UNIVERSITY OF CAPE COAST

COMPARATIVE STUDY OF THE LOGISTIC REGRESSION ANALYSIS AND THE DISCRIMINANT ANALYSIS

ERIC ABAYIE PREMPEH

JULY 2009

UNIVERSITY OF CAPE COAST

COMPARATIVE STUDY OF THE LOGISTIC REGRESSION ANALYSIS AND THE DISCRIMINANT ANALYSIS

BY

ERIC ABAYIE PREMPEH

Thesis submitted to the Department of Mathematics and Statistics of the School of Physical Sciences, Faculty of Science, University of Cape Coast, in partial fulfilment of the requirements for award of Master of Philosophy Degree in Statistics

JULY 2009

DECLARATION

Candidate's Declaration

I hereby declare that this thesis is the result of my own original work and that no part of it has been presented for another degree in this university or elsewhere.

Candidates Signature:…… Date:……...

Name:……

Supervisors' Declaration

We hereby declare that the preparation and presentation of the thesis were supervised in accordance with the guidelines on supervision of thesis laid down by the University of Cape Coast.

Principal Supervisor's Signature:…….. Date:……

Name:…….

Co-Supervisor's Signature:……..... Date:……..

Name:…….

ABSTRACT

This thesis considers logistic regression and discriminant analysis. Because the ordinary least regression requires that a dependent variable in a study cannot be categorical, the logistic regression and the discriminant analysis techniques are two major techniques that are often used to handle categorical dependent variable problems. This study, therefore, seeks to getting answer to the question of whether using the logistic regression and the discriminant analysis techniques, on the same data set, would yield the same result or not.

Since the logistic regression and the discriminant analysis methods relates in many ways, extensive review of the theories behind them and comparison, in terms of similarities and differences, were necessary and have therefore been captured in the study. The graphical nature of the two techniques, interpretation of results, in using the two techniques, and significant tests of the various aspects of the two techniques are also not left out. At the end, empirical comparison of the binary logistic regression and the two-group discriminant analysis was made and the result for this comparison suggests that logistic regression gives a better result than two-group discriminant analysis when all requirements and assumptions of the two techniques are met.

Finally, summary of the findings of the research is also captured in the study and consideration also given to discussion and conclusion about the findings.

ACKNOWLEDGEMENTS

It is a pleasure to thank the many people who made this thesis possible. It is difficult to overstate my gratitude to my Supervisor, Professor Benony K. Gordor of Department of Mathematics and Statistics, U.C.C, who with his enthusiasm, inspiration, insightful comments and great efforts explained things clearly that helped to make this thesis possible.

My sincere thanks also goes to all the other lecturers of Department of Mathematics and Statistics, U.C.C, who taught me statistics especially to Dr. Mrs. Natalie Mensah (Head, Department of Mathematics and Statistic, U.C.C.) and Dr. Nathaniel Howard, for their support and encouragement.

. I would also like to express my gratitude to Mr. Frederick Sam of Department of Physics, U.C.C, and also to Mr. Adjei Adjepong of Department of History, U.C.C, for their support.

Again, my sincere thanks goes to my entire family, especially Mrs. Henrietta Turkson and Mrs. Winifred Bettschen and her husband, Mr. Hansjorg Bettschen, for supporting me financially throughout my education and also providing me sound advice.

I am also grateful to all the non-teaching staff members of the Department of Mathematics and Statistics, University of Cape Coast, for their cordial relationship with me during my studies in the university.

DEDICATION

This work is dedicated to my entire family.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

Background of the Study

There are a number of statistical techniques available for handling various problems. Many of these techniques come as models such as linear, exponential and quadratic models. Some of these models have become integral components concerned with describing the relationship between a response variable and one or more explanatory variables. If there is a reason to believe that a linear relationship exists between a variable of interest (response variable) and other variables (predicator variables) in a study, the ordinary linear model is one technique that is often used for predicting outcomes. This technique is mostly adopted due to its flexibility for analysing the relationship between multiple independent variables and a single dependant variable. Much of its flexible is due to the way in which all sorts of independent variables can be accommodated.

However, the linear model has some limitations and can therefore not be used in certain situations, even if there is a linear relationship between the response variable and the explanatory variables. (Hosmer & Leweshow, 2000).

Statement of the Problem

A limitation of the ordinary linear models is the requirement that the dependent variable cannot be categorical. In many studies, however, variables that are of interest are usually categorical. For example, if a study is to determine whether a patient will recover from an ailment or die; or whether a student will pass an examination or fail; and so on, are situations where the response variables (i.e. "recover or die" and "pass or fail") are all categorical. In such situations, it is inappropriate to employ the ordinary linear model as the technique for finding the relationship between the variables of interest and the risk factors involved in order to make predictions or to make classification of cases. (Statgun, n.d).

A range of techniques have been developed for providing answers to this question but the two commonly adopted techniques for answering this question is discriminant analysis and logistic regression. Even though, these techniques are used for different purposes but they can sometimes be used to achieve a common object. (Statgun, n.d).

Objectives of the Study

The main objectives of the study are to:

1. Compare logistic regression and discriminant analysis both theoretically and empirically.
2. Find out, if analysing the same data set with the two techniques would yield the same result.

Scope of the Study

Even though using discriminant analysis and logistic regression to analyse the same data set reveal the same pattern in many cases, their ways of arriving at results are different and they also require different assumptions. It was, therefore, important to have thorough review of the two techniques to know how results are arrived at.

The research, therefore, reviewed some assumptions and requirements of each method and also the types of variables for each technique. It also covered methods of variable selection into the models and touched on the forms of the two techniques. Discussion of the graphical nature of the two techniques, various significant tests of the importance of independent variables, the reliability of the models and results interpretation in the two techniques were considered. The study also involved empirical analysis using the two techniques and a comparison of the results was made to establish the discrepancies in using the two techniques, if any.

Significance of the Study

The study provides differences in efficiency of the logistic regression analysis and the discriminant analysis techniques. It also provides insight to researchers on reliability of the results of their study when one technique is used instead of the order, under specific conditions.

Outline of the Study

The study is organised into six chapters. Chapter one of the study talks about the introduction of the study. The introduction consists of the background of

the study, the limitations of the linear regression, objectives, scope and significance of the study.

The chapter two is about literature review of the logistic regression and the discriminant analysis techniques. In this chapter, the researcher tries to find some studies that have been carried out by other researchers which are in relation to the objectives of the study. This offers the researcher the opportunity to know the consistencies or disparities in using the logistic regression or the discriminant analysis techniques and also providing information on other fields of study, other than statistic, where logistic regression and discriminant analysis can be adopted in solving problems.

The chapter three of the work highly concentrates on the review of the logistic regression. This review talks about the family of the logistic regression, the forms of the model, the graphical nature of the logistic model, the various methods of variable selection into the model, the types of the technique and the various significance tests.

The chapter four also deals with the review of the discriminant analysis. This review covers areas such as the form of the discriminant model, the types of the discriminant analysis, the assumptions of the discriminant analysis and the significant tests of the various factors of the model.

The chapter five of the study focuses on the comparison of the logistic regression and the discriminant analysis. The comparison is in two parts. The first part is theoretical comparison in terms of similarities and differences based on the theoretical background of the two techniques. The second part is empirical comparison of the logistic regression and the discriminant analysis using the same data set.

The sixth chapter comprises the summary, discussion, results and conclusion, based on the findings of the study.

CHAPTER TWO

LITERATURE REVIEW

This chapter is about review of some works or studies which have been carried out by other researchers in relation to the objectives of this study. It also aims at identifying areas that the logistic regression analysis and the discriminant analysis techniques can be applied.

The relevance of logistic regression and discriminant analysis is not limited only to the field of statistics. Application of the logistic regression analysis and the discriminant analysis in the clinical medicine has been widespread. Betensky and Williams (2001), for example, in a study on the lymphocyte proliferative assay (LPA) of immune competence on 52 subjects, analysed the resulting clustered binary data using logistic regression analyses.

Similarly, Clark et al. (1989), in their study on tumor progression used multi-variable logistic regression technique to develop a prognostic model for primary, clinical stage I cutaneous melanoma. The model, so developed, is 89% accurate in predicting survival of tumor patients.

In one instance, Gordon et al. (1984), used standard logistic regression analysis together with Cox hazard in their study on coronary risk factors and exercise test performance in asymptomatic hypercholesterolemic on 6850 whites to ascertain the prevalence of ischemic electrocardiographic in whites.

In another instance too, Geller et al. (2009), conducted a study on 227 males of age at least 40 years, who are with invasive melanoma. The aim was to determine the factors associated with physician discovery of early melanoma in middle-aged and older men. Odds ratio was adapted as one of the techniques in the analysis.

Also, Ito, Nishimura, Saito and Omori (1997), in their attempt to determine the level of erythrocyte aldose reductase protein (AR-p) in diabetic patients by a two-site enzyme-linked immunosorbent assay, classified 95 non-insulin-dependent diabetes mellitus (NIDDM) patients into two groups, based on the results of seven nerve function tests: Group I, without demonstrable neuropathy and Group II, with overt neuropathy. Multivariate logistic regression analysis was subsequently used to identify two independent risk factors for overt neuropathy in diabetic patients.

In one occasion, Takahashi et al. (1998), in their study to investigate the risk factors for diabetic severe neuropathy independent of glycemic control and duration of diabetes, used logistic regression analysis technique to establish that maximum body mass index (BMI) in the past minus present BMI and the level of erythrocyte aldose reductase protein together with measurement of erythrocyte AR level may be useful for predicting severe neuropathy in non-insulin-dependent diabetes mellitus (NIDDM).

In another occasion too, Smith (2005), suggested a simple approach to a study conducted by Platt on gestational-age-specific mortality entitled: "Research to date on perinatal outcomes has all but ignored the fact that gestational age is a time-to-event variable". Smith used logistic regression for intrapartum stillbirth and neonatal death to arrive at this simple approach.

Again, Kennedy et al. (1980), in their effort to better understand the clinical and angiographic characteristics predictive of operative mortality (OM), used multivariate discriminant analysis technique to analyse data on isolated coronary artery bypass grafting (CABG) operations on 6,176 patients carried out by "Collaborative Study in Coronary Artery Surgery" (a multi-institutional study of the medical and surgical treatment of coronary artery disease) to come out with the variables associated with OM.

In a similar situation, Wright et al. (1987), conducted a study on 9,000 patients from the operational database of Loyola Open-Heart Registry of those who had undergone coronary bypass or cardiac valve replacement from January 1970 to December 1984. The data was analysed using multivariate discriminant analysis to identify and quantify those factors that might predict operative mortality (OM) for patients undergoing coronary artery bypass grafts at Loyola University Medical Center.

Additionally, Wang, Xiao, Ren, Li and Zhang (2007), tried to assess and confirm the risk factors for mortality after coronary artery bypass grafting (CABG) operations so as to map out proper guidance of surgical strategy, especially in patients with low left ventricular ejection fraction (LVEF) in domestic polyclinics in China. A data of 5048 consecutive patients who underwent CABG from December 1999 through August 2005 at Peking University First Hospital, Beijing, was analysed using univariate and multivariate stepwise logistic regression analysis to identify 22 candidate factors for their association with perioperative death.

Another example of the use of logistic regression analysis is a study conducted by Abrahamowicz, du Berger and Graver (1997) - They examined

the impact of lipids and other risk factors on coronary heart disease. With a random sample of 2, 512 patients, from Lipid Research Clinics (1972–1987), who did not take lipid-lowering medications, logistic regression analysis was used to assess the potential impact of the lipids and continuous risk factors.

Pullinger, Seligman and Gornbein (1993), used multiple logistic regress analysis for common occlusal features for assumptomatic controls on 147 patients to establish that certain features such as anterior open bite in osteoarthrosis were consequence rather than etiological factors for the disorder.

Hirota (1999), also identified the discriminant analysis as a proper method to diagnose and treat liver diseases. In 1967, he applied the discriminant analysis by computer to diagnose liver diseases and it was, in that case, superior to diagnosis by physicians. He has, since 1990, applied the discriminant analysis in treadmill diagnosis and concluded the utility in prospective study.

In one breath, Perez (2006), used clustering and discriminant analysis of image quality to develop a mathematical model for activity optimization in Nuclear Medicine Studies (NMS). The application of this method yields results that is consistent with the application of Received Operating Characteristic (ROC) analysis, and has successful results in the reduction of the administered activity in planar studies of nuclear medicine.

Laika, (2003), also collected data of short-latency somatosensory-evoked potentials (SSEPs) on 91 patients with diabetes mellitus (DM) after median nerve stimulation. The patients were divided into three groups: (1) patients without neuropathy, (2) patients with mild neuropathy, and (3)

patients with severe neuropathy. The data which consisted of 26 independent variables was subsequently analysed, using discriminant analysis, to establish that six of the independent factors can best differentiate the groups.

Again, Walter, Feinstein and Wells (1987), presented a coding scheme for ordinal independent variables useful in dose-response analyses which can also be used in evaluating the survival of lung cancer patients. This scheme uses various forms of regression analysis including logistic regression.

Additionally, Abbott (1985), undertook a survival analysis on data from the Framingham Heart Study. He realised that when event times are grouped into intervals, logistic regression can be adapted to the analysis of such data by modeling the interval when an event occurs. It was furthermore shown that results from such an adaptation will often lead to parameter estimates close to those obtained by the proportional hazards model in the grouped event time setting.

Logistic regression analysis and discriminant analysis techniques have also found their way in social sciences. Kirschenbaum, Oigenblick and Goldberg (2000), conducted a study to examine the differences between two groups of Israeli workers. One of the groups comprised of 77 workers who had suffered a first-time work injury and the other comprised of 123 workers who had suffered work injuries on multiple occasions. It was clear that the multivariate relationship between independent variables and the work injuries were unlikely due to chance. Logistic regression analysis was used for the analysis to obtain a hit ratio of 70%.

Diehl. Elnick, Bourbeau and Labouvie-vief (1998) undertook classification analysis, given very different numbers. This analysis was executed using discriminant analysis.

Logistic regression and discriminant analysis have also gained some popularity in both accounting and agriculture management. Argilés (1998), examined the current use of accounting in agriculture with the aim of helping agricultural agents in decision making. Two logit models, one with non-financial variables and the other with financial ones, were applied to subsamples of viable and unviable farms in Catalonia, Spain.

Similarly, Ahmed, K. F. Alam, and M. Alam. (1997) conducted a survey on 295 students from five universities in New Zealand to examine the influence of intrinsic factors on whether accounting students choose to pursue a chartered accountancy (CA) career or a non-accounting career. The objective of the study was to help recruitment into the accounting profession in New Zealand. Discriminant analysis was employed to reveal the factors differentiating the two groups.

Again, Gestel, et al. (2004), conducted financial analysis of the creditworthiness of a potential client. The aim was to avoid taking wrong decisions that may result in foregoing their valuable clients if not given credit to them or, more severely, in substantial capital losses if the clients subsequently default. They observed that when nonlinear kernel-based classifiers is applied to a real-life data set concerning commercial credit granting to mid-cap Belgian and Dutch firms, it yields a better performance but acknowledged that many studies in this line focus on the use of financial

ratios in linear statistical models, such as linear discriminant analysis and logistic regression analysis.

CHAPTER THREE

REVIEW OF LOGISTIC REGRESSION ANALYSIS

.

Introduction

Logistic regression analysis (or simply logistic regression) is part of a category of generalised linear models. It is a type of multivariate regression that has a predictive model that can be used when the target variable is a categorical variable. The technique aims at modeling the relationship between a set of independent variables and the probability that a case is a member of one of the categories of the dependent variables.

Logistic regression has many uses- It is used to predict a dependent variable on the basis of continuous and/or categorical independents; to determine the percentage of variance in the dependent variable explained by the independents; to rank the relative importance of independents; to assess interaction effects; and to understand the impact of covariate control variable. (Garson, n.d)

Generally, the dependent or response variable in logistic regression is dichotomous, such as presence/absence or success/failure but the multinomial logistic regression also exists to handle situations with more than two dependent variables such as low/medium/high. (McCullagh, & Nelder, 1989)

Purpose for conducting Logistic Regression Analysis

The purpose of performing logistic regression analysis is to:

1. Analyze the relationship between metric independent variables and a dichotomous dependent variable.
2. Find out if there is a relationship, using the information in the independent variables, to improve the accuracy in predicting values for the dependent variable.
3. Find the chances of an object, subject or entity to be a member of a particular group.
4. To classify new objects, subjects and entities. (Luna, n.d)

Types of Logistic Regression

There are two types of logistic regression:

1. Binary logistic regression which is used for two groups.
2. Multinomial Logistic Regression that can be used with more than two groups.

