COMPARING DISCRIMINANT ANALYSIS AND LINEAR REGRESSION ANALYSIS TO PREDICT THE ALCOHOL CONSUMPTION BY HIGH SCHOOL STUDENTS

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF MASTERS IN SCIENCE

IN THE GRADUATE SCHOOL OF THE

TEXAS WOMAN'S UNIVERSITY

DEPARTMENT OF MATHEMATICS AND COMPUTER SCIENCE

COLLEGE OF ARTS AND SCIENCES

BY

TAMANNA HAMAL, B.S.

DENTON, TEXAS

MAY 2017

ACKNOWLEDGEMENTS

I would firstly like to thank Dr. David Marshall for his endless support and encouragement. I would also like to thank Dr. Brandi Falley for her valuable guidance and suggestion to complete my thesis. I would also like to thank Dr. Don Edwards for his encouragement and support on my every step, not only to complete my thesis, but also to complete my Master's Degree.

And lastly, I would also like to thank my family, especially my husband, for his love, encouragement, support, and patience to achieve my educational goal. Also to my children, you are pillar of my strength and your support and belief in me has been a driving factor to push myself.

ABSTRACT

TAMANNA HAMAL

COMPARING DISCRIMINANT ANALYSIS AND LINEAR REGRESSION ANALYSIS TO PREDICT THE ALCOHOL CONSUMPTION BY HIGH SCHOOL STUDENTS

MAY 2016

The purpose of the study is to compare the Discriminant Analysis and Linear Regression Analysis to predict the correlation between alcohol consumption by high school students and their social attributes and grades. Discriminant Analysis, developed by R. A. Fisher in 1936, is a statistical technique used to determine which variables discriminate between two or more mutually exclusive naturally occurring groups.

Linear Regression Analysis is the most widely used statistical technique where straight lines are fitted to patterns of data. In this model, the dependent variable, the variable of interest, is predicted from independent variables using a linear equation. Even though the earliest form of linear regression was the Method of Least Squares, which was published by Legendre in 1805, and by Gauss in 1809, the term *regression* was pioneered by Sir Francis Galton. Regression analysis is the process of finding out the relationship between one or more dependent variables and the independent variables.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION

Many studies show that alcohol is the most commonly used and abused drug among adolescents, especially in high school students, all over the world. Even though drinking under 21 years of age is illegal in the USA, 11% of the total alcohol is consumed by 12 to 20 year-olds and more than 90% of this is consumed via binge drinking. The students who drink alcohol are more likely to have some problems of higher absence, poor or failing grades in school and social, physical, and legal problems, such as less participation in youth activities, arguments with family, drinking and driving, accidents, physical and sexual assault, problems in brain development, etc. In my thesis, I am analyzing the major reasons of youth drinking through the use of two statistical techniques, discriminant analysis and linear regression analysis, so it may help in the reduction of youth drinking by minimizing the reasons which have direct effects on youth drinking. In addition, I am using these two statistical techniques to find out major reasons of youth drinking as well as evaluating the effectiveness of these techniques for the study.

In this paper I am using these two statistical techniques, discriminant analysis and linear regression analysis, in a statistical software SPSS, Statistical Package for Social Sciences, to find out the correlation between student alcohol usage and the social attributes, such as gender, age, family size, parents' cohabitation status, quality of family relationship, guardianship, number of absences in school, grades, etc., for each student.

My study is based on the data set “Student Alcohol Consumption” obtained from UCI Machine Learning Repository at the University of California. This data set was collected in April 2008 and donated by Fabio Pagnotta and Hossain Mohammad Amran, Department of Computer Science, University of Camerino, Italy on 03/03/2016. This data set contains 649 students’ records (observations) and 32 attributes with 12 binaries, 4 nominals and 16 numeric. Among these attributes, I am focusing on some major attributes, like students’ age, sex, family size, parents’ cohabitation status, parents’ education, quality of family relationship, guardianship, study time, internet access at home, absences, romantic relationship, motivation towards getting higher education, going out with friends, and grades to predict the workday alcohol consumption by the students.

CHAPTER II

LITERATURE REVIEW

My study is more focused on the use of statistical techniques, discriminant and linear regression analysis with the help of the computer program SPSS, so I have done a detailed study on these techniques and their derivations. In "Discriminant Analysis" by William R. Klecka, the author described when and how to use this technique, its assumptions, and derivation of the functions in detail. In "General Linear Model, A "New" Trend in Analysis of Variance" by Maurice M. Tatsuoka, the author described the history and different approaches of Linear Regression Analysis and their formulation. In "Linear Regression Analysis: Theory & Computing" by Yan, Xin, Su, and Xiao Gang, the authors described and formulated the linear regression model in detail. In addition, I am reviewing some other publications, like "Notes on Linear Regression Analysis" by Robert Nau at Proquest ebrary, "Linear Regression Analysis using SPSS Statistics," the website www.statistics.laerd.com, "SPSS Data Analysis Example: Discriminant Function Analysis," the website http://www.ats.ucla.edu/stat/spss/dae/discrim.htm, and "Discriminant Function Analysis" by John Poulsen and Aaron French for references. In addition, I am reviewing "Facts Sheets – Underage Drinking" published by the Centers for Disease Control and Prevention, and "Alcohol Use Beliefs and Behaviors among High School Students" by Feldman, Harvey, Holowaty, and Shortt which have explained the facts on drinking among youth, its consequences, the relationship between alcohol

consumption and various socio-demographic and lifestyle behaviors, and assist in the development and implementation of alcohol abuse prevention programs.

