

**A MULTIVARIATE ADAPTIVE REGRESSION SPLINES APPROACH TO
PREDICT THE TREATMENT OUTCOMES OF TUBERCULOSIS PATIENTS IN
KENYA.**

KIBET CHERUIYOT ERICK

**A RESEARCH PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENT FOR THE AWARD OF THE DEGREE OF MASTERS OF
SCIENCE IN BIOMETRY, SCHOOL OF MATHEMATICS, IN THE
UNIVERSITY OF NAIROBI.**

JULY 2012

Declaration

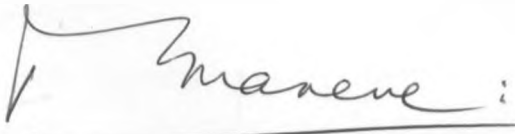
I declare that this research project is my own work. It is being submitted for the degree of Masters of Science in Biometry to The University of Nairobi, Kenya. It has not been submitted before for any degree or examination in this or any other university. Information derived from the published and unpublished work and others have been acknowledged in the text and a list of reference is given.

Signature:  Date 18/07/2012

Kibet Cheruiyot Erick

Student Reg No: I56/63935/2010

This project has been submitted as a partial fulfillment of the requirement for Masters of Science in Biometry of the University Of Nairobi with my approval as the university supervisor.

Signature:  Date: 19/07/2012

PROFESSOR MOSES M. MANENE

Head of Statistics and Operations Research

School of Mathematics

University of Nairobi

Dedication

This research project is dedicated to the Lord God Almighty. It is also dedicated to my parents Wilson and Bornice Bii, for the generous support they have granted me thus far. To the memory of my grandfather Julius Chirchir and my dear elder sister Rose Chepng'etich, I express my highest gratitude which no words can complement for they had shaped my thought processes to always strive for the best.

To my brothers and sisters, and to the friends and/or classmates, whom I sought pieces of advice and support from; THANK YOU for being there when I needed you.

Abstract

Precise and accurate predictive models are extremely important in tuberculosis (TB) treatment outcome modeling. Tuberculosis (TB) treatment with patient supervision and support is one of the elements of global plan to stop TB designed by World Health Organization in 2006 that requires prediction of patient treatment outcome in order to determine how intensive should be the level of supplying services and supports in DOTS (directly-observed treatment, short course). This study was aimed to develop a model using MARS technique to forecast TB cases treatment outcomes. MARS is a relatively new methodology, due to Friedman, for nonlinear regression modeling. MARS can be conceptualized as a generalization of recursive partitioning that uses spline fitting in lieu of other simple functions. Given a set of predictor variables, MARS fits a model in the form of an expansion in product spline basis functions of predictors chosen during a forward and backward recursive partitioning strategy. MARS produces continuous models for high dimensional data that can have multiple partitions and predictor variable interactions.

The five given outcomes included getting cured, completion of treatment courses, quit the treatment course or out of control, fail in treatment, and death. 16 predictor variables were applied as predictors. The data set with 4,605 Kenyan patients was divided as training to build a model and testing datasets to check the predictive ability of MARS model. Nine (9) variables were identified as important by the MARS algorithm and 8 basis functions were created for model building.

The Predictive model was developed by learning from given historical datasets, based on an MARS algorithm. After applying the developed model by training set, the validation set risk estimate was 0.1725. In conclusion, the good results obtained in this application suggest that the proposed MARS prediction model is highly reasonable, desirable and effective in producing a valid and transparent intelligent exploratory predictive model in predicting multiple response variables. To support TB patients actively, this valid model can support health workers to realize how intensive their follow up should be in frame of DOTS.

Acknowledgements

First and foremost I am very grateful to God the Almighty for the blessings and guidance He has bestowed upon me throughout the entire period of my masters undertaking.

I would like to express my greatest appreciation to Prof. Moses M. Manene for his careful supervision, guidance and invaluable support to my project work. It is my great fortune having him as my supervisor. I acknowledge the DLTLD for their data used in this study. I wish to acknowledge Dr. Bernard Langat of the DLTD in the programme of tuberculosis control for his technical advice and suggestions in health science and Tuberculosis control.

My deepest appreciation goes to my dear colleague, mentor and a wonderful friend in Nicasio Karani Migwi for his unmatched intelligence, encouragement and advice during trying moments.

Finally, I would thank all my family members for their support; their love is the most precious gift in my life.