Binary Logistic Regression

Binary Logistic Regression is a predictive model that can be used when the target variable is a categorical variable with two categories (dichotomous) – for example live/die, has disease/doesn't have disease, purchases product/doesn't purchase, wins race/doesn't win, etc., and the independents variables are of any type. Binary logistic regression has other application of combining the independent variables to estimate the probability that a particular event will occur, i.e. a subject will be a member of one of the

groups defined by the dichotomous dependent variable. The variate or value produced by binary logistic regression is a probability value between 0 and 1. If the probability for group membership in the modelled category is above some cut point (usually 0.5), the subject is predicted to be a member of the modelled group. If the probability is below the cut point, the subject is predicted to be a member of the other group.

For any given case, logistic regression computes the probability that a case with a particular set of values for the independent variable is a member of the modelled category. A case is predicted to belong to the group associated with the highest probability. Predicted group membership can be compared to actual group membership to obtain a measure of classification accuracy. (Luna, n.d)

Level of Measurements Required for Binary Logistic Regression

Logistic regression can be used only with two types of target variables:

1. A categorical target variable that has exactly two categories (i.e., a binary or dichotomous variable)
2. A continuous target variable that has values in the range 0.0 to 1.0 representing probability. The dependent variable in logistic regression is usually dichotomous, that is, the dependent variable can take the value say, 1 with a probability of success θ, or the value 0 with probability of failure $1-\theta$. This type of variable is a Bernoulli (or binary) variable. (Statgun, n.d)

Multinomial Logistic Regression

Multinomial logistic regression is an extension of binary logistic regression and it allows simultaneous comparison of more than two contrast (i.e. three or more contrasts are estimated simultaneously). The relationships between a non-metric dependent variable and metric or dichotomous independent variables can be analysed using the multinomial logistic regression.

This method compares multiple groups through a combination of binary logistic regressions and the group comparisons are equivalent to the comparisons for a dummy-coded dependent variable with the group having the highest numeric score used as the reference group. For example, if one wants to study differences in BSc, MSc, and PhD students using multinomial logistic regression, the analysis would compare BSc students to PhD students, MSc. students to PhD students and BSc students to MSc students. For each dependent variable, there would be two comparisons. (Statgun, n.d)

Level of Measurement for Multinomial Logistic Regression

1. Multinomial Logistic regression analysis requires that the independent variables be metric or dichotomous.
2. If an independent variable is nominal level and not dichotomous, the variable is dummy coded
3. Multinomial Logistic regression can also be applied to ordered categories (ordinal data), that is, variables with more than two ordered categories, as a dependant variable.

Methods for Including Variables in the Logistic Regression Analysis

The goal of logistic regression is to correctly predict the category of outcome for individual cases using the most parsimonious parameters. To accomplish this goal, a model is created that includes all predictor variables that are useful in predicting the response variable. Several different options are available for variables selection during model creation but the three most commonly used methods for selecting variables into the logistic regression equation are:

1. The standard or simultaneous method
2. The hierarchical method.
3. The stepwise method.

Standard or Simultaneous Regression Method

Standard or simultaneous regression method helps to evaluate the relationship between a set of independent variables and a dependent variable. This method enters all the independent variables into the logistic regression equation at the same time. Multiple R (coefficient of determination) or R^2 (coefficient of regression) is employed to measure the strength of the relationship between the set of independent variables and the dependent variable. An *F*-test is used to determine if the relationship can be generalised to the population represented by the sample. A *t*-test is used to evaluate the individual relationship between each independent variable and the dependent variable.

Hierarchical Method

Hierarchical method requires that the control variables are entered in the analysis before the predictors whose effects are the primary concern. In hierarchical method, the independent variables are entered in two stages. In the first stage, the independent variables that one wants to control for are entered into the regression equation. In the second stage, the independent variables whose relationship we want to examine, after the controls, are entered. A statistical test of the change in R^2 from the first stage is used to evaluate the importance of the variables entered in the second stage.

Stepwise Multiple Regression

Stepwise method attempts to identify the subset of independent variables that has the strongest relationship to a dependent variable. In this method, variables are selected in the order in which they maximize the statistically significant contribution to the model. Stepwise regression attempts to find the most parsimonious set of predictors that are most effective in predicting the dependent variable. Variables are added to the regression equation one at a time, using the statistical criterion of maximizing the R^2 of the included variables. The process is completed when none of the possible addition can make a statistically significant improvement in R^2.

(Statgun, n.d)

Assumptions of the Logistic Regression Analysis

1. Logistic regression does not assume a linear relationship between the dependents and the independents. It may handle nonlinear effects even

when exponential and polynomial terms are not explicitly added as additional independents because the logit link function in the logistic regression equation is non-linear. However, it is also possible and permitted to add explicit interaction and power terms as variables in the logistic equation.

2. The dependent variable in the logistic regression analysis need not be normally distributed (but does assume its distribution is within the Poisson, binomial or gamma).
3. The dependent variable need not be homoscedastic for each level of the independents; (thus), there is no homogeneity of variance assumption:- variances need not be the same within categories.
4. Normally distributed error terms are not assumed.
5. Logistic regression does not require that the independents be interval.
6. Logistic regression does not require that the independents be unbounded. (Statgun, n.d)

The Logistic Regression Model

The logistic model or formula computes the probability of the selected response as a function of the values of the predictor variables. If a predictor variable is a categorical variable with two values, then one of the values is assigned the value 1 and the other is assigned the value 0. If a predictor variable is a categorical variable with more than two categories, then a separate dummy variable is generated to represent each of the categories except for one which is excluded. The value of the dummy variable is 1 if the variable has that category, and the value is 0 if the variable has any other

category; hence, no more than one dummy variable will be 1. If the variable has the value of the excluded category, then all of the dummy variables generated for the variable are 0.

In summary, the logistic formula has each continuous predictor variable, each dichotomous predictor variable which are coded 0 or 1, and a dummy variable for every category of predictor variables with more than two categories less one category. (Statgun, n.d)

Distribution of the Logistic Regression Model

Since the response variable (y_i) for logistic regression is always binary (assuming only two values), its distribution is binomial.

i.e.

$$y_i \sim B(n_i, \pi_i)\text{, } (i = 0 \text{ or } 1),$$

where the numbers of Bernoulli trials, $n_{i,}$ are known and π_i which is unknown is the probability of success or being in one group(1), and $(1-\pi_i)$ is the probability of failure to be in modelled group or being in the other group(0).

The binomial distribution has distribution function

$$f_i(y_i) = \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i}$$

Taking natural log on the equation above gives

$$In\, f_i(y_i) = In \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i}$$

$$\Rightarrow In\, f_i(y_i) = y_i In\, \pi_i + (n_i - y_i) In (1-\pi_i) + In \binom{n_i}{y_i}$$

$$\Rightarrow In f_i(y_i) = y_i\, In\pi_i + n_i In(1-\pi_i) - y_i In(1-\pi_i) + In\binom{n_i}{y_i}$$

$$\Rightarrow In f_i(y_i) = y_i In\frac{\pi_i}{1-\pi_i} + n_i In(1-\pi_i) + In\binom{n_i}{y_i}$$

$$\text{Let } \theta_i = In\frac{\pi_i}{1-\pi_i}$$

$$\Rightarrow e^{\theta_i} = \frac{\pi_i}{1-\pi_i} \tag{1}$$

$$\Rightarrow e^{\theta_i} - e^{\theta_i}\pi_i = \pi_i$$

$$\Rightarrow \pi_i = \frac{e^{\theta_i}}{1+e^{\theta_i}} \tag{2}$$

and

$$1-\pi_i = 1-\frac{e^{\theta_i}}{1+e^{\theta_i}}$$

$$= \frac{1}{1+e^{\theta_i}} \tag{3}$$

Therefore,

$$In(1-\pi_i) = In\frac{1}{1+e^{\theta_i}}$$

$$\Rightarrow In(1-\pi_i) = -In(1+e^{\theta_i})$$

$$\text{Let } \; b(\theta_i) = n_i In(1+e^{\theta_i}) = -n_i In(1-\pi_i) \tag{4}$$

and

$$c(y_i,\phi) = In\binom{n_i}{y_i}, \text{ (since } In\binom{n_i}{y_i} \text{ is not a function of } \theta_i)$$

Therefore, $Inf_i(y_i)$ can be expressed in the form

$Inf_i(y_i) = \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_{i,}\phi)$, which is the form of the exponential family.

The Mean of the Distribution

From (4),

Differentiating $b(\theta_i)$ with respect to θ_i, the mean (μ_i), is given

$\mu_i = \frac{\partial}{\partial\theta_i}\left[n_i In(1+e^{\theta_i})\right] = n_i \frac{e^{\theta_i}}{1+e^{\theta_i}} = n_i\pi_i$, which is the mean of a binomial distribution

The Variance of the Distribution

The second differential of $b(\theta_i)$ with respect to θ_i yields the variance, v_i .

Thus,

$$v_i = \frac{\partial^2}{\partial\theta_i^{\,2}}\left[n_i In(1+e^{\theta_i})\right] = n_i \frac{(1+e^{\theta_i})e^{\theta_i} - e^{2\theta_i}}{(1+e^{\theta_i})^2}$$

$$= n_i \frac{e^{\theta_i}}{(1+e^{\theta_i})^2}$$

$$= n_i \frac{e^{\theta_i}}{(1+e^{\theta_i})} \cdot \frac{1}{(1+e^{\theta_i})}$$

$= n_i\pi_i(1-\pi_i)$, which is the variance, v_i

Derivation of the Logistic Regression Formula

From (2)

$$\pi_i = \frac{1}{(1+e^{\theta_i})e^{-\theta_i}}$$

$$\Rightarrow \pi_i = \frac{1}{(1+e^{-\theta_i})} \tag{5}$$

Thus, from (2) and (5)

$$f(\theta_i) = \frac{e^{\theta_i}}{1+e^{\theta_i}} \tag{6}$$

or

$$f(\theta_i) = \frac{1}{1+e^{-\theta_i}} \tag{7}$$

The variable θ_i is given by

$$\theta_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \tag{8}$$

From (6) and (8),

the model can equivalently be formulated as

$$f(\theta) = \pi_i = \frac{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}}{1+e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}} \tag{9}$$

Also from (7) and (8), the model takes the form

$$f(\theta_i) = \pi_i = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}} \tag{10}$$

Equation (9) or (10) is the general logistic regression model with risk factors $x_1, x_2, x_3, \ldots, x_k$.

The computed value π_i is a probability in the range from 0 to 1, β_0 is the intercept and $\beta_1, \beta_2, \beta_3, \ldots \beta_k$ are the regression coefficients of the variables $x_1, x_2, x_3, \ldots, x_k$ respectively.

The intercept is the value of θ_i when the value of all the other independent variables are zero (i.e., the value of θ_i in response with no independent variable). Each of the regression coefficients describes the size of

the contribution of that risk factor. A positive regression coefficient is an indication that that risk factor increases the probability of the outcome, while a negative regression coefficient means that that risk factor decreases the probability of the outcome. A large regression coefficient means that that risk factor strongly influences the probability of that outcome; while a near-zero regression coefficient means that the risk factor has little influence on the probability of that outcome. (Wikipedia, n.d)

The Logistic Regression Curve

The relationship between the predictor and response variables is not a linear function in logistic regression.

A plot o $f(\theta_i)$ against θ_i is as below:

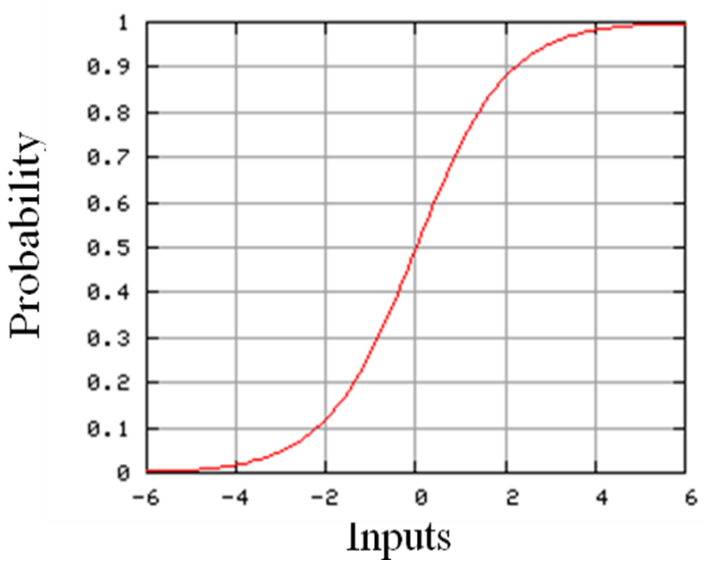

Figure 1: The Logistic Regression Curve

The logistic regression function is useful because it can take as an input, any value from negative infinity to positive infinity, whereas the output is confined to values between 0 and 1. The variable θ_i is the exposure to some

set of risk factors, while $f(\theta_i)$ represents the probability of a particular outcome, given that set of risk factors. The variable θ_i is a measure of the contribution of all the risk factors used in the model.

(Wikipedia, n.d).

Numerical Problems with Logistic Regression

The maximum likelihood method used to estimate the logistic regression coefficients is an iterative fitting process that attempts to cycle through repetitions to find an answer. Sometimes, the method breaks down and will not be able to converge or find an answer. Sometimes too, the method produces wildly improbable results, reporting that a one-unit change in an independent variable increases the odds of the modelled event by hundreds of thousands or millions. These implausible results can be produced by multi-collinearity, categories of predictors having no cases or zero cells, and complete separation whereby the two groups are perfectly separated by the scores on one or few independent variables. (Luna, n.d).

Just like linear regression, logistic regression gives each regressor a coefficient b_i which measures the regressor's independent contribution to variations in the dependent variable. But there are technical problems with logistic regression. What one wants to predict from knowledge of relevant independent variables is not a precise numerical value of a dependent variable, but rather the probability (π) that it is one group rather that the other. It is not possible to use this probability as the dependent variable in an ordinary regression, (i.e. as a simple linear function of regressors) due to these two reasons: Firstly, numerical regressors may be unlimited in range. If the idea

for one is to express π as a linear function of the predictors, the person may find himself predicting π that may be greater than 1 or less that 0, which cannot be true, as probabilities can only take values between 0 and 1.

Secondly, there is a problem of additivity. Imagine one tries to predict success at a task from two dichotomous variables, training and gender. Among untrained individuals, 50% of men from men population succeed and 70% of women from women population succeed. Among trained individuals, 90% of men from men population succeed and 40% of women from trained women succeed. If π is thought of as a linear function of gender and training, then the proportion of women who succeed in the study (i.e. 70% + 40% = 110%) would have to be estimated (which again cannot be true).

Due to the fact that using logistic regression to generate predicted probabilities can produce values outside the 0 to 1 range and also forcing linearity on what is more likely an S-shaped relation, violates the assumption that the components of the composite variable are additive, and violates the assumptions of normality and homoscedasticity required for statistical tests. Therefore, there is the need to seek for an alternative strategy. (Wikipedia, n.d).

Odds Function

The odds function is the strategy that is often used to streamline the work of logistic regression. The odds function makes use of which odds of an event is which is defined as the ratio of the probability that an event occurs to the probability that it fails to occur.

Thus,

$$\text{Odds}(\text{case}=1) = \frac{\text{p}(\text{case}=1)}{1-\text{p}(\text{case}=1)} = \frac{\pi_1}{1-\pi_1}$$

or

$$\text{Odds}(\text{case}=1) = \frac{\text{p}(\text{case}=1)}{\text{p}(\text{case}=0)} .= \frac{\pi_1}{1-\pi_1}$$

Link Function

The link function provides the relationship between the linear predictor and the mean of the distribution function. When the response variables, $f(\theta_i)$, are binary, taking on only values 0 and 1, (as in the case of the logistic regression), the distribution function is generally chosen to be the binomial distribution and the interpretation of μ_i is then the probability, π , of $f(\theta_i)$, being in the modelled group.

Logit Transformation of $\boldsymbol{\theta_i}$

From (1) above, taking natural logarithm on both sides gives

$$\theta_i = \ln\left[\frac{\pi_i}{1-\pi_i}\right] = \log(\text{odds})$$

but

$$\theta_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

$$\Rightarrow \ln\left[\frac{\pi_i}{1-\pi_i}\right] = \log(\text{odds}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k = \text{logit}(\pi),$$

which is the link function (logit link) for the logistic regression.

This indicates that the logistic regression requires that observations be independent and that the independent variables be linearly related to the logit of the dependent. (Menard, 2002).

The logistic regression strategy retains the goal of generating predicted probabilities when it is expressed in logit form and the problems that may arise, as a result of predicting a probability greater than 1 or less than 0, are eliminated. Thus, logistic regression involves fitting to a data an equation of the form:

$$\text{logit}(\pi) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_k x_k$$

Logit(π) ranges from negative infinity to positive infinity and its scale is symmetrical around the logit of 0.5 (which is zero).