CHAPTER III

PREDICTIVE MODELS

Discriminant Analysis

Discriminant Analysis, developed by R. A. Fisher in 1936, is a statistical technique which is used to determine which variables discriminate between two or more mutually exclusive naturally occurring groups. Data cases, such as people, animals, countries, economy at different times, etc., are the basic units of analysis. This technique is used in a wide variety of situations in the field of social sciences. Other areas where it is used are personal placement testing, roll call analyses of legislatures, psychological testing of children, effects of medical treatments, economic differences between geographic regions, predicting voting behavior, terrorist activity, etc. (Klecka, 1982, 8). It works on data sets where pre-specified and well-defined groups already exist and assesses the dependent relationship between independent and dependent variables. In discriminant analysis, some basic assumptions are applied. The characteristics used to differentiate between groups are called discriminating variables. These variables must be measured at a certain interval or ratio level of measurement, so that we can calculate means and variances and employ in mathematical equations. Generally, there is no limit on the number of discriminating variables if the total number of cases exceeds the number of variables by more than two. However, there are some limits on statistical

properties, for example, no variable can be used as the linear combination of other discriminating variables. A “linear combination” is the sum of one or more variables multiplied by weighted constant terms. Thus, neither the sum nor the average of several variables along with all these variables can be used because the variable defined after the linear combination does not contain any new information besides what is contained in the original variables. Similarly, two or more perfectly correlated variables cannot be used at the same time. Another assumption is that the population covariance matrices are equal for each group, and each group drawn from a population has a multivariate normal distribution on the discriminant variables (Klecka, 1982, 9-11).

These assumptions are:

1) There should be two or more groups,
2) There should be at least two cases per group,
3) The number of discriminating variables should be less than number of cases minus two,
4) Discriminating variables are measured at the interval level,
5) Discriminating variables cannot be a linear combination of other discriminating variables,
6) The covariance matrices for each group must be equal or approximately equal, and
7) Each group must be from a population with multivariate normally distributed discriminant variables.

Linear Regression Analysis

Linear Regression Analysis is the most widely used statistical technique where straight lines are fitted to patterns of data. In this model, the dependent variable (also called response variables, explained variables, predicted variables, or regressands, usually denoted by Y), the variable of interest, is predicted from independent variables (also called predictors, explanatory variables, control variables, or regressors, usually denoted by X_1, X_2, …, X_k) using a linear equation. Even though the earliest form of linear regression was the Method of Least Squares, which was published by Legendre in 1805, and by Gauss in 1809, the term *regression* was pioneered by Sir Francis Galton, a nineteenth century scientist to describe the biological phenomenon, heights of tall ancestors tend to regress down to an average height (Nau, 2014). Regression analysis is the process of finding out the relationship between one or more dependent variables and the independent variables. This technique is only appropriate if the data passes six assumptions which are required for this model to get a valid result. These assumptions are

1) Two variables should be in continuous level, i.e., they are either interval or ratio variables,
2) Two variables should have a linear relationship,
3) There should be no significant outliers,
4) There should be independence of observations,

5) The data should show homoscedasticity, i.e., the variances along the line of best fit should remain similar along the line, and
6) The residuals (errors) of the regression line should be approximately normally distributed.

Linear regression is used in the fields of biology, environmental and social sciences, and in the study of economic status of the country at different time periods. This is one of the most important tools used in these fields.

In this paper, we are using these two techniques, discriminant analysis and linear regression analysis, in SPSS program to find out the correlation between student alcohol usage and the social attributes, gender and study time for each student.

CHAPTER IV

MATHEMATICS OF THE MODELS

Discriminant Analysis

Computational Approach

Computationally, discriminant analysis is similar to analysis of variance (ANOVA). Specifically, we can find if two or more groups are significantly different from each other with respect to the mean of a particular variable. If the means for a variable are significantly different in different groups, we can conclude that this variable discriminates between the groups. In the case of a single variable, an *F* test, the ratio of the between-groups variance over the pooled (average) within-group variance, is the final significance test to find out whether or not a variable discriminates between groups. If the between-group variance is significantly larger, we can conclude that the means for a variable in different groups are significantly different and this variable discriminates between groups. However, there are multiple variables in the study and we must see which one(s) has strong contribution to discrimination between groups. In such cases, we use the matrix of total between-group variances and covariances, and the matrix of pooled within-group variances and covariances. Then, through the multivariate *F* test, we compare these two matrices to find out the significant difference between groups with respect to all variables. This process is similar to multivariate analysis of variance (MANOVA). If the multivariate *F* test is statistically significant, then we proceed to see

which of the variables have significantly different means among the groups ("Discriminant Function Analysis" 2-3).

Stepwise Discriminant Analysis

In stepwise discriminant analysis, a model of discrimination is built step by step. At each step, all the variables will be studied to find out the best variable which has the strongest contribution to the discrimination between groups. Then, that variable will be included in the model and the process is repeated. This is called forward stepwise analysis. In another case, we can include all the variables in the model, and at each step, the variable, which has the weakest contribution to the discrimination between groups, will be eliminated. This is called backward stepwise analysis. In this way we have those variables in the model which have a strong contribution to the discrimination between groups. We use respective *F* to enter and *F* to remove values, which indicates its statistical significance in the discrimination between groups, to do stepwise procedure.

Canonical discriminant functions. When there are two groups, discriminant analysis is called Fisher Linear Discriminant Analysis, which is similar to Multiple Regression. So, in the two groups, it has the following mathematical form of a linear equation:

$$\text{Group} = a + b_1*x_1 + b_2*x_2 + ... + b_m*x_m$$

where, a is a constant, and b_1 through b_m are regression coefficients.

When there are more than three groups, we use a canonical discriminant function, a linear combination of the discriminating variables. Its mathematical form is

$$f_{km} = u_0 + u_1 X_{1km} + u_2 X_{2km} + \ldots + u_p X_{pkm}, \quad [1]$$

where

f_{km} = the value on the canonical discriminant function for case m in the group k,

X_{ikm} = the value on discriminating variable X_i for case m in group k; and

u_i = coefficients which produce the desired characteristics in the function.