Abbreviations and Acronyms

TB= Tuberculosis

MDR-TB=Multi-Drug Resistant-Tuberculosis

XDR-TB= Extensive Drug Resistant-Tuberculosis

HIV=Human Immunodeficiency Virus

AIDS=Acquired Immunodeficiency Syndrome

WHO=World Health Organization

DOTs= Directly Observed Therapy Support

DLTD= Division of Leprosy Tuberculosis and Lung Disease

LTBI= Latent Tuberculosis Infection

MDGs= Millennium Development Goals

PPMDOTS = Public-Private Mix for DOTS

ROC= Receiver Operating Characteristic

AUC= Area Under the Curve

MARS= Multivariate Adaptive regression Splines

GCV= Generalized cross validation

CART= Classification and Regressions Trees

Table of Contents

Declaration.....	ii
Dedication	iii
Abstract.....	iv
Acknowledgements	v
Abbreviations and Acronyms	vi
Table of Contents	vii
List of Tables	ix
List of Figures.....	x

Chapter 1

1.1 Introduction	1
1.2 Background	2
1.2.1 Predictive analytics and MARS	2
1.2.2 Tuberculosis	5
1.3 Problem Statement.....	10
1.4 Objectives of the Study.....	11
1.4.1 Overall Objective	11
1.4.2 Specific Objectives.....	11
1.5 Justification	12
1.6 Research Questions.....	13
1.7 Research Significance and Contributions	14

Chapter 2

2.1 Literature Review	15
2.1.1 Tuberculosis treatment outcome prediction and related works.....	16
2.1.2 Summary	20

Chapter 3

3.1 Methodology	21
3.1.1 Introduction	21
3.1.2 Linear Regression Model.....	24
3.1.3 Generalized Linear Model.....	24
3.1.4 Nonparametric Regression Models and Accuracy Measures	25

3.1.5	Multivariate Adaptive Regression Splines (MARS).....	26
3.1.5.1	Basis functions	28
3.1.5.2	The MARS model	30
3.1.5.3	Hinge functions	32
3.1.5.4	The model building process: MARS approach	32
3.1.5.5	The forward pass	35
3.1.5.6	The backward pass	36
3.1.5.7	Generalized cross validation (GCV)	37
3.2	Advantages and Disadvantages of MARS.	39

Chapter 4

4.1	Data Analysis and Results.....	41
4.1.1	Data	41
4.1.2	Data collection.....	42
4.1.3	Variables.....	43
4.1.3.1	The outcome variables.....	43
4.1.3.2	Exposure variables	44
4.2	Data Processing Methods And Data Analysis.....	47
4.3	Variable selection and importance	48
4.3.1	Three criteria for estimating variable importance in MARS.....	48
4.3.2	Multivariate Model building and analysis.....	50
4.3.2.1	Basis functions	50
4.3.2.2	Treatments outcome models.....	51
4.4	Model Summary	52
4.5	Model Diagnostics.....	53
4.5.1	Model's partial response	54
4.6	Model Accuracy	57
4.6.1	Misclassification Matrix.....	58

Chapter 5

5.1	Conclusions and Recommendations.....	60
5.1.1	Summary	60
5.1.2	Limitations of the Study	61
5.1.3	Plan for Utilization and Dissemination of Results	61
5.1.4	Future Research.....	62

References:	63
-------------	-------	----

Appendix R code	66
-----------------	-------	----

List of Tables

Chapter 4

Table 4.1: Treatment outcomes proportion in the original data	41
Table 4.2: A summary of continuous predictor variables	45
Table 4.3: A summary of categorical predictors	45
Table 4.4: A table of variable importance	49
Table 4.5: A table of basis functions	50
Table 4.6: A table of model summary	52
Table 4.7 Classification Accuracy Matrix.....	58
Table 4.8 Misclassification matrix.....	58

List of figures

Chapter 3

Figure 3.1: The BFs used by MARS $(x-t)_+$ and $(t-x)_+$	29
Figure 3.2: Two-way interaction basis functions	30
Figure 3.3: A mirrored pair of hinge functions with a knot at $x=3.1$	32

Chapter 4

Figure 4.1: A plot of variable importance	49
Figure 4.2 The Cumulative Distribution graph	53
Figure 4.3 The QQ graph	54
Figure 4.4 A plot of how the dead treatment outcome Varies with the predictors	55
Figure 4.5 A plot of how the failed treatment outcome Varies with the predictors	55
Figure 4.6 A plot of how the out of control treatment Outcome varies with the predictors	56
Figure 4.7 A plot of how the transfer out treatment Outcome varies with the predictors	56
Figure 4.8 A plot of how the treatment success Treatment outcome varies with the predictors	57
Figure 4.9 A plot of the distribution of the predicted Values for each class	58

CHAPTER 1

INTRODUCTION

This section introduces the concept of predictive analytics and the tuberculosis (TB) control under the DOTS approach.