An example of some probabilities (π) and their corresponding logit probability values are as below.

Table 1: The Relationship Between Probability of Success (π) and Logit(π)

π	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Logit(π)	-2.20	-1.39	-0.85	-0.41	0.00	0.41	0.85	1.39	2.20

Because log(odds) take on any value between $-\infty$ and $+\infty$, the coefficients for logistic regression equations can be interpreted in the usual way. Thus, they represent the change in log(odds) of the response per unit change in the predictor.

A plot of log(odds) = logit(π) = $b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_k x_k$ is as below:

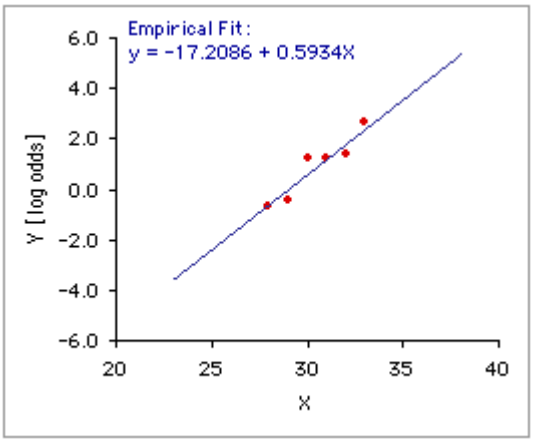

Figure 2: A Plot of Log (odds) against θ_i

Odds are asymmetric. When the roles of the two outcomes are switched, each value in the range 0 to 1 is transformed by taking its inverse (1/value) to a value in the range 1 to $+\infty$. For example, if the odds of having a disease is 1/4, the odds of not having the disease is 4/1. When the roles of the two outcomes are switched, the log(odds) are multiplied by -1, since *log*(a/b) = -log(b/a). For example, if the log(odds) of passing an examination is -1.39, the log(odds) of failing the examination is 1.39. As the probability of something increases, the odds and log (odd) too increase and vise versa. (Luna, n.d)

Loss Function

A loss function is a measure of fit between a mathematical model of data and the actual data. The parameters chosen for models are to minimize the badness-of-fit or to maximize the goodness-of-fit of the model to the data. With least squares method, parameters are chosen such that the sum of squares of residual (SS_{res}) is minimal or the sum of squares due to regression (SS_{reg},) is maximum. For logistic model, there is no mathematical solution that will

produce least squares estimates of the parameters. The loss function chosen is maximum-likelihood. (Luna, n.d)

The Maximum-Likelihood Function

Although logistic regression finds a "best fitting" equation just as linear regression does, the principles on which it does so are rather different. Instead of using a least-squared deviations criterion for the best fit, as in the case of linear regression, the logistic regression analysis uses a maximum likelihood method to compute the coefficients for the logistic regression equation. The maximum-likelihood parameters are chosen to maximize the likelihood. The maximum likelihood is conditional probability (i.e. $P(i/\theta_i)$, the probability of being in group *i* given the set of risk factors in group *i* (where *i* = 1 or 0).

The techniques actually employed to find the maximum likelihood estimates is iterative Newton-Raphson algorithm which attempts to find coefficients that match the breakdown of cases on the dependent variable. The maximum-likelihood estimation procedure successively tries to get closer and closer to the correct answer and it iterates until the absolute value of the largest parameter change is less than the value specified for “Tolerance” on the logistic regression modeling. (Wikipedia, n.d).

Measures and Significance Tests

Likelihood is probability under a specified hypothesis. Logistic regression is considered with two hypotheses that are of interest: The null hypothesis which is that all the coefficients in the regression equation take the

value zero, and the alternate hypothesis that the model currently under consideration is accurate. The likelihood of observing the exact data under each of these alternate hypotheses is nearly always a frighteningly small number, and to make it easier to handle. Natural logarithm is taken to give log likelihood. (Wikipedia).

Badness of Fit in Logistic Regression

Customarily, the natural logs of the probability (likelihood) of the results are multiplied by -2 to make the result positive. The statistic -2loglikelihood (-2 multiply by the log of the likelihood) is a badness-of-fit indicator. This indicates that a large number means a poor fit of the model to the data. The value $-2\text{loglikelihood}_{\text{R}}$ is a measure of the error associated in trying to predict the dependent variable without using any information from the independent variables.

Testing the Importance of Variables

The importance of variables in the logistic regression model can be tested using the hypothesis

$H_0 : \beta_i = 0$ (x_i not important in the model)

$H_1 : \beta_i \neq 0$ (x_i is important in the model), for the i^{th} independent variable.

The variable under consideration is first included in the model and the -2loglikelihood of the full model ($-2\text{loglikelihood}_{\text{F}}$) is found. The variable is then excluded and -2loglikelihood of the restricted model

$(-2\text{log likelihood}_R)$ of the model is again measured. The difference between the two results follows a χ^2 distribution. Thus,

$$\chi^2 = -2log\text{likelhood}_R - (-2\text{log likelihood}_F) = -2\ln\left(\frac{\text{likelihood}_R}{\text{likelihood}_F}\right) \qquad (11)$$

This chi-square (χ^2) statistic is used to statistically test whether including a variable reduces badness-of-fit measure. If chi-square is significant, the variable is considered to be a significant predictor in the equation. If not, the variable is considered unimportant and can therefore be excluded from the logistic regression model.

Overall Test of Relationship (Assessing Model Fit)

There are a number of statistics available for testing the relationship of the model and the variables. The null and alternative hypotheses for assessing overall model fit are given by

H_0 : The hypothesized model fits the data

H_1 : The hypothesized model does not fit for the data

The overall measure of how well the model fits to variables under consideration is given by the likelihood value. The overall test of relationship among the independent variables and groups defined by the dependent is based on the reduction in the likelihood values for a model which does not contain any independent variables and the model that contains the independent variables. A model that fits the data well has a small likelihood value. A perfect model would have a likelihood value of zero. The difference in likelihood follows a chi-square distribution. The significance test for the model chi-square (as in 11) is the statistical evidence of the presence of a

relationship between the dependent variable and the combination of the independent variables. (Wikipedia, n.d)

Other Methods for Significance Test

Hosmer and Lemeshow's goodness-of-fit test

Hosmer and Lemeshow's (H-L) goodness-of-fit test divides subjects into deciles based on predicted probabilities. It then computes a chi-square from observed and expected frequencies then a probability (*p*) value is computed from the chi-square distribution with 8 degrees of freedom to test the fit of the logistic model. If the H-L goodness-of-fit test statistic is greater than .05, (which indicates well-fitting models), we fail to reject the null hypothesis that there is no difference between observed and model-predicted values, implying that the model's estimates fit the data at an acceptable level. (Hosmer, et al., 1988).

Omnibus Tests of Model Coefficients

Omnibus test is an alternative to the Hosmer-Lemeshow test. It tests if the model with the predictors is significantly different from the model with only the intercept. The omnibus test is interpreted as a test of the capability of all predictors in the model jointly to predict the response (dependent) variable. Significance corresponds to the conclusion that there is adequate fit of the data to the model, meaning that at least one of the predictors is significantly related to the response variable. (Luna, n.d)

Wald Statistic

The Wald statistic is also another alternative test which can be used to test the significance of individual logistic regression coefficients for each independent variable (that is, to test the null hypothesis in logistic regression that a particular logit (effect) coefficient is zero). The Wald statistic is the squared ratio of the unstandardized logistic coefficient to its standard error.

Testing for Influential Case

It is important to determine whether a case is influential in the logistic model. Cook's distance statistic which is used as measure of influence of a case in linear multiple regressions is the technique used to measure the influence of a case on the solution in logistic regression. However, the criteria for determining that a case is an influential case in logistic regression differ from the criteria in multiple regressions. In logistic regression, a case is identified as influential if its Cook's distance is greater than 1.0. (Hosmer and Lemeshow, 2000).

Testing for Outliers

Logistic regression modelling requires the inclusion of only variables relevant to the course of the modeling. If there is an item which can unduly influence the model, it must be eliminated in the modeling process. The elimination can be done by using the residual process. The residual in predicting a case is the difference between the actual probability and the predicted probability for a case. For example, if the predicted probability for a case that actually belonged to the modelled category was 0.80, the residual

would be 1.00 – 0.80 = 0.20. The residual is standardised by dividing it by an estimate of its standard deviation. If a standardised residual is larger than 3.0 or smaller than -3, it is considered an outlier, and a candidate for exclusion from the analysis. (Wikipedia, n,d)

Strength of Logistic Regression Relation

Logistic regression computes correlation measures to estimate the strength of the relationship (pseudo R square measures, such as Nagelkerke's R^2 can be used). However, these correlations measures do not really tell us much about the accuracy or errors associated with the model. A more useful measure to assess the utility of a logistic regression model is classification accuracy, which compares predicted group membership based on the logistic model to the actual, known group membership, which is the value for the dependent variable.

Testing Accuracy (Efficiency) of Logistic Model (80 and 20 Strategy)

The accuracy of logistic regression model can be tested using the 80-20 strategy. In this validation strategy, the cases are randomly divided into two subsets: a training sample containing 80% of the cases and a holdout sample containing the remaining 20% of the cases. The training sample is used to derive the logistic regression model. The holdout sample is classified using the coefficients based on the training sample. The classification accuracy for the holdout sample is used to estimate how well the model based on the training sample will perform for the population represented by the data set. If the classification accuracy rate of the holdout sample is within 10% of the training

sample, it is deemed sufficient evidence of the utility of the logistic regression model.

In addition to satisfying the classification accuracy, it is required that the significance of the overall relationship of the dependent and the independent variables and the relationships with individual predictors for the training sample match the significance results for the model using the full data set. (Wikipedia, n.d)

Evaluating Usefulness of Logistic Model

Even if the independent variables had no relationship to the groups defined by the dependent variable, one would still expect to be correct in his predictions of group membership some percentage of the time. This is known as by chance accuracy. The estimate of "by chance accuracy" that is often used is the "proportional by chance accuracy rate", and its the sum of the squared percentage of chances of cases in each group**.** To characterise a model as useful, the overall percentage accuracy rate compared to the proportional by chance accuracy is some percentage more (i.e. logistic regression model is classified as useful if there is some percentage (set) improvement over the rate of accuracy achievable by chance alone. (Garson, n.d)

CHAPTER FOUR

REVIEW OF DISCRIMINANT ANALYSIS

Introduction

Discriminant analysis is a technique for predicting and classifying a set of observations into predefined classes. It is used to determine which continuous variables best discriminate between two or more natural occurring groups. The model is built based on a set of observations for which the classes are known and this discriminant function is used to predict the class of a new observation with unknown class.

Discriminant Analysis may be used either to assess the adequacy of classification, given the group memberships of the objects under study; or to assign objects to one of a number of (known) groups of objects. It thus has a descriptive or a predictive objective.

It is one of the available techniques for developing a rule or model to enable one identify or discriminate between cases based on the rule or the underlying principle. This technique is used for analyzing data when response variables are categorical and the predictor variables are interval scaled. (Wikipedia, n.d)

Objectives of Discriminant Analysis

The main objectives for performing discriminant analysis are:

1. Identify the variables that best discriminate between groups using the most parsimonious way (i.e. to determine most influential predictors).
2. To use the identified variables or factors to develop a good classification function that is linear combination of the predictor variables and would be reliable in classification cases.
3. Undertake discriminant classification analysis which classifies cases into groups and also to assign new objects to one of a number of known groups thereby validating the predictive function
4. To test theory by observing whether cases are classified as predicted.
5. To assess the relative importance of the independent variables in classifying the dependent variable
6. Examine whether significant differences exist among groups.
7. To determine the percentage of variance in the dependent variable explained by the independents over and above the variance accounted for by control variables. (Garson, n.d)

Types of Discriminant Analysis

There are two types of discriminant analysis namely

1. Linear Discriminant Analysis (LDA) also known as Discriminant Analysis (DA) or Two-group discriminant analysis.
2. Multiple Discriminant Analysis (MDA)/ Multiple-group discriminant analysis.

Linear Discriminant Analysis

Linear discriminant analysis or simply discriminant analysis (DA), is used to classify cases into the values of a categorical dependent, usually a dichotomy and it also explicitly attempts to model the difference between two classes of data. If the discriminant function analysis is effective for a set of data, the classification table estimates will yield a high percentage of correct classification. LDA is closely related to ANOVA (analysis of variance) and regression analysis, which also attempt to express one dependent variable as a linear combination of other features or measurements. In the ANOVA and regression analysis, however, the dependent variable is a numerical quantity, while for LDA it is a categorical variable (*i.e.* the class label).

LDA is also closely related to principal component analysis (PCA) and factor analysis in that they all look for linear combinations of variables which best explain the data. LDA explicitly attempts to model the difference between the classes of data. PCA on the other hand does not take into account any difference in class, and factor analysis builds the feature combinations based on differences rather than similarities. Discriminant analysis is also different from factor analysis in that it is not an interdependence technique. (Abdi, 2007)

Multiple-Group Discriminant Analysis

Multiple discriminant analysis (MDA) is an extension of linear discriminant analysis and relates to multiple analysis of variance (MANOVA), sharing many of the same assumptions and tests. Discriminant analysis is just the inverse of a one-way MANOVA. The levels of the independent variable

(or factor) for MANOVA become the categories of the dependent variable for discriminant analysis, and the dependent variables of the MANOVA become the predictors for discriminant analysis.

In MANOVA one asks whether group membership produces reliable differences on a combination of dependent variables. If the answer to that question is 'yes', then clearly, that combination of variables can be used to predict group membership. Mathematically, MANOVA and discriminant analysis are the same. We can compare those two matrices via multivariate *F* tests in order to determine whether or not there are any significant differences (with regard to all variables) between groups. This procedure is identical to multivariate analysis of variance or MANOVA. (Abdi, 2007).

Multiple discriminant analysis is used to classify a categorical dependent which has more than two categories, using as predictors a number of interval or dummy independent variables. For example, the management of a telephone company is interested in identifying characteristics that best discriminate among households that have one, two and three or more phone lines. Here, the interest is in identifying more than two groups.

The objectives of the multiple-group discriminant analysis are the same as that of two-group discriminant analysis except that in the case of two-group discriminant analysis, only one discriminant function is required to represent all of the differences between the two groups, but in the case of multiple group discriminant analysis, it may not be possible to represent or account for all of the differences among the groups by a single discriminant function, making it necessary to identify additional discriminant function(s). Thus, an additional objective of multiple-group discriminant analysis is to

identify the minimum number of discriminant functions that will provide most of the discrimination among the groups. (Abdi, 2007)

Methods for Selecting the Best Set of Variables for Discriminant Model

In some applications of discriminant analysis, there are data availability on a large number of variables. In this case, it is desirable to select relatively small subsets of variables that would contain almost as much information as the original collection.

One major objective for performing discriminant analysis is to come out with the best set of variables that can be used to develop a function or a model for future predictions and classifications. The question then is, how can one identify or select the best potential discriminator variables than can be used to form a discriminant function?

The following techniques are mostly used to select the best set of discriminating variables to form discriminant function(s) for future predictions and classifications:

1. The forward selection.
2. The backward selection.
3. The stepwise selection.

The Forward Selection

The forward selection method enters first the variable that provides the most discrimination between the groups as measured by a given statistical criterion. In the next step, the variable entered is the one that adds maximum amount of additional discriminating power to the discriminant function as

measured by the statistical criterion. The procedure continues until addition of new variable does not significantly change the model. (Statgun, n.d).

The Backward Selection

The backward selection begins with all the variables in the discriminant function. At each step, one variable is removed (that one being the one that provides the least amount of decrease in the discriminating power, as measured by statistical criterion). If the removal of that variable has no significant effect on the model, as revealed by the statistical criterion, it is excluded from the modelling. However, if its removal has a significant effect in the discriminating power, it is maintained in the modelling. The procedure continues until no more variables can be removed. (Statgun, n.d)

The Stepwise Selection

Stepwise selection is a combination of the forward and backward elimination procedures. It begins with no variables in the discriminating function, and then at each step a variable is either added or removed. A variable already in the discriminant function is removed if it does not significantly lower the discriminating power, as measured by the statistical criterion. If no variable is removed at a given step then the variable that significantly adds the most discriminating power, as measured by the statistical criterion, is added to the discriminant function. The procedure stops at a step when addition or removal of variable from the discriminant function does not increase the R-squared. (Statgun, n.d)

Criteria for Variable Selection for Discriminant Function

Since discriminant analysis requires several assumptions for group membership, it is important to determine whether a specific variable is good enough and also important to be included in the discriminant function. In other words, it is important to assess whether the variable is a member of the discriminant groups.