The coefficients for the first function are derived to maximize the difference between group means. The coefficients for the second function are also derived to maximize the difference between group means but the values on the second function are not correlated with the values on the first function. Similarly, the coefficients for the third function can be derived having maximum difference in group means while being uncorrelated with previous functions and so forth. The maximum number of unique functions we derived in this way is equal to the number of groups minus one or the number of discriminant variables, whichever is fewer (Klecka, 1982, 15-16).

To find the interrelations among variables, group means and standard deviations are not sufficient. So, we use the matrix of total sums of squares and cross-products, T, which is a square symmetric matrix. To derive T, we need some notation as follows:

g = number of groups

n_k = number of cases in group k

n. = total number of cases in all groups

X_{ikm} = the value of variable i for case m in group k

$X_{ik.}$ = mean value of variable i for those cases in group k

$X_{i..}$ = mean value of variable i for all cases (grand or total mean)

Now,

$$t_{ij} = \sum_{k=1}^{g} \sum_{m=1}^{n_k} (X_{ikm} - X_{i..}) \left(X_{jkm} - X_{j..}\right) \quad [2]$$

The terms in the parentheses are the value of a particular case deviated from the grand mean on the variable. When i = j, the two terms are the same and we just square the deviation. Thus, each diagonal element is the sum of squared deviations from the grand mean, which indicates how the cases are spread out on a single variable. When $i \neq j$, it will be the sum of the deviation on one variable multiplied by the deviation on the other variable. In this way, we can measure the correlation (covariation) between two variables, because it helps to find out how well a large deviation on one variable corresponds to a large deviation on the other. By taking the entire matrix, we can have the dispersion, a summary of the points' spread out around the total space by all the variables.

If each element of T is divided by (n.-1), we would get the total covariance matrix, which can also be used in computations of discriminant analysis. Covariance matrices are calculated for each group when they are based only on the cases for that

group. To get how strongly any two variables are related, we should examine the correlation between them. The correlation coefficient is more useful than the covariance for getting the strength of the relation of any two variables. It is standardized between -1 to +1. We can easily convert the T matrix into a correlation coefficient matrix by dividing each element by the square root of the product of the two diagonal elements falling on the same row and column. If the group locations, i.e. centroids, are not identical, the degree of dispersion within the group is less than the total dispersion, and this is measured by the W matrix, the within-group sums of squares and cross-products matrix. W is similar to T except that the deviations are measured from the mean of the group instead of grand mean. The elements of W are as follows:

$$w_{ij} = \sum_{k=1}^{g} \sum_{m=1}^{n_k} (X_{ikm} - X_{ik.}) \left(X_{jkm} - X_{jk.}\right) \quad [3]$$

If the elements of W are divided by (n.-g), the within group covariance matrix, which is a weighted average of the group covariance matrices, will be found. Similarly, we can convert W or the within-group covariance matrix into a within-groups correlation matrix by the same procedure as the total correlation matrix. Each correlation coefficient measures the strength of the relationship between the corresponding pair of variables within the group. These within-groups correlations are better estimates of the relationship between the variables than the total correlations. When group centroids are not different, all the elements of W will be equal to the corresponding elements of T. However, if the centroids are different, the elements of W will be smaller than the corresponding elements of T. This difference can be measured by matrix B which is called the between-

groups sums of squares and cross-products matrix. It is defined as B = T-W, i.e., ($b_{ij} = t_{ij} - w_{ij}$). The relative size of the elements of B to those in W gives the estimate of how distinct the groups are. These B and W matrices contain all the information of the relationship between the groups and within the groups, respectively. With the help of calculus and other mathematical operations, we can derive a function with the desired properties. Before that, we have to solve simultaneous equations defined by:

$$\sum b_{1i}v_i = \lambda \sum w_{1i}v_i$$

$$\sum b_{2i}v_i = \lambda \sum w_{2i}v_i$$

. .

. .

. . [4]

$$\sum b_{pi}v_i = \lambda \sum w_{pi}v_i \,,$$

where λ (lambda) is a constant called "eigenvalue," the v's are a set of p coefficients, and the b's and w's are known quantities calculated from the sample data. Thus, the objective is to solve the simultaneous equations given by the equation [4] for values of λ and the v's. To get the unique solutions, there should be a condition that the sum of the squared values of the v's must be equal to 1.0. There are a maximum of q unique, nontrivial solutions to these equations. Each solution, which has its own lambda and set of v's, corresponds to one canonical discriminant function. We can use these v coefficients as

the coefficients for the desired discriminant function. After a simple adjustment to the values, we can get the coefficients that give the function more desirable properties. The latter coefficients are the u's from Equation [1], and they are derived as:

$$u_i = v_i \sqrt{n. - g} \qquad \text{and} \quad u_0 = \sum_{i=1}^{p} u_i X_{i..} \qquad [5]$$

By employing the u's, the values of the f's, the "discriminant score," for the data will be in the standard form, that is, the discriminant scores over all the cases will have a mean zero and a within-group standard deviation of one (Klecka, 1982, 18-21).

Linear Regression Analysis

Linear regression, the most widely used statistical technique, is the study of linear, additive relationships between variables. Specifically, there are three types of linear regression. The first is the simple linear regression where we model the linear relationship between two variables, one independent variable and another dependent variable. This simple linear regression model is often written as:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon, \qquad [6]$$

where Y is the dependent variable, β_0 is Y intercept, β_1 is the slope of the regression line, and ε is the random error. Usually, it is assumed that the error ε is normally distributed with $E(\varepsilon) = 0$ and a constant variance $Var(\varepsilon) = \sigma^2$ in the simple linear regression.

