes	2	DF Between Classes	1

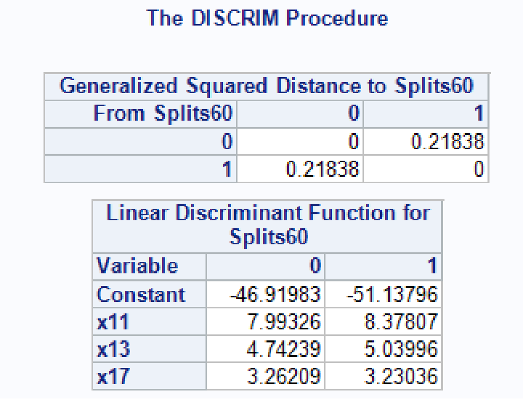
Number of Observations Read	60
Number of Observations Used	60

Class Level Information

Splits60	Variable Name	Frequency	Weight	Proportion	Prior Probability
0	0	50	50.0000	0.833333	0.500000
1	1	10	10.0000	0.166667	0.500000

Pooled Covariance Matrix Information

Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
3	1.15046



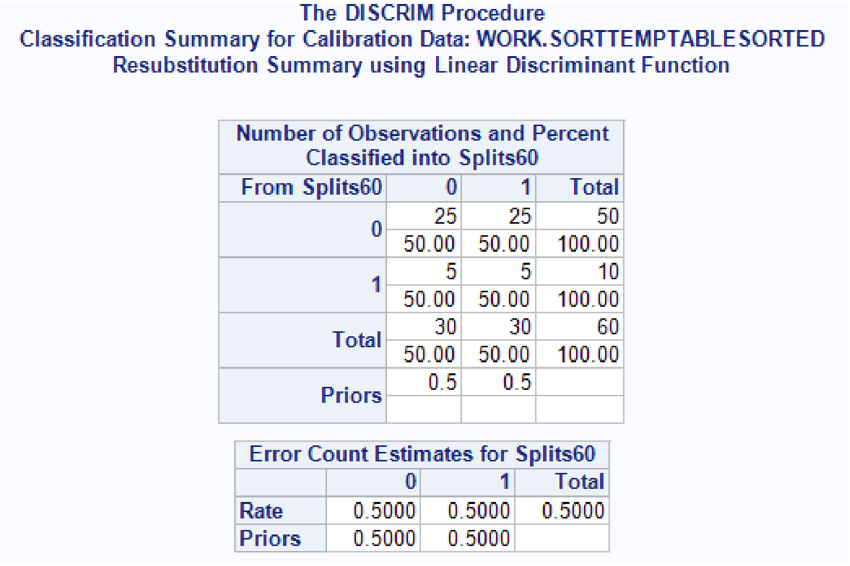
The DISCRIM Procedure

Generalized Squared Distance to Splits60

From Splits60	0	1
0	0	0.21838
1	0.21838	0

Linear Discriminant Function for Splits60

Variable	0	1
Constant	-46.91983	-51.13796
x11	7.99326	8.37807
x13	4.74239	5.03996
x17	3.26209	3.23036



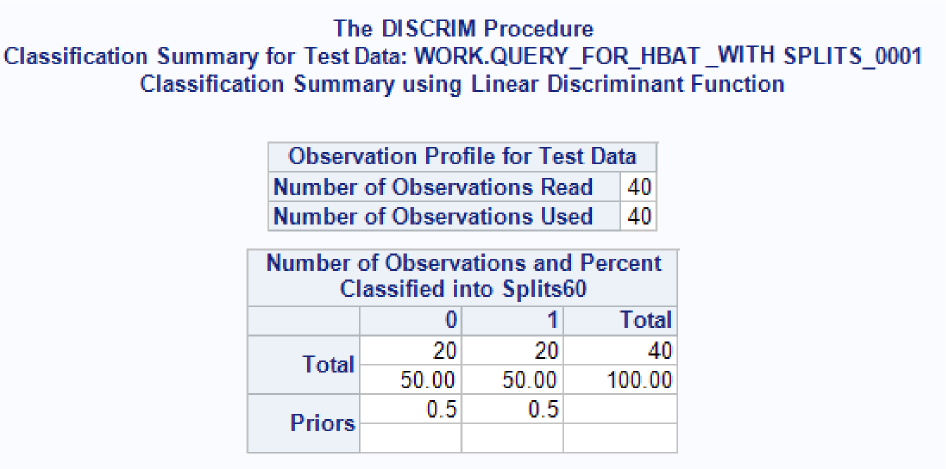
The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.SORTTEMPTABLESORTED
Resubstitution Summary using Linear Discriminant Function

Number of Observations and Percent Classified into Splits60

From Splits60	0	1	Total
0	25	25	50
	50.00	50.00	100.00
1	5	5	10
	50.00	50.00	100.00
Total	30	30	60
	50.00	50.00	100.00
Priors	0.5	0.5	

Error Count Estimates for Splits60

	0	1	Total
Rate	0.5000	0.5000	0.5000
Priors	0.5000	0.5000	



The DISCRIM Procedure
Classification Summary for Test Data: WORK.QUERY_FOR_HBAT _WITH SPLITS_0001
Classification Summary using Linear Discriminant Function

Observation Profile for Test Data	
Number of Observations Read	40
Number of Observations Used	40

Number of Observations and Percent Classified into Splits60			
	0	1	Total
Total	20	20	40
	50.00	50.00	100.00
Priors	0.5	0.5	