There are a number of statistical criteria for determining the addition or the removal of variables from the discriminant function. The most common ones are:

1. Wilks' lambda (Λ)
2. Mahalanobis square distance
3. Rao's V

Wilks' Lambda (Λ)

Wilks' lambda (Λ) is the ratio of the within-group sum of squares to the total sum of squares. At each step, the variable that is included in the function is the one with the smallest Wilks' lambda (Λ) after the effect of variables already in the discriminant function is removed. Since the Wilks' lambda (Λ) can be approximated by the *F*-ratio, Wilks' lambda (Λ) is equal to entering the variable that has the highest partial *F*-ratio. Wilks' lambda (Λ) is thus given by

$$\Lambda = \frac{SS_w}{SS_t} = \frac{SS_w}{SS_b + SS_w}$$

Where

SS_w = sum of squares within groups

SS_t = total sum of squares

SS_b = sum of squares between groups

The assessment of the Wilks' lambda is done by converting to *F*- ratio with the transformation

$$F = \left(\frac{1-\Lambda}{\Lambda}\right)\left(\frac{n_1 + n_2 - p - 1}{p}\right),$$

where p is number of variables for which the statistic is computed and Λ is the Wilks' lambda of the distribution. *F*-ratio follows an *F*-distribution with $n_1 + n_2 - p - 1$ degrees of freedom.

Wilks' lambda tests the significance of each discriminant function in DA specifically, the significance of the eigenvalue for a given function. Minimizing Wilks' lambda is an indication that the within-group sum of squares is minimized and the between-group sum of squares is maximized. That is, the Wilks' lambda selection criterion considers between-groups separation and within-group homogeneity. The larger the lambda, the more likely it is significant. (Statgun, n.d)

A significant lambda means one can reject the null hypothesis that the two groups have the same mean discriminant function scores and conclude the model is discriminating. It is a measure of the difference between groups of the centroid (vector) of means on the independent variables. Wilks' Lambda varies from 0 to 1, with 0 meaning group means differ (thus the variable highly differentiates the groups), and 1 meaning all group means are the same.

The Bartlett's V transformation of lambda is used to compute the significance of lambda. Wilks' lambda is used, in conjunction with Bartlett's

V, as a multivariate significance test of mean differences in MDA, for the case of multiple interval independents and multiple. (Statgun, n.d)

Mahalanobis Distances

Mahalanobis distance (D^2) is another technique that is used in analyzing cases in discriminant analysis. For instance, one might wish to analyze a new and unknown set of cases in comparison to an existing set of known cases. Mahalanobis distance is the distance between a case and the centroid for each group (of the dependent) in attribute space (n-dimensional space defined by n variables). A case will have one Mahalanobis distance for each group, and it will be classified as belonging to the group for which its Mahalanobis distance is smallest. Thus, the smaller the Mahalanobis distance, the closer the case is to the group centroid and the more likely it is to be classified as a member of that group.

Mahalanobis distance for two group discriminant analysis is related to the squared multiple correlation coefficient (R^2) by

$$D^2 = \frac{(n_1 + n_2)(n_1 + n_2 - 2)R^2}{n_1 n_2 (1 - R^2)},$$

where n_1 is the number of cases in group one and n_2 is the number of cases in group two.

Rao's V

Rao's V is based on the Mahalanobis distance and concentrates on the separation between the groups, as measured by the distance of the centroid of each group from the centroid of the total sample. Rao's V is used to determine

the extent to which the discriminant functions discriminate between criteria groups. A measure from this group is mostly used in stepwise discriminant analysis to determine if adding an independent variable to the model will significantly improve classification of the dependent variable. Rao's V and the change in it while adding or deleting a variable is approximately χ^2 statistic and thus follows a χ^2 distribution.

Although Rao's V provides information about between-groups separation, it does not take into consideration group homogeneity. Therefore, the use of Rao's V may produce a discriminant function that does not have maximum within-group homogeneity. (Statgun, n.d)

The Discriminant Functions

The discriminant functions or models (also called canonical discriminant functions), are the set of equations that are used to find the relationship between the predictor variables and the response variable. They are built based on a set of observations for which the classes are known (training set). Based on the training set, the technique constructs a set of linear functions of the predictors or discriminant functions which are the heart of discriminant analysis. The discriminant functions are the linear combinations of the standardised independent variables which yield the biggest mean differences between the groups.

Number of Discriminant Functions for DA and MDA

The discriminant function is a linear function of the form

$$D_t = \lambda_{t_0} + \lambda_{t_1} x_1 + \lambda_{t_2} x_2 + \ldots + \lambda_{t_k} x_k$$

where,

D_t = the predicted discriminant score for group t .

t = the number of groups differentiated by the t discriminant functions

$\lambda_{t_0}, \lambda_{t_1}, \lambda_{t_2}, \ldots, \lambda_{t_k}$ are the weights of the independent variables $x_1, x_2, \cdots, x_k$ respectively and is the constant term in group *t*.

There is one discriminant function for 2-group discriminant analysis (i.e. if the dependent variable is a dichotomy), but for higher order DA (k dependent variables), up to *k*-1 discriminant functions can be extracted. Thus the maximum number of functions is the lesser of k - 1 (number of dependent groups minus 1).

A first function is computed on which the group means are as different as possible. A second function is then computed uncorrelated with the first, then a third function is computed uncorrelated with the first two, and so on, for as many functions as possible. Each discriminant function is orthogonal to the others. The first function maximizes the differences between the values of the dependent variable. The second function maximizes the differences between values of the dependent variable uncontrolled for by the first factor, the third function maximizes the differences between values of the dependents uncontrolled for by the first two, and so on. (Luna, n.d)

Though mathematically different, each discriminant function is a dimension which differentiates a case into categories of the dependent based on its values on the independents. The first function will be the most powerful differentiating dimension, but later functions may also represent additional significant dimensions of differentiation. (Luna, n.d)

Calculation of the Discriminant Function for Two-Grouped DA

Considering two populations with observed values on p random variables $X_1, X_2, \ldots, X_p$ for each of the n_1 individuals selected from population 1 and for each of the n_2 individual from population 2. In particular, for the i^{th} population $(i = 1, 2)$, supposing that we let X_{ijk} denote the observed value of variable $X_j (j = 1, 2, \ldots, p)$ for the k^{th} sampled individual $(k = 1, 2, \ldots, n_i)$. Thus, the set of variable values $(x_{i1k}, x_{i2k}, \ldots, x_{1pk})$ represent the group of measurements obtained for the k^{th} individual selected from population *i*. The main objective of discriminant analysis is to develop a model *L* that is a linear combination of the independent variables say

$$L = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

with values for $\beta_1, \beta_2, \ldots, \beta_p$ chosen so as to provide maximum discrimination between the two populations (thus, the variation in the values of *L* between the two groups would be much greater that the variation in the values of *L* within the two groups).

For any k^{th} individual from population i, if the $\beta's$ are known, the associated *L* value would be

$$L_{ik} = \beta_1 x_{i1k} + \beta_2 x_{i2k} + \cdots + \beta_p x_{ipk}.$$

In the analysis-of-variance framework, the total variation in the scores is measured by

$$\sum_{i=1}^{2} \sum_{k=1}^{n_i} (L_{ik} - \bar{L})^2$$

where

$$\bar{L} = \frac{1}{n_1 + n_2} \sum_{i=1}^{2} \sum_{k=1}^{n_i} L_{ik} = \frac{1}{n_1 + n_2} (n_1 \bar{L}_1 + n_2 \bar{L}_2)$$

The total sum of squares can be broken down into two interpretable components, a between-groups sum of squares, *B*, given by

$$B = \sum_{i=1}^{2} n_i (\bar{L}_i - \bar{L})^2 = \frac{n_1 n_3}{n_1 + n_2} (\bar{L}_1 - \bar{L}_2)^2$$

and a within-group sum of squares, *W*, given by

$$W = \sum_{i-1}^{2} \sum_{k=1}^{n_i} (L_{ik} - \bar{L}_i)^2 .$$

The ratio $\frac{\mathbf{B}}{\mathrm{W}}$ is a measure of the discriminant power of *L* due to the fact that the larger the value of *B* relative to *W*, the more *L* is reflecting between-population variation as opposed to within–population variation.

Also, let

$$\bar{x}_{ij} = \sum\nolimits_{k=1}^{n_j} x_{ijk} / n_j$$

be the observed mean value of variable j in the two sample of n_j individuals from population j and let also $d_j = \bar{x}_{1j} - \bar{x}_{2j}$

be the observed differences between values of variable, then if

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} \ldots . s_{1p} \\ s_{21} & s_{22} \ldots s_{2p} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ s_{p1} & s_{p2} \ldots & s_{pp} \end{bmatrix}$$

is the covariance matrix of the $x's$ so that

$$\mathbf{S}^{-1} = \begin{bmatrix} s^{11} & s^{12} & \dots & s^{1p} \\ s^{21} & s^{22} & \dots & s^{2p} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ s^{p1} & s^{p2} & \dots & s^{pp} \end{bmatrix}$$

is the inverse matrix of **S**, then the values of $b_1, b_2, \ldots, b_p$ which maximizes $\frac{\mathbf{B}}{\mathrm{W}}$ are given as follows:

$$b_1 = s^{11}d_1 + s^{12}d_2 + \cdots + s^{1p}d_p$$

$$b_2 = s^{21}d_1 + s^{22}d_2 + \cdots + s^{2p}d_p$$

$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$

$$\cdot$$

$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$

$$b_p = s^{p1}d_1 + s^{p2}d_2 + \cdots + s^{pp}d_p$$

The linear combination of the $x's$ based on the $b's$ is given by

$$l = b_1x_1 + b_2x_2 + \cdots + b_px_p$$

and this linear combination maximizes the quantity $\frac{\mathbf{B}}{\mathrm{W}}$ based on the sample at hand. The $b's$ are estimates of the $\beta's$ and l is also an estimate of the optimal linear combination of *L*. (Bartlett, 1951)

Testing for Importance of Discriminant Functions

Assessing Relative Importance

The eigenvalue, which is the characteristic root of each discriminant function, reflects the ratio of importance of the dimensions which classify cases of the dependent variable. There is one eigenvalue for each discriminant

function. For two-group DA, there is one discriminant function and one eigenvalue, which accounts for 100% of the explained variance. If there is more than one discriminant function, the first will be the largest and most important, the second next most important in explanatory power, and so on.

The eigenvalues assess relative importance because they reflect the percents of variance explained in the dependent variable, cumulating to 100% for all functions. That is, the ratio of the eigenvalues indicates the relative discriminating power of the discriminant functions. If the ratio of two eigenvalues is 1.4, for instance, then the first discriminant function accounts for 40% more between-group variance in the dependent categories than does the second discriminant function. (Luna, n.d)

Relative Percentage of Discriminant Functions

The relative percentage (RP) tells how many functions are important. The relative percentage of a discriminant function equals a function's eigenvalue divided by the sum of all eigenvalues of all discriminant functions in the model. Thus it is the percent of discriminating power for the model associated with a given discriminant function. The relative percentage is given by

$$\text{RP} = \frac{\lambda_i}{\sum_{i=1}^{k} \lambda_i},$$

where λ_i = the eigenvalue of the i^{th} discriminant function and k is the number of discriminant functions. (Luna, n.d)

Testing for Association between Groups and Discriminant Function

The association between the groups formed by the dependent and the given discriminant function is measured by the canonical correlation **(**R*). The canonical correlation of each discriminant function is also the correlation of that function with the discriminant scores .When R* is zero, it means there is no relation between the groups and the function. When the canonical correlation is large, there is a high correlation between the discriminant functions and the groups. R* is used to tell how much each function is useful in determining group differences. A canonical correlation close to 1 means that nearly all the variance in the discriminant scores can be attributed to group differences. Squared canonical correlation, $(R^*)^2$, is the percentage of variation in the dependent discriminated by the set of independents in DA or MDA. For two-group DA, the canonical correlation is equivalent to the Pearsonian correlation of the discriminant scores with the grouping variable.(Wikipedia, n.d).

Assessing Independent Variables in the Discriminant Function

The discriminant score, also called the DA score, is the value resulting from applying a discriminant function formula to the data for a given case. If the discriminant score of the function is less than or equal to cutoff (the mean of the centroids of two groups), the case is classed as 0, or if above it is classed as 1. (Wikipedia, n.d).

Discriminant Coefficients

There are two types of discriminant coefficients:

1. unstandardized discriminant coefficients.
2. standardized discriminant coefficients

Unstandardized Discriminant Coefficients

Unstandardized discriminant coefficients are used in the formula for making the classifications in DA much as *b* coefficients are used in regression in making predictions. The constant plus the sum of products of the unstandardized coefficients with the observations yields the discriminant scores. That is, discriminant coefficients are the regression-like *b* coefficients in the discriminant function, in the form $L = b_0 + b_1x_1 + b_2x_2 + ... + b_nx_n$ where *L* is the latent variable formed by the discriminant function, the *b*'s are discriminant coefficients, the *x*'s are discriminating variables, and b_0 is a constant. The discriminant function coefficients are partial coefficients, reflecting the unique contribution of each variable to the classification of the criterion variable.

Standardized Discriminant Coefficients

Standardized discriminant coefficients, also termed the standardized canonical discriminant function coefficients, are used to compare the relative importance of the independent variables, much as beta weights (*b*) are used in regression. The standardized discriminant coefficients are also used to assess the relative classifying importance of the independent variables. Importance of variables is assessed relative to the model being analyzed. Addition or deletion of variables in the model can change discriminant coefficients markedly. In situations where there are more than two groups of the dependent, the

standardized discriminant coefficients do not give any information about which groups the variable is most or least discriminating. For this reason, group centroids and factor structure are examined.

Testing Efficiency of Discriminant Functions

The mean discriminant scores for each of the dependent variable categories for each of the discriminant functions in discriminant analysis are know as functions at group centroids. Two-group discriminant analysis has two centroids, one for each group. If the means of the two groups are well apart, it shows that the discriminant function is clearly discriminating. The closer the means, the more errors of classification there likely will be.

Discriminant function plots, also called canonical plots, can be created in which the two axes are two of the discriminant functions (the dimensional meaning of which is determined by looking at the structure coefficients), and circles within the plot locate the centroids of each category being analyzed. The farther apart one point is from another on the plot, the more the dimension represented by that axis differentiates those two groups. Thus, these plots depict discriminant function space.

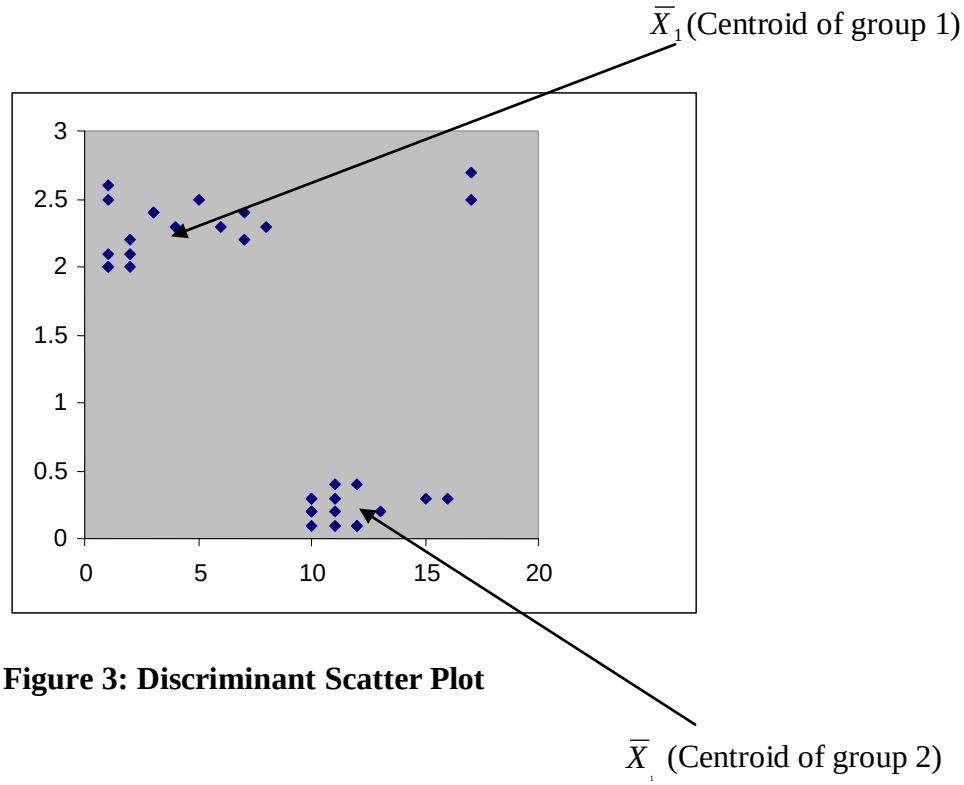

Figure 3: Discriminant Scatter Plot

Wilks' Lampda Significance Tests

Models Wilks' Lambda

Model Wilks' lambda is used to test the significance of the discriminant function as a whole. The larger the lambda, the more likely it is significant. A significant lambda means one can reject the null hypothesis that the two groups have the same mean discriminant function scores and conclude the model is discriminating.