The second type of regression is the multiple linear regression which has one dependent variable and more than one independent variables. Let Y be the dependent

variable, whose values we are going to predict, and let X_1, …, X_k be the independent variables from which we are going to predict the value of Y, with the value of variable X_i in period t (or row t), denoted by X_{it}. Then the equation for the predicted value of Y_t is:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \ldots + \beta_k X_{kt} + \varepsilon_i \quad [7]$$

where the betas are constants and the epsilons are independent and identically distributed normal random variables with mean zero. β_0, the intercept of the model, is the expected value of Y when all the X's are zero, and β_i is the regression coefficient of the variable X_i . The betas with the mean and standard deviation of the epsilons are the parameters of the model. Thus, we will get the following equation for predicting Y_t from the corresponding values of the X's:

$$\hat{Y}_t = b_0 + b_1 X_{1t} + b_2 X_{2t} + \ldots + b_k X_{kt} \quad [8]$$

In this formula, it is assumed that the prediction for Y is a linear function of each of the X variables, holding the others fixed, and the contributions of different X variables is additive to the predictions. The constants, b_1, b_2, …, b_k assumed to be the same, are the slopes of the individual straight-line relationship with Y, which are the coefficients of the variables. That is, b_1 is the change in the predicted value of Y per unit change in X_1, if all X-variables are held constant. The additional constant b_0, called the intercept, is the prediction if all the X's are zero. The coefficients and intercept are estimated by least squares, i.e., these are set to the unique values which minimize the sum of squared errors within the sample data to which the model is fitted. It is assumed that the model's

prediction errors are independently and identically normally distributed, the total effect of the X's on the prediction of Y is the sum of their separate effects, all X's have the same variance (homoscedasticity) and are normally distributed, and the unexplained variations of Y are independent random variables (Nau, 2014).

The third type of regression is nonlinear regression, which assumes that the relationship between the dependent variable and independent variable is not linear in the regression parameters, which may be written as:

$$Y = \alpha/(1 + e^{\beta t}) + \varepsilon, \quad [9]$$

where Y is the growth of a organism as a function of period t, α and β are model parameters, and ε is the random error. A nonlinear regression model is more complicated in the estimation of model parameters, model selection, model diagnosis, variable selection, outlier detection and influential observation identification (Yan, 2009). Thus, I am not using this model in my study.

1. Least Squares Estimation for Simple Linear Regression:

The purpose of least squares estimation for the simple linear regression is to estimate the b_0 and b_1 such that the sum of the squared distance from the actual response Y_i and the predicted response $\hat{Y}_i = \beta_0 + \beta_1 X_i$ will be the minimum among all the possible choices of the regression coefficients β_0 and β_1, i. e.,

$$(b_0, b_1) = \arg\min_{(\beta_0,\beta_1)} \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2 \quad [10]$$

The goal of the least squares method is to find parameter estimates by choosing the regression line that is the closest line to all data points (X_i, Y_i). Mathematically, the least squares estimate of the simple linear regression are derived by solving the following system:

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2 = 0 \quad [11]$$

$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2 = 0 \quad [12]$$

Suppose b_0 and b_1 are the solutions of the above systems, the relationship between X and Y can be described by the regression line $\hat{Y} = b_0 + b_1 X$, which is called the fitted regression line. It is more appropriate to solve for b_0 and b_1 using the following model:

$$Y_i = \beta_0^* + \beta_1 (X_i - \bar{X}) + \varepsilon_i,$$

where $\beta_0 = \beta_0^* - \beta_1 \bar{X}$, then the equations [11] and [12] become

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} [Y_i - (\beta_0^* + \beta_1 (X_i - \bar{X}))]^2 = 0$$

$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} [Y_i - (\beta_0^* + \beta_1 (X_i - \bar{X}))]^2 = 0$$

Taking the partial derivatives with respect to β_0 and β_1, we get

$$\sum_{i=1}^{n} [Y_i - (\beta_0^* + \beta_1 (X_i - \bar{X}))] = 0 \quad [13]$$

$$\sum_{i=1}^{n} [Y_i - (\beta_0^* + \beta_1 (X_i - \bar{X}))] (X_i - \bar{X}) = 0 \quad [14]$$

Note that

$$\sum_{i=1}^{n} Y_i = n\beta_0^* + \sum_{i=1}^{n} \beta_1 (X_i - \bar{X})$$

$$n\beta_0^* = \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \beta_1 (X_i - \bar{X})$$

$$\beta_0^* = \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{1}{n}\beta_1[\sum_{i=1}^{n} X_i - n\bar{X}]$$

$$= \frac{1}{n}\sum_{i=1}^{n} Y_i - \beta_1(\bar{X} - \bar{X})$$

Therefore, we have $\beta_0^* = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$, substituting β_0^* by $\bar{Y}$ in equation [14], we have

$$\sum_{i=1}^{n}[Y_i - (\bar{Y} + \beta_1(X_i - \bar{X}))] (X_i - \bar{X}) = 0$$

$$\sum_{i=1}^{n} Y_i (X_i - \bar{X}) - \sum_{i=1}^{n} \bar{Y} (X_i - \bar{X}) - \beta_1 \sum_{i=1}^{n}(X_i - \bar{X})^2 = 0$$

$$\beta_1 \sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} Y_i (X_i - \bar{X}) - \sum_{i=1}^{n} \bar{Y} (X_i - \bar{X})$$

$$\beta_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Therefore,

$$b_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \quad [15]$$

and

$$b_0 = \bar{Y} - b_1 \bar{X} \quad [16]$$

The fitted value of the simple linear regression is $\hat{Y}_i = b_0 + b_1 X_i$, and the difference between Y_i and $\hat{Y}_i$, $e_i = Y_i - \hat{Y}_i$, is called the regression residual. Regression residuals can be calculated from the observed Y_i's and the fitted values $\hat{Y}_i$'s, thus the residuals can be observed, whereas the error, ε_i, in the regression model cannot be observed. Regression error is the difference between the observed value and expected value, the average of the entire population.