Precise and accurate predictive models are very important in screening initiatives. Advances in soft computing and data mining techniques like artificial intelligence in prediction provides an alternative to the existing statistical prediction techniques in disease susceptibility studies (Garibaldi and Ozen, 2007; Catto et al., 2006; Muzio et al., 2005; Abbod et al., 2004; Dreiseitl and Ohno-Machado, 2003; Speight and Hammond, 1998; Speight et al., 1995;).

Many machine learning algorithms that have been developed to analyze medical datasets, have revealed immense information about the underlying datasets and therefore proved indispensable tools for intelligent data analysis (Garibaldi and Ifeachor, 2000; Butcher, 2004; Mendonca, 2004; Liu et al., 2005). However, there are limitations of the scientific foundation of decision making algorithms and unfounded resistance by practitioners that have been identified among the factors that contribute to the slow widespread of such systems (Mendonca, 2004; Goggin et al., 2007). Therefore, some criteria is reckoned important in order for computer technology and the associated robust statistical techniques so as to be acceptable and useful in medical diagnosis and prognosis; such studies and models must depict good performance as measured by accuracy, sensitivity and specificity; transparency in explaining decisions achieved; and ability to work with small sample and ambiguous data (Dreiseitl and Binder, 2005; Kononenko, 2000; Bradley, 1996).

BACKGROUND

PREDICTIVE ANALYTICS AND MARS

Predictive modeling is a process that deals with extracting information from data and using it to make forecast about future trends and behavior patterns, in this process multiple predictors are combined into a predictive model, which, when subjected to analysis, can be used to estimate future probabilities with an acceptable level of reliability. Data mining is a component of predictive analytics that entails analysis of data to identify trends, patterns, or relationships among the data. This information can then be used to develop a predictive model. Predictive analytics, along with most predictive models and data mining techniques, rely on increasingly sophisticated statistical methods, including multivariate analysis techniques such as advanced regression or time series models. Such techniques enable institutions to determine trends and relationships that may not be readily apparent, but still enable it to better predict future events or behaviors. One can think of data mining as gathering knowledge about relationships, and the resulting predictive analytics models as applying that knowledge. Data mining catalogs all relationships or correlations that may be found among data, regardless of what causes that relationship. The strength areas of predictive modeling are: (a) an ability to incorporate any type of variable into the analysis, (b) Dynamic, as they can easily accommodate any information as they become available to adjust the model accordingly. Regression models are the backbone of predictive modeling. The goal of regression, as in many competing techniques, is to model the relationship between predictor variables and the desired outcome variables so that in the future, when the outcome variable is unknown, it can be estimated or predicted. Therefore, regression is the process of establishing a mathematical model as a function to represent the relation between the different predictor variables and the expected outcome. The method of arriving at the mathematical formulation depends on the structural assumptions of the relationship between predictors and expected outcome, as well as distributional assumptions regarding the outcome variable. A predictive model captures relationships between explanatory variables (predictors) and the predicted variables (dependent variables) from past occurrences, and exploits it to predict future outcomes. Predictive modeling has a wide application for example in marketing, a customer's gender, age, and purchase history might predict the likelihood of a future sale, other applications include customer

relationship management (CRM), meteorology, credit scoring, insurance, healthcare and pharmaceuticals.

Successful modeling of a complex data is part science, part statistical methods, and part experience and common sense and thus modeling is part a science and an art. Multivariate analysis encompasses all methods that simultaneously analyse multiple measurements on each individual or object under investigation. It is an appropriate method of analysis when the research problem involves a single dependent variable that is presumed to be related to one or more independent variables. The objective is to predict the changes in the dependent variable using the changes in the independent variable.

Predictive analytics and data mining is concerned with constructing statistical models from historical data. These models are applied to predict future unknown data values, and/or to help gain an insight of the predictive relationships inherent in the data. The measured variables (denoted by y) is designated as the one to be predicted, given future values of the other variables denoted by $x = \{x_1, x_2, \dots, x_n\}$. Depending on the field of study, y is referred to as the response variable in Statistics, output variable in neural networks, or concept in Machine Learning techniques. The x -variables are referred to as predictor variables, input variables, or attributes in these respective fields.