Model Wilks' Lambda Difference Tests

This is also used in a second context to assess the improvement in classification when using sequential discriminant analysis. There is an *F*- test of significance of the ratio of two Wilks' lambdas, such as between a first one for a set of control variables as predictors and a second one for a model

including both control variables and independent variables of interest. The second lambda is divided by the first (where the first is the model with fewer predictors). (Statgun, n.d)

Purposes of Discriminant Analysis

This concern for the classification ability of the linear discriminant function or model is even confused due to the fact that two very distinct purposes for conducting discriminant analysis exist. These are:

1. discriminant analysis for predictive purposes (discriminant prediction analysis).
2. discriminant analysis for classification purposes (discriminant classification analysis).

Discriminant Analysis for Prediction

Discriminant analysis for prediction is used to optimize the predictive function. The discriminant analysis for predictive purposes maximizes the amount of subject variance explained by the linear function. It uses a set of *k* variables with associated weights (λ_{t_k}) that are derived in a best fit, linear unbiased fashion to predict the score of the dependent variable, *D*. These discriminant scores are predictors of group membership that can be used to classify groups of observations that are of either known or unknown group membership. (Luna, n.d)

Discriminant analysis conducted for predictive purposes formulates a linear discriminant function describing the importance of the independent variables in differentiating observations of known group membership. Given

two previously identified groups, predictive discriminant analysis formulate predictive function from the independent variables to explain differences between the members of the two groups. Example, a predictive analysis was used to differentiate a sample of purchasers of a certain product and non-purchasers of the products, or innovators and non-innovators. In these situations, group membership was known prior to the analysis and the sole purpose was to derive the predictive function, using the set of independent variables, to predict consumers of unknown group membership or innovators and non-innovators. A predictive analysis is possible in many situations where prior designation of groups exists. (Statgun, n.d)

Discriminant Analysis for Classification of Observations

Discriminant classification analysis uses the predictive functions derived in the predictive analysis to classify fresh sets of data of known group membership, thereby validating the predictive function; or if the function has previously been validated, to classify new sets of observations of unknown group membership. The purpose of classification of observations of known grouping is merely to see how well the derived function predicts group membership using the subject data from which it was derived. The classification procedure associated with the predictive analysis may be thought of as a base line analysis that establishes a standard of comparison for future discriminant classification analysis. (Luna, n.d).

Classification of Two Populations

There are several methods of actually classifying cases in MDA. The aims or objectives of all these methods are to classify cases with a minimum error. A good classification procedure should result in few misclassifications. In other words, the chances, or probabilities, of misclassification should be small.

One aspect of classification is cost. Suppose that classifying a case from population one (π_1) as belonging to population two (π_2) represents a more serious error than classifying a π_2 object as belonging to a π_1, then one should be cautious about making the former assignment. An optimal classification procedure should, whenever possible, account for the costs associated with misclassification. (Luna, n.d)

Cost of Misclassification

Consider probability density functions $f_1(x)$ and $f_2(x)$ associated with random variable **x** for the population π_1 and π_2, respectively. If an object with an associated measurement **x** has to be assigned to either π_1 or π_2, then consider Ω to be the sample space; that is, collection of all possible observations **x**. Let R_1 be the set of **x** for which an object is classified as π_1 and $R_2 = \Omega - R_1$ be the remaining **x** values for which an object is classified as π_2. Since every object must be assigned to one and only one of the two populations, the sets R_1 and R_2 are mutually exclusive and exhaustive.

The conditional probability, $P(2 \mid 1)$, of classifying an object as π_2 when, in fact, it is from π_1 is

$$P(2 \mid 1) = P(\mathbf{x} \in R_2 \mid \pi_1) = \int_{R_2=\Omega-R_1} f_1(\mathbf{x})\mathrm{d}\mathbf{x} \qquad (12)$$

Similarly, the conditional probability, $P(1 \mid 2)$, of classifying an object as π_1 when it is really from π_2 is

$$P(1 \mid 2) = P(\mathbf{x} \in R_1 \mid \pi_2) = \int_{R_1} f_2(\mathbf{x})\mathrm{d}\mathbf{x} \qquad (13)$$

Equation (12) represents the volume formed by the density function $f_1(\mathbf{x})$ over the region R_2. Similarly, equation (13) represents the volume formed by $f_2(\mathbf{x})$ over the region R_1. If p_1 is prior probability of π_1 and p_2 is the prior probability of π_2, where $p_1 + p_2 = 1$. The overall probabilities of incorrectly classifying an object is given by

$$P(2 \mid 1) + P(1 \mid 2) = \int_{R_2=\Omega-R_1} f_1(\mathbf{x})\mathrm{d}\mathbf{x} + \int_{R_1} f_2(\mathbf{x})\mathrm{d}\mathbf{x}$$

The costs are:

(1) zero for correct classification,

(2) $c(1 \mid 2)$ when an observation from π_2 is incorrectly classified as π_1,

(3) $c(2 \mid 1)$ when π_1 observation is incorrectly classified as π_2.

For any rule, the average, or expected cost of misclassification (ECM) is provided by the product of the off-entries by their probabilities of occurrence. Hence,

$$\text{ECM} = c(2 \mid 1)\, P(2 \mid 1)\, p_1 + c(1 \mid 2)\, P(1 \mid 2)\, p_2 .$$

A good classification rule should have an ECM as small as possible. The regions R_1 and R_2 that minimize the ECM are defined by the values $\mathbf{x}$ for which the inequalities

$$R_1 : \frac{f_1(x)}{f_2(x)} \geq \left(\frac{c(1 \mid 2)}{c(2 \mid 1)} \right) \left(\frac{p_2}{p_1} \right) \qquad (14)$$

i.e

$$\begin{pmatrix}\text{density}\\ \text{ratio}\end{pmatrix} \geq \begin{pmatrix}\text{cost}\\ \text{ratio}\end{pmatrix}\begin{pmatrix}\text{prior}\\ \text{probability}\\ \text{ratio}\end{pmatrix}$$

and

$$R_2 : \frac{f_1(\mathbf{X})}{f_2(x)} < \left(\frac{c(1\mid 2)}{c(2\mid 1)}\right)\left(\frac{p_2}{p_1}\right) \tag{15}$$

$$\begin{pmatrix}\text{density}\\ \text{ratio}\end{pmatrix} < \begin{pmatrix}\text{cost}\\ \text{ratio}\end{pmatrix}\begin{pmatrix}\text{prior}\\ \text{probability}\\ \text{ratio}\end{pmatrix}$$

Minizing the Total Probability of Misclassification (TPM)

TMP = P(misclassifying a π_1 observation or misclassifying a π_2 observation)

= P(observation comes from π_1 and is misclassified as from π_2)

+ P(observation comes from π_2 and is misclassified as from π_1)

$= p_1 \int_{R2} f_1 d\mathbf{x} + p_2 \int_{R_1} f_2 d\mathbf{x}$

This problem is equivalent to minimizing the expected cost of misclassification when the costs of misclassification are equal.

A new observation $\mathbf{x_o}$ could be allocated to the population with the largest "posterior" probability. Thus classifying an observation in population π_1 is given as below:

$$P(\pi_1 \mid \mathbf{x_o}) = \frac{P(\pi_1 \text{ occurs and observe } x_o)}{P(\text{observe } x_o)}$$

$$= \frac{P(\text{observe } x_o \mid \pi_1)P(\pi_1)}{P(\text{observe } x_o \mid \pi_1)P(\pi_1) + P(\text{observe } x_o \mid \pi_2)P(\pi_2)} \tag{16}$$

$$= \frac{P_1 f_1(x_o)}{P_1 f_1(x_o) + P_2 f_2(x_o)} \quad (17)$$

Also classifying an observation $\mathbf{x_0}$ in population π_2 is given as below:

$$P(\pi_2 \mid x_o) = 1 - P(\pi_1 \mid \mathbf{x_0}) = \frac{P_2 f_2(x_o)}{P_1 f_1(x_o) + P_2 f_2(x_o)} \quad (18)$$

Classification with Two Multivariate Normal Populations

Classification procedures based on normal populations is a major practice in statistics. Assuming $f_1(x)$ and $f_2(x)$ are multivariate normal densities with mean vectors μ_1 and μ_2, and covariance matrix $\sum_1$ and $\sum_2$ respectively.

Classification of Normal Population when $\sum_1 = \sum_2 = \sum$

Assume the joint densities of $X^1 = [X_1, X_2 \ldots X_p]$ for populations π_1 and π_2 are given by

$$f_i(x) = \frac{1}{(2\pi)^{p/2} \left|\sum\right|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_i)' \sum^{-1} (x-\mu_i)\right], \quad \text{for } i = 1, 2 \quad (19)$$

If the population parameters μ_1, μ_2 and $\sum$ are known, then after cancellation of $(2\pi)^{p/2} \left|\sum\right|^{1/2}$ (since it is common to both populations), the minimum expected cost of misclassification (ECM) are

$R_1 =$

$$\exp\left[-\frac{1}{2}(x-\mu_i)' \sum^{-1} (x-\mu_i) + \frac{1}{2}(x-\mu_2)' \sum^{-1} (x-\mu_2)\right] \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \quad (20)$$

$$R_2 =$$

$$\exp\left[-\frac{1}{2}(x-\mu_i)'\sum{}^{-1}(x-\mu_i)+\frac{1}{2}(x-\mu_2)'\sum{}^{-1}(x-\mu_2)\right] < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \quad (21)$$

Taking natural logarithm of (1) gives

$$R_1 = \left[-\frac{1}{2}(x-\mu_1)'\sum{}^{-1}(x-\mu_1)+\frac{1}{2}(x-\mu_2)'\sum{}^{-1}(x-\mu_2)\right] \geq \ln\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$\Rightarrow (\mu_1-\mu_2)'\sum{}^{-1}x-\frac{1}{2}(\mu_1-\mu_2)^1\sum{}^{-1}(\mu_1+\mu_2) \geq \ln\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

Hence the allocation rule is given as below:

Allocate x_o to π_1 if

$$(\mu_1-\mu_2)'\sum{}^{-1}x-\frac{1}{2}(\mu_1-\mu_2)^1\sum{}^{-1}(\mu_1+\mu_2) \geq \ln\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \quad (22)$$

or

Allocate x_o to π_2 if

$$(\mu_1-\mu_2)'\sum{}^{-1}x-\frac{1}{2}(\mu_1-\mu_2)'\sum{}^{-1}(\mu_1+\mu_2) < \ln\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \quad (23)$$

In most practical situations, the population quantity μ_1, μ_2 and $\sum$ are not known, so, Wald A. and Anderson T.W. suggested that the sample counterparts of the parameters be used instead.

Since it is assumed that the parent populations have the same covariance matrix $\sum$, the sample covariance matrices S_1 and S_2 are combined (pooled) to derive a single unbiased estimate of $\sum$. The weighted average, S_{pooled} is given by

$$S_{pooled} = \left[\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}\right] S_1 + \left[\frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)}\right] S_2$$

$$= \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{(n_1 + n_2 - 2)}$$

is an unbiased estimate of $\sum$ and the data matrices X_1 and X_2 contain random sample from the population π_1 and π_2, respectively.

Substituting $\bar{X}_1$ for μ_1, $\bar{X}_2$ for μ_2 and S_{pooled} for $\sum$ in (11) gives the classification rule below:

Allocate X_0 to π_1 if

$$(\bar{X}_1 - \bar{X}_2)^1 S_{pooled}^{-1} X_0 - \frac{1}{2}(\bar{X}_1 - \bar{X}_2)^1 \sum^{-1} (\bar{X}_1 + \bar{X}_2) \geq \ln\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \quad (24)$$

Allocate X_0 to π_2 otherwise.

Classification of Normal Population when $\sum_1 \neq \sum_2$

From equations (19) and (20) above, substituting multivariate normal densities with different covariance matrices $\sum_1$ and $\sum_2$ in (1)**,** we have

$$f_1(x) = \frac{1}{(2\pi)^{p/2}\left|\sum_1\right|^{1/2}} \exp\left[-\frac{1}{2}(x - \mu_1)' \sum_1^{-1} (x - \mu_1)\right] \quad (25)$$

$$f_2(x) = \frac{1}{(2\pi)^{p/2}\left|\sum_2\right|^{1/2}} \exp\left[-\frac{1}{2}(x - \mu_2)' \sum_2^{-1} (x - \mu_2)\right] \quad (26)$$

Hence from (14),

$$\frac{f_1(x)}{f_2(x)} = \frac{\frac{1}{(2\pi)^{\frac{p}{2}}\left|\sum_1\right|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_1)^1 \sum_1^{-1}(x-\mu_1)\right]}{\frac{1}{(2\pi)^{\frac{p}{2}}\left|\sum_2\right|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_2)^1 \sum_2^{-1}(x-\mu_2)\right]} \qquad (27)$$

Taking natural logarithm of (20) and simplifying the results yield the below classification regions:

$$R_1: \quad -\frac{1}{2}X^1(\sum_1^{-1}-\sum_2^{-1})X + (\mu_1^1\sum_1^{-1}-\mu_2^1\sum_2^{-1})X - k \geq \ln\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$R_1: \quad -\frac{1}{2}X^1(\sum_1^{-1}-\sum_2^{-1})X + (\mu_1^1\sum_1^{-1}-\mu_2^1\sum_2^{-1})X - k < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

Where

$$k = \frac{1}{2}\ln\left(\frac{\left|\sum_1\right|}{\left|\sum_2\right|}\right) + \frac{1}{2}(\mu_1^1\sum_1^{-1}\mu_1 - \mu_2^1\sum_2^{-1}\mu_2)$$

The allocation rule that minimizes the expected cost of misclassification is given by

Allocate $\mathbf{x_0}$ to π_1 if

$$-\frac{1}{2}X_0{}^1(\sum_1^{-1}-\sum_2^{-1})X_0 + (\mu_1^1\sum_1^{-1}-\mu_2^1\sum_2^{-1})X_0 - k \geq \ln\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

Allocate $\mathbf{x_0}$ to π_2 otherwise.

From Luna, according to Wald A. and Anderson T.W for practical conditions, the classification rule is implemented by substituting the sample

quantities $\bar{X}_1, \bar{X}_2, S_1,$ and S_2 for $\mu_1, \mu_2, \sum_1,$ and $\sum_2,$ respectively. Hence the allocation rule is given as below:

Allocate $\mathbf{x_0}$ to π_1 if

$$-\frac{1}{2} X_0^{\ 1}(S_1^{-1} - S_2^{-1})X_0 + (\bar{X}_1^1 S_1^{-1} - \bar{X}_2^1 S_2^{-1})X_0 - k \geq \ln\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

Allocate $\mathbf{x_0}$ to π_2 otherwise.

Fisher's Discriminant Function - Separation of Two Populations

Fisher's idea of linear classification was to transform the multivariate observation **x** to univariate observation *y* such that the *y*'s derived from population π_1 and π_2 were separated as much as possible. Fisher suggested taking linear combinations of *x* to create *y*'s because they are simple enough functions of the *x* to be handled easily. Fisher's approach does not assume that the populations are normal. It does, however, implicitly assume that population covariace matrix is used. (Luna, n.d)

A fixed linear combination of the *x*'s takes the values $y_{11}, y_{12}, \ldots, y_{1n}$ for the observations from the first population and the values $y_{21}, y_{22,} \ldots, y_{2n}$ for the observations from the second population. The separation of these two sets of univariate *y*'s is assessed in terms of the difference between $\bar{y}_1$ and $\bar{y}_2$ expressed in standard derivation units. That is,

$$\text{Separation} = \frac{|\bar{y}_1 - \bar{y}_2|}{s_y},$$

where

$$s_y^2 = \frac{\sum_{j=1}^{n_1}(y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_2)^2}{n_1 + n_2 - 2}$$

is the pooled estimate of the variance. The objective is to select the linear combination of the *x* to achieve maximum separation of the sample means $\bar{y}_1$ and $\bar{y}_2$.