2. Multiple Linear Regression:

A multiple linear regression model helps to find out the linear relationship between a dependent variable (Y) and more than one independent variables (X_1, X_2, …, X_k) at the same time. This multiple linear model is usually introduced in matrix form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \mathcal{E} \quad [17]$$

where ε is the normally distributed random error with mean 0 and the constant variance σ^2.

2.1 Least Square Estimation for Multiple Linear Regression

A typical multiple linear regression model is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \mathcal{E}_i,$$

where Y_i is the dependent variable, $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are the regression coefficients, and $\mathcal{E}_i$'s are random errors, which are assumed to be normally distributed with $E(\mathcal{E}_i) = 0$ and $Var(\mathcal{E}_i) = \sigma^2$ for i = 1, 2, 3, …, n. We can also express the multiple linear regression in the form of matrix:

$$Y = X\beta + \varepsilon,$$

where

$$X = \begin{pmatrix} x_{11} & x_{12} \dots & x_{1k} \\ x_{21} & x_{22} \dots & x_{2k} \\ \dots & \dots & \dots \\ x_{n1} & x_{n2} \dots & x_{nk} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \dots \\ \beta_{k-1} \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \dots \\ \varepsilon_n \end{pmatrix} \quad [18]$$

This matrix form of the multiple regression model is a more convenient and effective model.

CHAPTER V

DATA ANALYSIS

Discriminant Analysis

In the previous two chapters, we discussed the assumptions and requirements of the models and derived their mathematical equations and matrices. In this chapter, we are going to perform an analysis of the data using the computer program SPSS.

Table 5.1

Tests of Equality of Group Means

Variables	Wilk's Lamda	F	df1	df2	Sig.
Sex	.911	15.733	4	644	.000
Age	.975	4.143	4	644	.003
Guardian	.979	3.517	4	644	.007
Study Time	.976	3.917	4	644	.004
Family Support	.984	2.606	4	644	.035
Wants to take Higher Education	.980	3.235	4	644	.012
Romantic Relationship	.985	2.428	4	644	.047
Family relationship	.981	3.078	4	644	.016
Free Time	.983	2.740	4	644	.028
Going Out	.939	10.489	4	644	.000
Absences	.968	5.279	4	644	.000
1st period grade	.962	6.406	4	644	.000
2nd period grade	.962	6.440	4	644	.000
Final grade	.952	8.193	4	644	.000

In Table 5.1, if Wilk's Lambda is smaller, the independent variables are more important to the discriminant variables. Wilk's Lambda is significant by a F-test if the significance level is less than 0.05 for the independent variables. All the variables shown in the table are important by this reasoning.

From the SPSS output of correlation matrix, it is also found that most of the independent variables are not related to each other. However, some of the variables, such as father and mother's education are correlated. Similarly, first period grade, second period grade, and final grade are also correlated. (see Appendix A)

Homogeneity of Covariance Matrices

We can test the homogeneity of covariance matrices

Table 5.2

Box's Test of Equality of Covariance Matrices-Log Determinants

Workday Alcohol Consumption	Rank	Log Determinant
1- Very low	4	0.832
2- Low	4	0.878
3- Medium	4	0.472
4- High	4	2.169
5- Very high	4	0.348
Pooled within groups	4	0.922

Table 5.3

Box M Test

Box's M	65.326
F approx.	1.536
df1	40
df2	13372.862
Sig.	0.016

In Table 5.2, if the log determinants of the groups are roughly equal, it is more likely the homogeneity of covariance matrices. In the table 5.3, Box's M test also tests the homogeneity of the covariance matrices. If the significance level is more than 0.05, the groups do not differ significantly in their covariance matrices. But for our data, it is concluded that the groups do differ in their covariance matrices which violates our assumption of homogeneity of variances. However, Discriminant Analysis is robust even though the homogeneity of variances is not met, as long as the data does not contain important outliers. There may be small deviations from homogeneity if the sample size is larger, so Box's M is interpreted with the combination of log determinants. Our sample size is 649, which is large enough to assume the homogeneity of variances.

Table 5.4

Variables in the Analysis

Step	Variables	Tolerance	F to Remove	Wilk's Lambda
1	Sex	1.000	15.733	
2	Sex	1.000	15.149	0.939
	Going out with friends	1.000	9.931	0.911
3	Sex	0.993	13.537	0.899
	Going out with friends	0.998	9.088	0.876
	Final grade	0.992	5.624	0.858
4	Sex	0.979	14.684	0.880
	Going out with friends	0.984	10.091	0.857
	Final grade	0.989	5.031	0.831
	Family relationship	0.970	4.545	0.829

In the stepwise statistics, the independent variables which are important and have correlation with the dependent variable are entered in the model and others are removed. The variables which are entered in the model are sex, going out with friends, final grades,

and quality of family relationships. Tolerance is the proportion of a variable's variance not accounted for by other independent variables in the equations. If it is less than 0.10, it can cause computational problems.

Table 5.5

Wilk's Lambda

Step	Number of variables	Lambda	df1	df2	df3	Statistics	df1	Exact F df2	Sig.
1	1	0.911	1	4	644	15.733	4	644	0.00
2	2	0.858	2	4	644	12.795	8	1286	0.00
3	3	0.829	3	4	644				
4	4	0.806	4	4	644				

Table 5.5 shows that the Wilk's Lambda is significant. That is, as the number of independent variables, which have correlation with the dependent variable, are increased, the value of Wilk's Lambda is decreased on each step showing that function with all four independent variables is the best model.

Table 5.6

Summary of Canonical Discriminant Functions Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative	Canonical Correlation
1	0.213[a]	90.5	90.5	0.419
2	0.015[a]	6.4	96.9	0.122
3	0.007[a]	2.9	99.8	0.083
4	0.000[a]	0.2	100.0	0.019

Note. a. First 4 canonical discriminant functions were used in the analysis.

Table 5.7

Summary of Canonical Discriminant Functions Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	Df	Sig.
1 through 4	0.806	138.727	16	.000
2 through 4	0.978	14.280	9	.113
3 through 4	0.993	4.654	4	.325
4	1.000	.238	1	.625

The first model of discriminant function nearly explains all the variance. The second, third, and fourth models have p-values more than 0.05, so these functions contribute little to the model.

Table 5.8

Standardized Canonical Discriminant Function Coefficients

Variables	Function 1	Function 2	Function 3	Function 4
Sex	.679	.515	-.322	-.438
Family Relation	-.337	.547	.769	-.163
Going out	.582	.094	.196	.794
Final grade	-.360	.616	-.588	.397

This structure matrix in Table 5.8 shows the strength of the correlation of each variable with each discriminant function. As the sex of the students increases (Female = 0, Male = 1), the workday alcohol consumption increases too. Similarly, if the going out with friends is high, the workday alcohol consumption is also high. But there are adverse effects of family relationship and final grade to the alcohol consumption. As the quality of family relationship increases, alcohol consumption by the students decreases and as the final grades of the students increase, there is low consumption of the alcohol on workday.

Table 5.9

Canonical Discriminant Function Coefficients

Variables	Function 1	Function 2	Function 3	Function 4
Sex	1.441	1.092	-.684	-.929
Family Relation	-.354	.576	.810	-.172
Going out	.509	.083	.172	.695
Final grade	-.114	.195	-.186	.126
(Constant)	.535	-5.295	-1.237	-2.652

Now we have the following functions:

Function 1: 0.54 + 1.44 sex – 0.35 family relation + 0.51 going out – 0.11 final grade

Function 2: -5.30 + 1.09 sex + 0.58 family relation + 0.08 going out + 0.20 final grade

Function 3: -1.24 – 0.68 sex + 0.81 family relation + 0.17 going out – 0.19 final grade

Function 4: -2.65 – 0.93 sex – 0.17 family relation + 0.70 going out + 0.13 final grade

Table 5.10

Classification Results

		Predicted group membership					
	Dalc	1	2	3	4	5	Total
Original Count	1	**283**	30	71	36	31	451
	2	42	**8**	26	21	24	121
	3	10	4	**10**	3	16	43
	4	2	1	3	**7**	4	17
	5	3	0	3	0	**11**	17
%	1	**62.7**	6.7	15.7	8.0	283	100.0
	2	34.7	**6.6**	21.5	17.4	19.8	100.0
	3	23.3	9.3	**23.3**	7.0	37.2	100.0
	4	11.8	5.9	17.6	**41.2**	23.5	100.0
	5	17.6	.0	17.6	.0	**64.7**	100.0

Note. 49.2% of the original cases correctly classified.

A discriminant model is created which classifies workday alcohol consumption into one of five groups based on the alcohol consumption amount. The classification results show that out of 451 students, who were categorized as low workday alcohol consumers, 283 are correctly categorized as low consumers of alcohol during the workday, that is, 62.7% of the low-level consumption of alcohol is correctly categorized. Similarly, out of 17 students who are categorized as high-level of alcohol consumers, 11 were correctly categorized as high-level of alcohol consumers during the workday, that is, 64.7% is correctly categorized as high-level of alcohol consumers during the workday.

Linear Regression Analysis

In stepwise regression analysis, the independent variables, such as sex of the student, going out with friends, final grade, number of school absences, student's guardian, quality of family relationships, and age of the student are entered in the analysis and other are removed. This indicates that these seven independent variables have a correlation with the workday alcohol consumption by high school students. Among these seven variables, sex of the student is the single best predictor, going out with friends is the second-best predictor, final grade of the student is the third best predictor and fourth, fifth, sixth, and seventh predictors are the number of school absences, student's guardian, quality of family relationships, and the age of the student, respectively. Likewise, they are included from first to seventh in the model respectively.

Table 5.11

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.283[a]	.080	.078	.888
2	.364[b]	.132	.130	.863
3	.394[c]	.155	.151	.852
4	.417[d]	.174	.169	.843
5	.430[e]	.185	.179	.838
6	.440[f]	.193	.186	.834
7	.446[g]	.199	.190	.832

a. Predictors: (Constant), sex. b. Predictors: (Constant), sex, goout. c. Predictors: (Constant), sex, goout, G3. d. Predictors: (Constant), sex, goout, G3, absences. e. Predictors: (Constant), sex, goout, G3, absences, guardian. f. Predictors: (Constant), sex, goout, G3, absences, guardian, famrel. g. Predictors: (Constant), sex, goout, G3, absences, guardian, famrel, age.

This model summary shows that with sex alone, 8% of the variance was accounted for; with sex and going out with friends together, 13.2% of the variance was accounted for; with sex, going out with friends, and final grade together, 15.5% of the variance was accounted for; with sex, going out with friends, final grade, and number of school absences together, 17.4% of the variance was accounted for; with sex, going out

with friends, final grade, number of school absences, and guardian together, 18.5% of the variance was accounted for; with sex, going out with friends, final grade, number of school absences, guardian, and quality of family relationship together, 19.3% of the variance was accounted for; and with sex, going out with friends, final grade, number of school absences, guardian, quality of family relationships, and age together, 19.9% of the variance was accounted for.

Table 5.12

ANOVA

Model		Sum of Squares	Df	Mean square	F	Sig.
7	Regression	110.098	7	15.728	22.699	.000
	Residual	444.148	641	.693		
	Total	554.247	648			

From the model summary, we find that with Model 7, 19.9% of the variance was accounted for. ANOVA Table 5.10 shows that it has a significant result 0.000 which is less than 0.05.