The linear combination $y = \hat{\ell}'x = (\bar{x}_1 - \bar{x}_2)S_{pooled}^{-1}x$ maximizes the ratio

$$\frac{\begin{pmatrix}\text{Square distance}\\ \text{between sample means of y}\end{pmatrix}}{(\text{sample variance of y})} = \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2}$$

$$= \frac{(\hat{\ell}'\bar{x}_1 - \hat{\ell}\bar{x}_2)^2}{\hat{\ell}'S_{pooled}\hat{\ell}} \qquad (28)$$

$$= \frac{(\hat{\ell}'d)^2}{\hat{\ell}'S_{pooled}\hat{\ell}}$$

The over all possible coefficient vectors is $\hat{\ell}$ and $d = (\bar{x}_1 - \bar{x}_2)$. The maximum of the ratio (1) is $D^2 = (\bar{x}_1 - \bar{x}_2)'S_{pooled}^{-1}(\bar{x}_1 - x_2)$, where D^2 is the sample squared distance between the two means.

Allocation Rule (Classification) using Fisher's Discriminant Function

Allocate $\mathbf{x_0}$ to π_1 if

$$y_0 \geq \hat{m},$$

where

$$y_0 = (\bar{x}_1 - \bar{x}_2)'S_{pooled}^{-1}x_0 \quad \text{and} \quad \hat{m} = \tfrac{1}{2}(\bar{x}_1 - \bar{x}_2)'S_{pooled}^{-1}(\bar{x}_1 + \bar{x}_2)$$

or

$$y_0 - \hat{m} \geq 0$$

Allocate $\mathbf{x_0}$ to π_2 if

$$y_0 < \hat{m} \quad \text{or} \quad y_0 - \hat{m} < 0$$

This Fisher's linear discriminant function was developed under the assumption that the two populations, whatever their form, have a common covariance matrix *S*.

Thus

$$w = (\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} x - \tfrac{1}{2}(\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} (\bar{x}_1 + \bar{x}_2)$$

$= (\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} - [\, x - \tfrac{1}{2}(\bar{x}_1 + \bar{x}_2)\,]$ which is frequently called Anderson's classification function. (Luna, n.d)

In summary, Fisher's discriminant function says, for two populations, the maximum relative separation that can be obtained by considering linear combinations of the multivariate observations is equal to the distance D^2. The D^2 can be used to test whether the population means μ_1 and μ_2 differ significantly. Consequently, a test for differences in mean vectors can be viewed as a test for the "significance" of the separation that can be achieved.

Supposing the populations π_1 and π_2 are multivariate normal with a common covariance matrix $\sum$. Then, a test of $H_0 : \mu_1 = \mu_2$ versus $H_0 : \mu_1 \neq \mu_2$ with a corresponding test statistic

$$\left(\frac{n_1 + n_1 - p - 1}{(n_1 + n_2 - 2)p}\right)\left(\frac{n_1 n_2}{n_1 + n_2}\right) D^2$$

to an *F*-distribution with $\upsilon_1 = p$ and $\upsilon_2 = n_1 + n_2 - p - 1$ degrees of freedom. If H_0 is rejected, we can conclude that the separation between the two populations π_1 and π_2 is significant. (Luna, n.d)

Differences between Discriminant Predictive and Discriminant Classification Analysis

The discriminant classification analysis is in sharp contrast to the discriminant predictive analysis.

Discriminant analysis conducted for predictive purposes formulates a linear discriminant function describing the importance of the independent variables in differentiating observations of known group membership. Discriminant analysis conducted for classification purposes validates the predictive discriminant function as a means of classifying fresh observations of unknown group membership sampled from the same population. In the event of previous validation of the predictive function, the classification analysis is purely for classification purposes. The result of the classification analysis is evaluated in the light of the specific objectives, if optimization rather than maximization is to result.

Discriminant analysis conducted for predictive purposes uses an initial data set having known group membership to both derive the discriminant functions and predict group classification. This classification of observation is an extension of the predictive discriminant analysis in that the predictive Discriminant scores, D_{it}, form the basis of the decision rule used to classify this same set of objects into the *t* groups. In contrast to the classification of the

initial data set, where group membership is known, the same decision rule may be applied to other sets of data.

However, when one classifies data sets other than the initial set from which the predictive analysis was conducted, the person is no longer engaged in predictive discriminant analysis, but rather in discriminant classification analysis. Predictive discriminant analysis requires no validation procedures to be implemented, since derivation of an optimal discriminant function is the only relevant issue. However, if fresh sets of data with either known or unknown grouping are classified, then the discriminant function must be validated to be generalisable to these data sets centroids. (Wikipedia, n.d)

Validation of Discriminant Function

Generalised distance functions are based on the Mahalanobis distance, D-square, of each case to each of the group but other methods such as the holdout method can also be used for validation. The holdout sample method is a split halves test, where a portion of the cases are assigned to the analysis sample for purposes of training the discriminant function, then it is validated by assessing its performance on the remaining cases in the hold-out sample.

Another method that is also used for validating the discriminant function is the U-method. The U-method was proposed by Lachenbruch in 1967. This method holds one observation from n samples at a time, estimates the discriminant function using the remaining $n-1$ observations, and classifies the held-out observation. (Wikipedia, n.d).

Testing Importance of Independent Variables

The derived discriminant coefficients may be interpreted as indicative of the importance of the respective *p* independent variables entered into the discriminant analysis. Although these coefficients indicate importance, they are not appropriate for assessing the relative importance or discriminatory power of the variables, (i.e., the proportion of total discriminating power attributable to a specific variable). Relative importance of the independent variables entered in the predictive function is defined in part by:

$$\text{RP} = \frac{I_p}{\sum_{i=1}^{k} I_i},$$

$$I_p = \left|\Lambda_p(\bar{x}_{pt} - \bar{x}_{p_i})\right|$$

Where

I_p = the importance of the p^{th} variable

Λ_p = the unstandardized discriminant coefficient for the p^{th} variable;

$\bar{x}_{pt}$ = the mean of the p^{th} variable for the t^{th} group.

The bigger the value of the relative importance of a variable is, the more important that variable is in the function. (Wikipedia, n.d).

Requirements and Assumptions for Discriminant Analysis

Proper Specification

The discriminant coefficients can change substantially if variables are added to or subtracted from the model. It is therefore important to specify the variables which are necessary to give a reliable equation.

True Categorical Dependent Variable

The dependent variable is a true dichotomy. When the range of true underlying continuous variable is constrained to form a dichotomy, correlation is attenuated (biased toward underestimation). It is therefore important not to dichotomize a continuous variable simply for the purpose of applying discriminant function analysis, (the same considerations apply to trichotomies and higher). All cases must belong to a group formed by the dependent variable. The groups must be mutually exclusive, with every case belonging to only one group.

Homogeneity of Variances (Homoscedasticity)

Within each group formed by the dependent, the variance of each interval independent should be similar between groups. That is, the independents may (and will) have different variances one from another, but for the same independent, the groups formed by the dependent should have similar variances and means on that independent.

Homogeneity of Covariances/Correlations

Within each group formed by the dependent, the covariance/correlation between any two predictor variables should be similar to the corresponding covariance/correlation in other groups. DA will tend to classify cases in the group with the larger variability.

Assumption of Linearity

Discriminant analysis does not take into account exponential terms unless such transformed variables are added as additional independents.

Low Multicollinearity of the Independent Variables

To the extent that independent variables are correlated, the standardized discriminant function coefficients will not reliably assess the relative importance of the predictor variables. Multicollinearity is looking at the "pooled within-groups correlation.”Pooled" is the average across groups formed by the dependent but this can be very different from normal (total) correlation when two variables are less correlated within groups than between groups.

Assumption of Additivity

Discriminant analysis does not take into account interaction terms unless new cross-product variable are added as additional independents.

For Purposes of Significance Testing

Predictor variables follow multivariate normal distributions. That is, each predictor variable has a normal distribution about fixed values of all the other independents. When non-normality is caused by outliers rather than skewness, violation of this assumption of normality would have more serious consequences as discriminant analysis is highly sensitive to outliers.

Sample Size

Unequal sample sizes are acceptable. The sample size of the smallest group needs to exceed the number of predictor variables. As a “rule of thumb”, the smallest sample size should be at least 20 for a few (4 or 5) predictors. The maximum number of independent variables is n - 2, where n is the sample size. While this low sample size may work, it is not encouraged, and generally it is best to have 4 or 5 times as many observations as the independent variables.

Assumption of Normality of Distribution

It is assumed that the data (for the variables) represent a sample from a multivariate normal distribution. One can examine whether or not variables are normally distributed with histograms of frequency distributions. However, violations of the normality assumption are not "fatal" and the resultant significance tests are still reliable as long as non-normality is caused by skewness and not outliers (Tabachnick, et al., 2001).

Outliers

Discriminant analysis is highly sensitive to outliers. Lack of homogeneity of variances may indicate the presence of outliers in one or more groups and it will also mean significance tests are unreliable. A test for univariate and multivariate outliers for each group has to be carried out and transform or eliminate them.

If one group in the study contains extreme outliers that impact the mean, they will also increase variability. Overall significance tests are based on pooled variances, that is, the average variance across all groups. Thus, the significance

tests of the relatively large means (with the large variances) would be based on the relatively smaller pooled variances, resulting erroneously in statistical significance.

CHAPTER FIVE

COMPARISON OF LOGISTIC REGRESSION AND DISCRIMINANT ANALYSIS

Introduction

Logistic regression and discriminant analysis share a lot of common characteristics but sometimes do have some differences. It is, therefore, important to make comparisons of the two techniques to determine the areas of their similarities and also where they have differences.

Theoretical Differences between Logistic Regression and Discriminant Analysis

Logistic regression analysis may be considered as an alternative technique to discriminant analysis in some cases. However, there are some differences between the two techniques. While the Logistic regression analysis bears the same logic as ordinary least square regression, discriminant analysis, on the other, works in line with multiple analysis of variance (MANOVA).

Difference in Purpose

Instead of classifying an observation into one group or the other, as in the case of discriminant analysis, logistic regression predicts the probability

that an object is a member of one of the groups. Thus, it predicts the probability (π) that an object belongs to one group rather than the other.

Differences in Requirements

The discriminant analysis situation has been a more integral part of the historical development of multivariate statistics when the dependent variables come from a normal population and the homoscedastic for each level of the independents. Also, discriminant analysis requires that the independent variables come from a normal population but the logistic regression does not strictly condition the independent variables to be normally distributed.

Also, discriminant analysis requires homoscedasticity for each level of the independent variables or makes requirement of equal group membership but logistic regression analysis does not make any of such requirements.

Difference in Computation of Predictor Coefficients

Discriminant analysis tries to find coefficients for the independent variables that minimize the within group variability and maximize between group variability. Logistic regression uses maximum-likelihood estimation to compute the coefficients for the logistic regression equation. This method attempts to find coefficients that match the breakdown of cases on the dependent variable. But the coefficients of the independent variables for discriminant analysis are computed using calculations.

Difference in Model

The dependent variable, for discriminant analysis, is linearly related to the independent variables but the response variable for logistic regression is not linearly related to the independent variable.

Thus, the discriminant analysis has the linear equation of the form

$$D_t = \lambda_{t_0} + \lambda_{t_1} x_1 + \lambda_{t_2} x_2 + \cdots + \lambda_{t_k} x_k$$

and the logistic regression has the equation of the form

$$\pi_i = \frac{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}}{1+e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}}$$

Differences in Assumptions

Logistic regression usually involves fewer violations of assumptions (independent variables need not be normally distributed, linearly related to the dependent variable, or have equal within-group variances) as compared to discriminant analysis.

Theoretical Similarities between Logistic Regression and Discriminant Analysis

Both logistic regression analysis and discriminant analysis are techniques for separating groups with similar characteristics. Also, both techniques have a similar purpose of predicting group membership and also classifying new cases into one of predefined groups. Again, all the two techniques aim at coming out with a model that can best be used to describe relationship between dependent and independent variables. Logit function of logistic regression and discriminant function attempt to express dependent

variable as a linear combination of other features or measurements. Also the two techniques estimate coefficients of independent variables that best explain the data.

Table2: Comparison of the Errors and Different Types of the Two Methods

---

Logistic Regression

Types of the technique

1. Binary logistic regression: - is used when dependent variable involves two categories.
2. Multinomial logistic regression: - is used when dependent variable involves three or more categories.

Assumptions and requirements

1. Does not make assumptions such as normality, linearity, equal variance of independents.
2. Has disadvantage of requirement of large data set to attain stability.

Discriminant Analysis

Types of the technique

1. Linear discriminant analysis :- is used when the dependent variable involves two categories and resembles ANOVA.
2. Multiple discriminant analysis:- is used when the dependent variable involves three or more categories.

Table 2 (*Continued*)

Requirements and assumptions
1. Requires that independent variables have to be linearly related to the dependent variable and also assume equal variance for groups formed by dependents.
2. Does not require large data set to attain stability as compared to logistic regression.
3. Requires that independent variables have to be linearly related to the dependent variable and also assume equal variance for groups formed by dependents.
4. Does not require large data set to attain stability as compared to logistic regression.
5. Requires that independent variables have to be linearly related to the dependent variable and also assume equal variance for groups formed by dependents.

Advantages of Logistic Regression Analysis over Discriminant Analysis

Logistic regression handles categorical as well as continuous independent variable and also the independent variables do not have to be normally distributed. Discriminant analysis, on the other hand, makes these requirements. Also, logistic regression analysis does not assume a linear relationship between the independent variable and dependent variable and can therefore handle nonlinear effects. Again, one can add explicit interaction and

power terms. These make logistic regression analysis flexible to use, as compared to discriminant analysis.

Other advantages of logistic regression analysis over discriminant analysis are that it makes no assumption of homogeneity of variance and normally distributed error terms are also not assumed. Again, logistic regression does not require that the independents be interval and it does not also require that the independents be unbounded.

The fewer assumptions and requirements by logistic regression analysis give it an advantage over discriminant analysis. In fact, the lack of statistical emphasis on discriminant analysis may be due to the greater inherent difficulty of the technical problems associated with it.

(Statgun, n.d)

Advantages of Discriminnant Analysis over Logistic Regression Analysis

With all the flexibility of assumptions and requirements by logistic regression analysis enumerated, one might wonder why anyone would ever use discriminant analysis rather than logistic regression analysis when analyzing data involving categorical dependent variable.

The unfortunate aspect of logistic regression analysis is that all the advantages come at a cost: it requires much more data to achieve stable and meaningful results. For logistic regression, the ratio of independent variable to sample of at least 1 to10 is necessary to achieve stable results, while that of the discriminant analysis is about 4 or 5 to 20. (Wikipedia, n.d).

Again, the discriminant analysis has more statistical power than logistic regression (less chance of type 2 errors - accepting a false null

hypothesis), therefore, it is preferred to logistic regression analysis when the assumptions of linearity, equal within-group variances, linearity between dependent and independent variables are met,

Choosing between Logistic Regression Analysis and Discriminant Analysis

Until the development of generalized linear model by John Nelder and Robert Wedderburn in 1972 as a way of unifying various other statistical models including linear regression, logistic regression and Poisson regression under one framework, discriminant analysis technique, which was developed by A.A Fisher in 1936, was the common technique which was mostly used to solve problems involving categorical latent variables.

Both logistic regression analysis and discriminant analysis are now often used to solve categorical dependent variable problems. However, there is the possibility that results produced by applying logistic regression and discriminant analysis for analysing the same set of data would be the same. In such a situation, the question of which of the two methods to use to derive the best results arises. The choice between the two techniques is dictated by the pertaining advantage(s) of one technique over the other at that instant. (Press & Wilson, 1978)

Empirical Comparison of Logistic Regression Analysis and Discriminant Analysis

This section focuses on the empirical analysis using both logistic regression and discriminant analysis to assess if the two techniques would

produce the same results so that in the absence of one of the techniques, the other can be used with the certainty of obtaining the same results. The empirical comparison analysis would be based on binary logistic regression and two-group discriminant analysis.

The Data

The data consists of operational performance of 47 financial industries in the United States of the years1968, 1969, 1970, 1971 and 1972. The data was collected by Moody's Industry to assess their performance. The aim was to check if a financial industry would go bankrupt or not two years after information about its operations had been gathered. Out of the 47 financial institutions sampled, it was observed that 22 of the institutions went bankrupt and 25 survived or did not go bankrupt, two years after the data was collected on them. (See appendix I for the data).

Statistical Tool for the Analysis

SPSS software was used in performing both the discriminant analysis and the logistic regression analysis of the data.