Table 5.13

Coefficients

Model		Unstandardized B	Coefficients Std. error	Standardized Coefficients Beta	T	Sig.
7	(Constant)	.335	.505		.664	.507
	Sex	.491	.068	.261	7.266	.000
	Goout	.166	.028	.212	5.871	.000
	G3	-.036	.010	-.126	-3.474	.001
	Absences	.022	.007	.113	3.126	.002
	Guardian	.140	.056	.091	2.518	.012
	Famrel	-.091	.035	-.094	-2.607	.009
	Age	.057	.028	.074	2.031	.043

a. Dependent variable Dalc.

Table 5.13 gives beta coefficients so the regression equation can be constructed:

Function 7: $Y = 0.335 + 0.491(sex) + 0.166\ (goout) – 0.36\ (G3) + 0.022\ (absences) + 0.14\ (guardian) – 0.091\ (famrel) + 0.057\ (age)$

From this model, we can conclude that as the sex of the students increases (i.e., from 0 = female to 1 = male), alcohol consumption during the workday increases. Similarly, as the going out time, number of absences in the schools, age, or the guardianship (from 0 = mother, 1 = father to 2 = other) increases, alcohol consumption

during the workday increases. Whereas, as the final grades or quality of family relationship increases, alcohol consumption during the workday decreases.

CHAPTER VI

CONCLUSION

The purpose of this thesis is to determine the effectiveness of the two statistical techniques, discriminant analysis and linear regression analysis, in the prediction of workday alcohol consumption by high school students. We discussed the output of both discriminant analysis and linear regression analysis in Chapter V. When we compared these two techniques, we found some differences in prediction model even though most of the predictor variables were common. First of all, we will discuss the results of both analyses and then the reasons for the differences found.

Results

From the discriminant analysis, it is found that most of the independent variables are not related to each other, however, some of the variables, such as father and mother's education are correlated and the first period grade, the second period grade, and the third period grade are correlated. This technique also provides the homogeneity of the covariance matrices. Whereas, linear regression analysis did not show any relationship between independent variables and we could not test the homogeneity through this technique. However, our sample size is 649, which is large enough to assume the homogeneity of the variances. In the stepwise statistics of both techniques, the independent variables, which are important and have correlation to the dependent variables, are entered in the models and other variables are removed. The variables,

which are entered in the discriminant analysis model, are sex of the students, going out with friends, final grades, and quality of family relationships, whereas, the variables, which entered in the linear regression analysis, are the sex of the students, going out with friends, final grades, number of absences, students' guardian, quality of family relationships, and the age of the students. The model from discriminant analysis explains that as the sex of the students increases (Female = 0, Male = 1), the workday alcohol consumption increases too. Similarly, if the going out with friends is high, the workday alcohol consumption is also high. But there are adverse effects of family relationship and final grade to the alcohol consumption. As the quality of family relationship increases, alcohol consumption by the students decreases and as the final grades of the students increase, there is low consumption of the alcohol during the workday. Whereas, the model from linear regression analysis explains that as the sex of the students increases (i.e., from 0 = female to 1 = male), alcohol consumption during the workday increases. Similarly, as the going out time, number of absences in the schools, age or the guardianship (from 0 = mother, 1 = father to 2 = other) increases, alcohol consumption during the workday increases. Whereas, as the final grades or quality of family relationship increases, alcohol consumption during the workday decreases. Both models from both techniques showed the similar best three predictor variables for the weekday alcohol consumption by high school students, whereas the fourth predictors are different in these two models.

In conclusion, even though there are some differences in the models from two different techniques, we can conclude that both models provided us the similar predictor variables for the weekday alcohol consumption by high school students.

REFERENCES

"Discriminant Function Analysis." *Documents.software.dell.com/statistics/textbook/discriminant-function-analysis*. N.p., 8 May 2015. Web. 25 Apr. 2016. <documents.software.dell.com>.

"Fact Sheets - Underage Drinking." Centers for Disease Control and Prevention, 20 Oct. 2016. Web. 22 May 2016. <https://www.cdc.gov>.

Frost, Jim. "How to Interpret Regression Analysis Results: P-values and Coefficients." *Blog.minitab.com/blog/adventure-in-statistics*. N.p., 1 July 2013. Web. 23 May 2016. <blog.minitab.com>.

Klecka, William R. *Discriminant Analysis*. Beverly Hills, London: Sage Publications, 1982. Print. Quantitative Applications in the Social Sciences, 07-019.

“Linear Regression Analysis using SPSS Statistics.” *https://statistics.laerd.com/spss-tutorials/linear-regression-using-spss-statistics.php*. Web. 24 May 2016. <statistics.laerd.com>.

Nau, Robert. "Notes on Linear Regression Analysis." (2014): n. pag. 26 Nov. 2014. Web. 7 June 2016. <people.duke.cdu>.

Poulsen, John, and Aaron French. "Discriminant Funstion Analysis (DA)." (n.d.): n. pag. *sfsu.edu/efc/classes/biol710/discri/discrim.pdf*. Web. 7 June 2016. <sfsu.edu>.

"SPSS Data Analysis Examples: Discriminant Function Analysis." *http://www.ats.ucla.edu/stat/spss/dae/discrim.htm.* Web. 23 May 2016. <ats.ucla.edu>

Tatsuoka, Maurice M. *The General Linear Model A "New" Trend in Analysis of Variance*. USA: Institute for Personality and Ability Testing, 1975. Print.

Yan, Xin, and Su, Xiao Gang. N.p.: "Linear Regression Analysis Theory and Computing." World Scientific Publishing Co. Pte. Ltd., 2008. *ProQuest Ebrary*. June 2009. Web. 30 June 2016.