Analysis of the Data using Logistic Regression Analysis

Out of the 47 financial companies sampled, 45 were used as training sets and 2 (1 from each class) were used as holdup sample. The 45 samples were run with SPSS software with the simultaneous selection method. The dependent variable was bankruptcy status of a company. The companies

which went bankrupt were coded 0 and the companies which did not go bankrupt were coded 1. The independent variables were

1. The ratio of Cash Flow to the total debt of the financial institution, $\left(\frac{CF}{TD}\right)$.

2. The ratio of Net Income to Total Debt of the financial institution, $\left(\frac{NI}{TA}\right)$.

3. The ratio of Current Assets of the institution to the institution's Current Liabilities $\left(\frac{CA}{CL}\right)$.

4. The ratio of Current Assets to Net Sales of the financial institution, $\left(\frac{CA}{NS}\right)$.

In the analysis, the coefficients obtained for the independent variables $\left(\frac{CF}{TD}\right)$, $\left(\frac{NI}{TA}\right)$, $\left(\frac{CA}{CL}\right)$ and $\left(\frac{CA}{NS}\right)$ were 7.683, -4.13, 3.17 and -0.925 respectively. The constant term was also found to be -5.832. The resulting logistic regression equation was therefore given by

$$\pi_i = \frac{e^{-5.832+7.683x_1-4.13x_2+3.17x_3-0.925x_4}}{1+e^{-5.832+7.683x_1-4.13x_2+3.17x_3-0.925x_4}}, \text{ where } i = 0,1$$

and

$$x_1 = \left(\frac{CF}{TD}\right), \quad x_2 = \left(\frac{NI}{TA}\right), \quad x_3 = \left(\frac{CA}{CL}\right) \text{ and } x_4 = \left(\frac{CA}{NS}\right).$$

Table 3: Output of the Analysis using Simultaneous Method

Variable	B	S.E	Wald	D.f	Sig.	Exp(B)
x_1	7.683	6.026	1.625	1		2171.554
x_2	-4.130	13.855	0.089	1	0.766	0.016
x_3	3.170	1.159	7.473	1	0.006	23.797
x_4	-0.925	2.682	0.119	1	0.730	0.396
Constant	-5 .832	2.561	5.185	1	0.023	0.023

Multicollinearity among the independent variables shows that complete separation of the two groups of the dependent event variable can be perfectly done by scores on one or few of the independent variables.(See appendix I for correlation matrix)

In this analysis, multicollinearity in the data was detected. This was due to the fact that standard errors for the beta coefficients of $\left(\frac{CF}{TD}\right)$, $\left(\frac{NI}{TA}\right)$, $\left(\frac{CA}{CL}\right)$ and $\left(\frac{CA}{NS}\right)$ were found to be 6.026, 13.855, 1.159 and 2.682 respectively, some of which were greater than 2. This was again confirmed by the correlation matrix which shows a high correlation between $\left(\frac{CF}{TD}\right)$ and $\left(\frac{NI}{TA}\right)$ with a value of -0.872, giving an indication that the full model developed is not the best for prediction.

Due to the presence of multicollinearity in the data set, it was inappropriate to proceed with the analysis as it may give us misleading information about the data. The data was, therefore, re-run using the forward-

stepwise method which uses the parsimonious set of variables. This time, the variables $\left(\frac{CF}{TD}\right)$ and $\left(\frac{CA}{CL}\right)$ were identified as the significant variables, with corresponding coefficients of 6.495 and 2.978 respectively. The constant term was also found to be -5.875. The resulting reduced logistic regression equation was therefore given by

$$\pi = \frac{e^{-5.875+6.497x_1+2.978x_3}}{1+e^{-5.875+6.497x_1+2.978x_3}}, \text{ where } i = 0, 1.$$

Testing the Importance of the Independent Variables

The Log Likelihood Method

Even though, two independent variables were identified as the best set of variables for the model, it was still important to test for their association with the dependent variables.

The presence of a relationship between the dependent variable and combination of independent variables can be tested using the log likelihood method. This is based on the statistical significance of the model chi-square after the independent variables have been added to the analysis. It is obtained by finding the difference of -2loglikelihood of the model without any independent variable $(-2\text{loglikelihood}_R)$ and model with all the independents $(-2\text{loglikelihood}_F)$.

The hypothesis for the relationship between the dependent and the independent variables is given by

H_o = There is no association between the dependent and the independent variables.

H_1 = There is an association between the dependent and the independent variables.

The chi-square value for this analysis was 33.874, with a corresponding probability value of 0.000, which is significant at $\alpha = 0.05$. Thus, the null hypothesis that there was rejected. Hence, the significance of the overall relationship between the individual independent variables and the dependent variable supported the interpretation of the model using the independent variables in the data set.

Classification using the Logistic Regression Model.

Table 4: Classification Table with no Independent Variables

	Institution		Percentage Correctly
Observed Values	0	1	Classified
0	0	3	0
1	1	24	100
Overall Percentage			53.3

Table 5: Classification Table with Independent Variables

	Institution		Percentage Correctly
Observed Values	0	1	Classified
0	18	3	85.7
1	1	23	95.8
Overall Percentage			91.1

The classification table shows that 18 out of 21 bankrupt companies were correctly classified as bankrupt companies and 3 were misclassified as non-bankrupt companies (i.e 85.7% correct classification and 14.3% misclassification). Again, 23 out of 24 non-bankrupt companies were correctly classified as non-bankruptcy companies with 1 misclassified as a bankrupt company (i.e. 95.8% correct classification and 4.2% misclassification). The overall correct classification was found to be 91.1%.

This means that there is a high reliability in classifying a member into one of the two groups. However, correct classification of a company as a non-bankrupt company is slightly better than correct classification as bankrupt companies. In other words, it is slightly less to commit an error by classifying a non-bankrupt company as a bankrupt company than classifying a bankrupt company as a non-bankrupt company.

Also, the independent variables could be characterised as useful predictors which could perfectly distinguish between a non-bankrupt financial companies and a bankrupt financial companies if the classification accuracy rate was substantially higher than the accuracy attainable by chance alone. Operationally, the classification accuracy rate should be 25% (or more) higher than the proportion by chance accuracy rate. The proportion by chance accuracy rate is computed by first calculating the proportion of cases for each group based on the number of cases in each group in the classification table at the stage where no independent variable has been included in the modelling.

The proportion of the bankrupt company for this analysis was 0.467 and the proportion of non-bankrupt group was 0.533. The proportional by chance accuracy rate is the sum of the squares of the proportion of cases in

each group. And in this analysis, the proportion by chance accuracy rate is 0.5022. The accuracy rate computed by SPSS was 91.1% which was greater than the proportion by chance accuracy criteria of 25% or more improvement in the proportion by chance accuracy rate (62.77%). The criteria for classification accuracy is, thus, satisfied. Therefore, some or all the independent variables are useful for distinguishing between a bankrupt company and a non-bankrupt company.

The exp(B) of $\left(\frac{CF}{TD}\right)$ and $\left(\frac{CA}{CL}\right)$ were 663.272% and 19.654% respectively. Meaning, a unit increase in $\left(\frac{CF}{TD}\right)$ would increase the odds that a company would not go bankrupt by about 663 times and vise versa, when all other variables are held constant. Also, a unit increase in $\left(\frac{CA}{CL}\right)$ would increase the odds that a financial institution would not go bankrupt in the next two years by about 20 times, when all other variables are held constant.

The constant term of -5.875 means that the log(odds) of the logistic model when there was no independent variable was -5.875.

Analysis of the Data using Discriminant Analysis Technique

The analysis was carried out using the same data set which was used in the case of the logistic regression analysis. Once again, out of the 47 companies sampled, 45 were used as training sets and 2 (1 from each class) were used as a holdout sample. The 45 samples were run with SPSS software with the simultaneous selection method.

The variables $\left(\frac{NI}{TA}\right)$ and $\left(\frac{CA}{CL}\right)$ were identified as the most discriminating factors with conical coefficients of 5.41 and 0.87 respectively. The constant term was found to be -1.732. The discriminant function was, thus, given by

$$D_i = -1.732 + 5.41x_2 + 0.87x_3,$$

where

$$x_2 = \left(\frac{NI}{TA}\right)$$

$$x_3 = \left(\frac{CA}{CL}\right)$$

The discriminant analysis has several purposes but it is mostly used for separation of groups, prediction group membership of events or cases and above, for classification

Below the classification table of the discriminant analysis base the analysis above

Table 6: Classification Table of Discriminant Analysis

	Institution		Percentage Correctly
Observed Values	0	1	Classified
0	18	3	85.7
1	2	22	91.7
Overall Percentage			88.9

The classification table shows that 18 out of 21 bankruptcy companies were correctly classified as bankruptcy companies and 3 were misclassified as

not bankruptcy companies (85.7% correct classification and 14.3% misclassification). Again, 22 out of 24 non-bankruptcy companies were correctly classified as non-bankruptcy companies with 2 misclassified as bankruptcy companies (91.7% correct classification and 8.3% misclassification). The hit ratio (overall correct classification) was found to be 88.9%.

In the output, variable Wilks' λ s for , $\left(\frac{NI}{TA}\right)$, and $\left(\frac{CA}{CL}\right)$ were found to be , 0.696 and 0.627 respectively, with corresponding *p*-values of 0.000 for each.

The eigenvalue was 0.863, indicating that 86.3% of the discrimination between the bankruptcy company and the non-bankruptcy companies were accounted for by the discriminant model. The standard deviation of the variables $\left(\frac{NI}{TA}\right)$ and $\left(\frac{CA}{CL}\right)$ were found to be 0.1248, 1.0170 respectively. This shows that there was more variability in the sample of $\left(\frac{CA}{CL}\right)$ than the variable $\left(\frac{NI}{TA}\right)$.

The presence of a relationship between the dependent variable and combination of independent variables is based on the statistical significance of the model chi-square which is 26.9 and a corresponding *p*-value of 0.000. This shows that there existed a relation between the dependent variable and the two independent variables.

CHAPTER SIX

SUMMARY, DISCUSSION AND CONCLUSION

Introduction

In this research, so many things were observed. This chapter is to offer the researcher the opportunity to summarize, discuss and draw conclusions of his findings.

Summary

This research discussed logistics regression and discriminant analysis. The research showed that the logistic regression is a member of the exponential family while the discriminant analysis belongs to the linear family. Both techniques require that the response variable be categorical.

The discriminant analysis technique is for separation and classification while the logistic regression is for prediction and also for classification of cases. The two techniques have a common purpose selecting the “best” (and as few as possible) set of variables for developing a model that can be used to separate groups and also classify new cases.

Some major problems associated with discriminant analysis are that it requires that the distribution regarding the independent variables is normal. This makes logistic regression more flexible to use than discriminant as it goes with few assumptions and requirements. However, logistic regression also has

some disadvantages as it requires large data set for stability and also computationally cumbersome. It was also observed that all the techniques require that only independent variables that a relevant to the objectives of the study should be included in the selection process and the significance of these variables have to be tested with an appropriate statistical technique, even though the techniques for testing may be different.

It was observed that the independent variable that may be important for the purposes of separating groups or predicting group membership or for classifying a case in logistic regression analysis might not necessarily be important for predicting or classifying a case in the discriminant analysis. This is inferred from the empirical analysis as in the same data set, the logistic regression analysis and discriminant analysis came out with differences in choosing between the ratio of cash flow to total debt and the ratio of net income to total asset.

On the issue of classification capabilities of the two techniques, logistic regression can predict better than discriminant analysis. This is justified by correct classification of 91.1% by the logistic model and 88.9% by the discriminant function in the empirical analysis.

In the variable selection, the two techniques did not agree in choosing between the variable $\left(\frac{CF}{TD}\right)$ and $\left(\frac{NI}{TA}\right)$. Since the logistic regression does not put any assumption of homogeneity, its basic objective is to come up with a model which can predict best. It noticed that dropping $\left(\frac{NI}{TA}\right)$ and going for $\left(\frac{CF}{TD}\right)$seems to generate a model which can predict better. However, the

discriminant is more concerned with assumptions such as homogeneity of group variables. The standard deviation of the variable $\left(\frac{NI}{TA}\right)$ (i.e 0.1248) was less than the standard deviation of the variable $\left(\frac{CF}{TD}\right)$ (i.e 0.2636). Meaning that there was less variability in data for $\left(\frac{NI}{TA}\right)$ and this might have been the reason why the discriminant analysis chose the variable $\left(\frac{NI}{TA}\right)$ at the expense of the $\left(\frac{CF}{TD}\right)$. This is because when logistic regression made use of the independent, $\left(\frac{CF}{TD}\right)$ which was less homogeneous at the expense of $\left(\frac{NI}{TA}\right)$, it was able to come out with a model which can make 91.1% correct classification of cases as against the discriminant analysis which selected $\left(\frac{NI}{TA}\right)$ at the expense of $\left(\frac{CF}{TD}\right)$, and was able to come out with a model which made 88.9% correct classification of cases..

Discussion

The logistic regression is a member of the exponential family with a general formula of the form

$$\pi_{\mathrm{i}} = \frac{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}}{1+e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}},$$

where $\beta_1, \beta_2, \ldots, \beta_k$ and the coefficients of the independent variables $x_1, x_2, \ldots, x_k$ and β_0 is the constant term of the logistic regression equation but

the discriminant analysis belong to the linear family with a general formula of the form

$$L = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p,$$

where $\beta_1, \beta_2, \ldots, \beta_k$ are the coefficients of the independent variables $x_1, x_2, \ldots, x_k$ that best discriminate between the two groups, and β_0 is the constant term of the discriminant function.

Even though the two techniques have different models, the models are all formulated with a basic objective of showing the predictive capability of the individual independent variables and also classifying new cases into any of the two predefined groups. The coefficients of the independent variables determine the independents contribution in classifying a case. A high value of the coefficients suggests a high contribution of that independent. A positive value of the coefficients indicates that that independent variable contributes positively while a negative value indicates a negative contribution of that individual variable in predicting the group membership of subjects or classification of cases.

A reliable or good model should be able to come up with high correct classification with less or few misclassification. The question of whether the application of logistic regression and discriminant analysis techniques on the same data set would yield the same result, or whether independent variables of the same data set would be significant in the two cases is clarified by the empirical analysis above which shows that logistic regression can perform better than discriminant analysis.

Validation of Empirical Results

It is always important to assess the utility of generated logistic model. In this research, the validation of generated logistic model would be carried out using the logistic regression model developed above and the holdout samples from the data.

From the analysis above, the generated logistic regression model was given by

$$\pi_i = \frac{e^{-5.875+6.497\,x_1+2.978\,x_3}}{1+e^{-5.875+6.497\,x_1+2.978\,x_3}}$$

In linear terms, the model was given by

$$\text{logit}\,(\pi_i) = -5.875 + 6.497x_1 + 2.978x_3,$$

where

$$x_1 = \frac{\text{Cash Flow}}{\text{Total Debt}}$$

and

$$x_3 = \frac{\text{Current Assets}}{\text{Current Liabilities}}$$

In the holdout sample, a financial company which went bankrupt had the following observed variable values:

1. $\frac{\text{Cash Flow}}{\text{Total Debt}}\ (x_1) = 0.07,$ 2. $\frac{\text{Net Income}}{\text{Total Assets}}\ (x_2) = 0.02$

3. $\frac{\text{Current Assets}}{\text{Current Liabilities}} (x_3) = 1.31$ and 4. $\frac{\text{Current Assets}}{\text{Net Sales}}\ (x_4) = 0.25$

The classification of this company into a bankrupt company or a non-bankrupt company by the logistic model was given by

$$\pi_i = \frac{e^{-5.875+6.497(0.07)+2.978(1.31)}}{1+e^{-5.875+6.497(0.07)+2.978(1.31)}}, \text{ where i = 0,1}$$

$$= 0.18$$

Since 0.18 < 0.5 it means that the company belongs to the bankrupt companies (0 groups). Hence the logistic model was able to correctly classify the company as a bankrupt company.

Also, a financial company which did not go bankrupt two years after the data was collected on its operation had the following observed variable values:

1. $\frac{\text{Cash Flow}}{\text{Total Debt}}(x_1) = 0.14,$ $\frac{\text{Net Income}}{\text{Total Assets}}(x_2) = 0.07$

$\frac{\text{Current Assets}}{\text{Current Liabilities}}(x_3) = 2.61$ and $\frac{\text{Current Assets}}{\text{Net Sales}}(x_4) = 0.52$

The classification of this company into a bankrupt company or a non-bankrupt company by the logistic model was given by

$$\pi_i = \frac{e^{-5.875+6.497(0.14)+2.978(2.61)}}{1+e^{-5.875+6.497(0.14)+2.978(2.61)}}, \text{ where i = 0,1}$$

$$= 0.9393$$

Since 0.9393 > 0.5 it implies that the company belonged to the non-bankrupt companies. Hence the model was able to correctly classify the financial company as non-bankrupt two years after its operation.

It was also important to assess the utility of the generated discriminant function. The model was validated by testing the two holdout samples with the model. The centroid of the bankrupt group and non-bankrupt group by the discriminant output were respectively 0.173and 0.849. The mean of these centroid was 0.51. That meant that a financial company with a discriminant

score more than 0.51 was unlikely to go bankrupt two year after its operations, and a company with a discriminant score less than 0.51 was likely to go bankrupt two years.