APPENDIX A

List of Variables

#	Attributes	Information
1	School	Student's school (binary: 'GP' – Grabiel Pereira or 'MS'-Mousinho da Silveira)
2	Sex	Student's sex (binary: 'F' – female or 'M' – male
3	Age	Student's age (numeric: from 15 to 20)
4	Address	Student's home address (binary: 'U' – urban or 'R' – rural)
5	Famsize	Family size (binary: 'LE3' – less or equal to 3 or 'GT3' – greater than 3)
6	Pstatus	Parent's cohabitation status (binary: 'T' – living together or 'A' – apart)
7	Medu	Mother's education (numeric: 0 – none, 1 – primary education (4^{th} grade), 2 – lower secondary education (5^{th} to 9^{th} grade), 3 – secondary education or 4 – higher education)
8	Fedu	Father's education (numeric: 0 – none, 1 – primary education (4^{th} grade), 2 – lower secondary education (5^{th} to 9^{th} grade), 3 – secondary education or 4 – higher education)
9	Mjob	Mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10	Fjob	Father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11	Reason	Reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12	Guardian	Student's guardian (nominal: 'mother', 'father' or 'other')
13	Traveltime	Home to school travel time (numeric: 1 - < 15 min, 2 – 15 to 30 min, 3 – 30 to 1 hour, or 4 - > 1 hour)

14	Studytime	Weekly study time (numeric: 1 - < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours, 4 - > 10 hours)
15	Failures	Number of past class failures (numeric: n if 1<= n < 3, else 4)
16	Schoolsup	Extra educational support (binary: yes or no)
17	Famsup	Family educational support (binary: yes or no)
18	Paid	Extra paid classes within the course subject (math or Portuguese) (binary: yes or no)
19	Activities	Extra-curricular activities (binary: yes or no)
20	Nursery	Attended nursery school (binary: yes or no)
21	Higher	Wants to take higher education (binary: yes or no)
22	Internet	Internet access at home (binary: yes or no)
23	Romantic	With a romantic relationship (binary: yes or no)
24	Famrel	Quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
25	Freetime	Free time after school (numeric: from 1 – very low to 5 – very high)
26	Goout	Going out with friends (numeric: from 1 – very low to 5 – very high)
27	Dalc	Workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
28	Walc	Weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
29	Health	Current health status (numeric: from 1 – very bad to 5 – very good)
30	Absences	Number of school absences (numeric: from 0 to 93)
31	G1	First period grade (numeric: from 0 to 20)
32	G2	Second period grade (numeric: from 0 to 20)
33	G3	Final grade (numeric: from 0 to 20)

Appendix B

SPSS Output of Correlation Matrix

		sex	Age	famsize	Pstatus	Medu	Fedu	guardian	studytime	schoolsup	Famsup	romantic	famrel	freetime	goout	health	absences	G1	G2	G3
Correlation	Sex	1.000	-.080	.081	-.056	.127	.085	-.013	-.172	-.104	-.129	-.125	.113	.116	-.015	.123	-.033	-.051	-.055	-.080
	Age	-.080	1.000	-.007	.009	-.107	-.122	.160	.005	-.170	-.102	.169	-.016	-.015	.086	-.012	.133	-.153	-.084	-.081
	Famsize	.081	-.007	1.000	.245	-.015	-.041	-.030	.001	-.052	-.034	-.032	.014	-.029	-.021	-.005	-.006	.061	.053	.060
	Pstatus	-.056	.009	.245	1.000	.058	.031	.001	.001	.008	-.010	.059	-.062	-.033	-.020	-.009	.127	-.024	-.030	-.012
	Medu	.127	-.107	-.015	.058	1.000	.648	-.113	.098	-.022	.122	-.031	.026	-.019	.011	.004	-.007	.264	.269	.246
	Fedu	.085	-.122	-.041	.031	.648	1.000	.008	.052	.025	.138	-.065	.019	.006	.028	.044	.030	.222	.229	.215
	Guardian	-.013	.160	-.030	.001	-.113	.008	1.000	.036	.007	.056	.060	-.020	-.013	-.056	-.012	.043	-.024	-.018	.001
	Studytime	-.172	.005	.001	.001	.098	.052	.036	1.000	.083	.139	.037	-.021	-.053	-.042	-.044	-.096	.241	.220	.229
	Schoolsup	-.104	-.170	-.052	.008	-.022	.025	.007	.083	1.000	.072	-.099	-.017	-.011	-.052	.028	-.055	-.080	-.063	-.073
	Famsup	-.129	-.102	-.034	-.010	.122	.138	.056	.139	.072	1.000	-.031	.010	.003	.022	.025	.044	.034	.033	.054
	Romantic	-.125	.169	-.032	.059	-.031	-.065	.060	.037	-.099	-.031	1.000	-.042	.024	-.015	-.016	.072	-.066	-.088	-.077
	Famrel	.113	-.016	.014	-.062	.026	.019	-.020	-.021	-.017	.010	-.042	1.000	.140	.115	.120	-.079	.034	.073	.042
	Freetime	.116	-.015	-.029	-.033	-.019	.006	-.013	-.053	-.011	.003	.024	.140	1.000	.330	.077	-.042	-.075	-.090	-.107
	Gout	-.015	.086	-.021	-.020	.011	.028	-.056	-.042	-.052	.022	-.015	.115	.330	1.000	-.033	.044	-.027	-.035	-.040
	Health	.123	-.012	-.005	-.009	.004	.044	-.012	-.044	.028	.025	-.016	.120	.077	-.033	1.000	-.042	-.040	-.072	-.089
	Absences	-.033	.133	-.006	.127	-.007	.030	.043	-.096	-.055	.044	.072	-.079	-.042	.044	-.042	1.000	-.118	-.097	-.061
	G1	-.051	-.153	.061	-.024	.264	.222	-.024	.241	-.080	.034	-.066	.034	-.075	-.027	-.040	-.118	1.000	.861	.822
	G2	-.055	-.084	.053	-.030	.269	.229	-.018	.220	-.063	.033	-.088	.073	-.090	-.035	-.072	-.097	.861	1.000	.916
	G3	-.080	-.081	.060	-.012	.246	.215	.001	.229	-.073	.054	-.077	.042	-.107	-.040	-.089	-.061	.822	.916	1.000