The formulated discriminant function was given by

$$D_i = -1.732 + 5.41x_2 + 0.87x_3$$

In the holdout sample, information about the bankrupt financial company which was used to cross-validate the logistic regression was also used to cross-validate the discriminant function. The discriminant score for this company was therefore given by

$$D_i = -1.732 + 5.41(0.02) + (0.87)(1.31)$$

$$D_i = -0.4841$$

Since $D_i = -0.4841 < 0.51$ it meant that the company belonged to the bankrupt companies. Hence the model was able to correctly classify the financial company as a bankrupt company.

Also, in the holdout sample, information about the non-bankrupt financial company which was used to cross-validate the logistic regression was also used to cross-validate the discriminant function. The discriminant score for this company was therefore given by

$$D_i = -1.732 + 5.41(0.07) + (0.87)(2.61)$$

$$D_i = 0.9174$$

Since $0.9174 > 0.51$ it meant that the company belonged to the non-bankrupt companies. Hence the model was able to correctly classify the financial company did not go bankrupt two years after its operation.

In using the holdout sample to test the efficient of the models built, both the logistic regression and the discriminant analysis models were able to make correct classification of the bankrupt companies and the non- bankrupt companies.

Conclusion

Based on the research, it is evident that logistic regression yields better classification results than discriminant analysis when there is a problem of multicollinearity in the independent variable. This is because empirical comparison of the two techniques shows that logistic regression was able to make 91.1% correct classification of the data as compared to 88.9% correct classification of the data by discriminant analysis where the data has multicollinearity among some of the independent variables.

Another interesting observation, based on the analysis, was that despite the obvious revelation of both techniques showing very reliable model for prediction and classification of cases, they do not, in principle, always agree in variable selection for development of models. This disagreement might be due to consideration given to factors such as variability in the data set. Hence in adopting any of the two techniques for analysis, consideration has to be given to the assumptions, requirements and any circumstance surrounding the data at that moment.

However, it is obvious that logistic regression analysis can yield better results than discriminant analysis. In this wise, the researcher can conceive that when all the assumptions and requirements of logistic regression analysis

and discriminant analysis are met, logistic regression technique should be adopted rather than discriminant analysis.

REFERENCES

Abbott R.D.(1985). Logistic regression in survival analysis. *American Journal of Epidemiol*, 121(3): 465 - 471. , National Heart, Lung, and Blood Institute Bethesda, MD 20205.

Abdi, H.(2007). Encyclopedia of measurement and statistics. Thousand oak(CA) sage, 270-275.[Retrieved October 15, 2008 from http;//www.ut dallas.edu]

Abrahamowicz, M., du Berger , R. & Graver, S.A.(1997). Flexible Modeling of the Effects of Serum Cholesterol on Coronary Heart Disease Mortality . *American Journal of Epidemiol*, 145, 714–29.

Ahmed,K., Alam, K.F &Alam, M. (1997). Working papers of School of Business. *Journal of Accounting Education,* 6, 325-335.

Argilés, J. M. (1998). Accounting information and prediction of farm viability. Economic working papers, Dept. of Economic and Business, Universitat Pompeu Fabra, 277.

Bartlett, M. S. (1951). An inverse matrix adjustment arising in discriminant analysis (1951), *Annals of Mathematical Statistics*, 22**,** 107−111.

Betensky, A. and Williams, P.L.(2001). A comparison of models of clustered binary outcomes: analysis of a designed immunology experiment. Appl.Statistics, 50, (1), 43-61.

Clark, W. H., Elder,D.E., Guerry, D., Braitman,L.E., Trock, B.J., Schultz, D., et al. (1989). Model for predicting survival in stage I melanoma based on tumor progression. *Journal of the National Cancer Institute*, 81,(24), 1893-1904.

Garson, G.D.(n.d). Discriminant analysis . Retrieved September 29, 2008 from http://faculty.chass.ncsu.edu/Garson/pa765/discrim.htm.

Garson, G.D.(n.d). Logistic regression . Retrieved September 29, 2008 from http://faculty.chass.ncsu.edu/Garson/pa765/logistic.htm.

Geller , A C., Johnson, M.D., Miller, D.R., Brooks, K.R., Layton, C.J., Susan, M. et al. (2009). Factors associated with physical discovery early melanoma in middle-aged and older men. *Arch Dermatol*, 145(4), 409-414.

Gestel,T.V., Baesens, B., Suykens, J.A.K., Poel, D.V. Baestaens & Willekens, B.M. (2004). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. Working Papers of Faculty of Economics and Business Administration, Ghent University, Belgium, 04/247.

Gordon D.J., Probstfield, J. L., Rubenstein.C., Bremner, W. F., Leon, A.S., Karon, J.M. et al. (1984). Coronary risk factors and exercise test performance in asymptomatic hypercholesterolemic men: Application of Proportional Hazards Analysis. *American Journal of Epidemiology,* 120, (2), 210-224.

Hirota, S.(1999). Application of multivariate analysis to clinical medicine. *American Journal of Epidemiology,* 134, (2), 232-244.

Hosmer, D. W. & Lemeshow, S., & Klar, J. (1988). A goodness-of- fit test for the multiple regression model. *Communications in Statistics*, A10, 1043-1069.

Hosmer, D. W & Lemeshow, S. (2000). Applied Logistic Regression, 2nd ed, 1-2 . Wiley and sons inc.

Ito, T., Nishimura, S., Saito, M. & Omori, Y. **(**1997). The level of erythrocyte aldose reductase: a risk factor for diabeticneuropathy? *Diabetes Research and Clinical Practice*, 36(3) 161-167.

Kennedy, J.W. , Kaiser, G.C., Fisher, L.D. ,Maynard,C., Fritz, J.K., Myers, W. et al. (1980) -Multivariate discriminant analysis of the clinical and angiographic predictors of operative mortality from the Collaborative Study in Coronary Artery Surgery. *The Journal of Thoracic and Cardiovascular Surgery*, 80, 876-887.

Kirschenbaum, M., Oigenblick K., & Goldberg, S. (2000). Analysis of the predictors or work accident proneness. Chest, 118-125.

Laika, S.H. et al. (2003). Discriminant analysis of different levels of diabetic neuropathies recorded by somatosensory-evoked potentials. *American Journal of Epidemiology*.67(5), 143-153.

Luna (n.d) Statnotes. Retrieved 29 September, 2008 from Http//Luna.cas. us.edu/~mbrannic/files/regression/logistic.html.

McCullagh, P. & Nelder, J. A. (1989). Generalized Linear Models, 2nd ed. 82-87. London:Chapman & Hall.

Menard, S (2002). Applied logistic regression analysis, 2nd ed. 1-16. Thousand Oaks, CA: Sage Publications. Series:

Nelder, J. & Wedderburn, R. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society.* Series A (General), 135, 370-384.

Pérez, S. (2006). Activity Optimization method for Nuclear Medicine Planar Studies based on Discriminant Analysis: A comparison with ROC Curve. *Alasbimn Journal* 8(31): Article N°AJ31-2.

Press, S. J. & Wilson, S. (1978). Choosing between logistic regression and

discriminant analysis. *Journal of the American Statistical Association*, :85, 699-705.

Pullinger, A.G., Seligner, D.A. & Gornbein, J.A. (1993). A multiple logistic regression analysis of the risk and relative odds temporamandibular disorder as function of common occlusal features. *Journal of dental research*, 72. (6) 968-979.

Smith G. C. S.(2005). A proportional hazards model with time-dependent covariates and time-varying effects for analysis of fetal and infant death. *American Journal of Epidemiol* ,161, 100-102.

Statgun (n.d). Statgun statistics. Retrieved October 6,2008 from www.statgun.com/tutorials/logistic-regression.html.

Tabachnick, Barbara G. & Fidell, L.S. (2001). *Using multivariate statistics*, 4th.ed. Boston: Allyn and Bacon. 53-74.

Takahashi, Y., Tachikawa, T., Ito, T., Takayama, S., Omori, Y., & Iwamoto, Y.(1998). Erythrocyte aldose reductase protein: a clue to elucidate risk factors for diabetic neuropathies independent of glycemic control. *Diabetes Research and Clinical Practice*, 42 (2), 101-107.

Walter, S.D., Feinstein, A. R. & Wells, C. K. (1987). Coding ordinal independent variables in multiple regression analyses. *American Journal of Epidemiol*, 125, 319 – 323.

Wang, J., Xiao, F., Ren, J., Li, Y & Zhang, M.L. (2007). Risk factors for mortality after coronary bypass grafting with patients with low left ventricular ejection fraction. *Chinese Medical Journal*,120(4), 317-322. Department of Cardiac Surgery, Peking University First Hospital, Beijing, China. 100034.

Wikipedia (n.d).Logistic regression notes. Retrieved October 7, 2008 from www.en.wikipedia.org/wiki.

Wright, J.G., Pifarré ,R., Sullivan., H.J., Montoya, A., Bakhos, M., Grieco., J, et al.(1987), Multivariate discriminant analysis of risk factors for operative mortality following isolated coronary artery bypass graft. *Loyola University Medical Centre experienc*e. 91(3), 394-399.

APPENDIX I

Some Outputs of Logistic Regression Analysis

Classification Table[a,b]

			Predicted		
			X1		Percentage
	Observed		.0	1.0	Correct
Step 0	X1	.0	0	21	.0
		1.0	0	24	100.0
	Overall Percentage				53.3

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	.134	.299	.200	1	.655	1.143

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	X1	15.557	1	.000
		X2	13.688	1	.000
		X3	16.766	1	.000
		X4	.005	1	.943
	Overall Statistics		21.266	4	.000

Correlation Matrix

		Constant	X1	X2	X3	X4
Step 1	Constant	1.000	-.598	.518	-.837	-.427
	X1	-.598	1.000	-.872	.473	.210
	X2	.518	-.872	1.000	-.470	-.103
	X3	-.837	.473	-.470	1.000	-.093
	X4	-.427	.210	-.103	-.093	1.000

Iteration History[a,b,c,d]

Iteration		-2 Log likelihood	Coefficients				
			Constant	X1	X2	X3	X4
Step 1	1	37.416	-1.377	1.079	3.774	.84	-.626
	2	31.024	-2.902	2.706	3.179	1.70	-1.002
	3	28.651	-4.666	5.371	-.528	2.61	-1.052
	4	28.317	-5.631	7.246	-3.391	3.08	-.964
	5	28.309	-5.825	7.669	-4.103	3.16	-.927
	6	28.309	-5.832	7.683	-4.130	3.17	-.925

a. Method: Enter

b. Constant is included in the model.

c. Initial -2 Log Likelihood: 62.183

d. Estimation terminated at iteration number 6 because log-likelihood decreased b less than .010 percent.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	33.874	4	.000
	Block	33.874	4	.000
	Model	33.874	4	.000

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	4.487	7	.722

Classification Table[a]

Observed			Predicted		
			X1		Percentage Correct
			.0	1.0	
Step 1	X1	.0	18	3	85.7
		1.0	1	23	95.8
	Overall Percentage				91.1

a. The cut value is .500

Correlation Matrix

		Constant	X1	X2	X3	X4
Step 1	Constant	1.000	-.598	.518	-.837	-.427
	X1	-.598	1.000	-.872	.473	.210
	X2	.518	-.872	1.000	-.470	-.103
	X3	-.837	.473	-.470	1.000	-.093
	X4	-.427	.210	-.103	-.093	1.000

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	X1	15.557	1	.000
		X2	13.688	1	.000
		X3	16.766	1	.000
	Overall Statistics		21.127	3	.000

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	X1	7.683	6.026	1.625	1	.202	2171.554	.016	2.9E+08
	X2	-4.130	13.855	.089	1	.766	.016	.000	1.0E+10
	X3	3.170	1.159	7.473	1	.006	23.797	2.453	230.902
	X4	-.925	2.682	.119	1	.730	.396	.002	75.971
	Constant	-5.832	2.561	5.185	1	.023	.003		

a. Variable(s) entered on step 1: X1, X2, X3, X4.

The data for the analysis

X1	X2	X3	X4	P
-0.45	-0.41	1.09	0.45	0
-0.56	-0.31	1.51	0.16	0
0.06	0.02	1.01	0.4	0
-0.07	-0.09	1.45	0.26	0
-0.1	-0.09	1.56	0.67	0
-0.14	-0.07	0.71	0.28	0
0.04	0.01	1.5	0.71	0
-0.06	-0.06	1.37	0.4	0
0.07	-0.01	1.37	0.34	0
-0.13	-0.14	1.42	0.44	0
-0.23	-0.3	0.33	0.18	0
0.07	0.02	1.31	0.25	0
0.01	0	2.15	0.7	0
-0.28	-0.23	1.19	0.66	0
0.15	0.05	1.88	0.27	0
0.37	0.11	1.99	0.38	0
-0.08	-0.08	1.51	0.42	0
0.05	0.03	1.68	0.95	0
0.01	0	1.26	0.6	0
-0.28	-0.27	1.27	0.51	0
0.12	0.11	1.14	0.17	0
0.51	0.1	2.49	0.54	1

0.08	0.02	2.01	0.53	1
0.38	0.11	3.37	0.35	1
0.19	0.05	2.25	0.33	1
0.32	0.07	4.24	0.63	1
0.31	0.05	4.45	0.69	1
0.12	0.05	2.52	0.69	1
-0.02	0.02	2.05	0.35	1
0.22	0.08	2.35	0.4	1
0.17	0.07	1.8	0.52	1
0.15	0.05	2.17	0.55	1
-0.1	-0.01	2.5	0.58	1
0.14	-0.03	0.46	0.52	1
0.15	0.06	2.23	0.56	1
0.16	0.05	2.31	0.2	1
0.29	0.06	1.84	0.38	1
0.54	0.11	2.33	0.48	1
-0.33	-0.09	3.01	0.47	1
0.48	0.09	1.24	0.18	1
0.56	0.11	4.29	0.45	1
0.2	0.08	1.99	0.3	1
0.47	0.14	2.92	0.45	1
0.17	0.04	2.45	0.14	1
0.58	0.04	5.06	0.13	1

APPENDIX II

SOME OUTPUTS OF DISCRIMINANT ANALYSIS

Variables Entered/Removed[a,b,c,d]

Step	Entered	Wilks' Lambda							
						Exact F			
		Statistic	df1	df2	df3	Statistic	df1	df2	Sig.
1	X2	.627	1	1	43.000	25.535	1	43.000	.000
2	X3	.537	2	1	43.000	18.120	2	42.000	.000

At each step, the variable that minimizes the overall Wilks' Lambda is entered.

a. Maximum number of steps is 8.

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.

Variables in the Analysis

Step		Tolerance	F to Remove	Wilks' Lambda
1	X3	1.000	25.535	
2	X3	.961	12.441	.696
	X2	.961	7.089	.627

Variables Not in the Analysis

Step		Tolerance	Min. Tolerance	F to Enter	Wilks' Lambda
0	X1	1.000	1.000	22.721	.654
	X2	1.000	1.000	18.798	.696
	X3	1.000	1.000	25.535	.627
	X4	1.000	1.000	.005	1.000
1	X1	.890	.890	6.583	.542
	X2	.961	.961	7.089	.537
	X4	.981	.981	.361	.622
2	X1	.339	.339	.487	.531
	X4	.980	.945	.396	.532

Wilks' Lambda

Step	Number o Variables	Lambda	df1	df2	df3	Exact F			
						Statistic	df1	df2	Sig.
1	1	.627	1	1	43	25.535	1	43.000	.000
2	2	.537	2	1	43	18.120	2	42.000	.000

Summary of Canonical Discriminant Functions

Classification Table[a]

			Predicted		
			X1		Percentage
	Observed		.0	1.0	Correct
Step 1	X1	.0	18	3	85.7
		1.0	2	22	91.7
	Overall Percentage				88.9

a. The cut value is .500

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.863[a]	100.0	100.0	.681

a. First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.537	26.129	2	.000

Standardized Canonical Discriminant Function Coefficients

	Function
	1
X2	.570
X3	.717

Structure Matrix

	Function
	1
X3	.830
X2	.712
X1[a]	.689
X4[a]	.131

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

a. This variable not used in the analysis.

Canonical Discriminant Function Coefficients

	Function
	1
X2	5.410
X3	.879
(Constant)	-1.732

Unstandardized coefficients

Functions at Group Centroids

	Function
X1	1
.0	.173
1.0	.849

Unstandardized canonical discriminant functions evaluated at group means

Classification Statistics

Prior Probabilities for Groups

		Cases Used in Analysis	
X1	Prior	Unweighted	Weighted
.0	.500	21	21.000
1.0	.500	24	24.000
Total	1.000	45	45.000