The data base consists of a collection of N previously solved cases;

$$(y_i, x_{i1}, \dots, x_{in}) \quad i = 1, \dots, N \quad (1.1)$$

The predictive model takes the abstract form:

$$\hat{y} = \hat{f}(x_1, \dots, x_n), \quad (1.2)$$

where \hat{f} is a prediction rule that maps a set of predictor variable values to a response value. The goal is to use the data to produce an accurate mapping. The notion of accuracy depends on the type of the response variable y in terms of the set of values it can assume. If y assumes numeric values the problem is known as regression and lack of accuracy is defined in terms of a distance

measure $d(y, \hat{y})$ between the predicted value \hat{y} and the unknown true value y . Common measures of inaccuracy are average absolute error:

$$AAE = \text{ave} |y - \hat{y}| \quad (1.3)$$

or root mean squared error

$$RSME = \sqrt{\text{ave}(y - \hat{y})^2} \quad (1.4)$$

Here *ave* represents the average over future values to be predicted. If y assumes unorderable categorical values (class labels), the problem is called classification. In this case inaccuracy is generally defined to be the fraction of incorrect future predictions (error rate).

There is an ever-growing number of high dimensional, super large databases across different industries that require effective analysis techniques to mine interesting information from such data. This project work applies a flexible Multivariate Adaptive Regression Splines (MARS) method for modeling of high dimensional data to model the treatment outcomes of tuberculosis patients under the directly observed therapy support (DOTS) strategy. The MARS model takes the form of an expansion in product spline basis functions, where the number of basis functions as well as the parameters associated with each one (product degree and knot locations) are automatically determined by the data. This procedure is an improvement of the recursive partitioning approach to regression since unlike recursive partitioning; it produces continuous models with continuous derivatives. MARS is a generalization of the recursive partitioning regression strategy, or the additive modeling approach of Friedman and Silverman (1989). An adaptive computation is one that dynamically adjusts its strategy to take into account the behaviour of the particular problem to be solved, e.g. the behaviour of the function to be approximated. Adaptive algorithms have been in long use in numerical quadrature [Lyness (1970); Friedman and Wright (1981).] In statistics, adaptive algorithms for function approximation have been developed based on two paradigms, recursive partitioning [Morgan and Sonquist (1963), Breiman, et al. (1984)], and projection pursuit [Friedman and Stuetzle (1981), Friedman, Grosse, and Stuetzle (1983), and Friedman, (1985)].

Below is a very simple example of a model predicting annual income (I) using the categorical variables Education Level (E) and Region (R) and the continuous predictor Age (A):

$$I = 10.0 + 0.5 \cdot A + [-5.0|R=rural] + [5.0|R=urban \text{ and } E=H.S.] + [10.0|R=urban \text{ and } E>H.S.]$$

In this example, a 30 year old rural resident with a high school education would have a predicted annual income of \$20,000 ($10+.5*30-5$) a year, while a 50 year old urbanite with a college degree would have a predicted annual income of \$45,000 ($10+.5*50+10$). Note that the first two terms are applied to the entire population, while the last three terms are applied only to specific regions of the data. The relationship between Region and Education Level, where the weights differ depending on the level of the two variables, is an example of an interaction.

TUBERCULOSIS

A modern medical dictionary defines tuberculosis as: "a specific disease caused by the presence of *Mycobacterium tuberculosis*, which may affect almost any tissue or organ of the body, the most common seat of the disease being the lungs; the anatomical lesion is the tubercle....; local symptoms vary according to the part affected; general symptoms are those of sepsis: hectic fever, sweats, and emaciation; often progressive with high mortality if not treated." [1]. TB is spread through the air from one person to another when a person with active TB disease of the lungs or throat coughs, sneezes, speaks, or sings. People nearby may breathe in these bacteria and become infected.

Further the disease has been defined as a **disease of poverty** affecting mostly young adults in their most productive years, 95% of TB deaths occur in the developing world. TB bacteria can live in the body without making a person sick. This is called **latent TB infection (LTBI)**. In most people who breathe in TB bacteria and become infected, the body is able to fight the bacteria to stop it from growing. People with latent TB infection do not feel sick and do not have any symptoms of the disease, the TB bacteria become active if the immune system can't stop it from growing. When TB bacteria are **active** i.e. multiplying in the body, this is called **active TB disease**. The sign of TB infection is a positive reaction to the tuberculin skin test or special TB blood test.

Kenya has a large and rising TB disease burden and is ranked 15th among the 22 countries that collectively share about 80 percent of the world's TB cases. In 1991, the 44th World Health Assembly set two key targets for global tuberculosis (TB) control to be reached by the year 2000: 70% case detection of acid-fast bacilli smear-positive TB patients under the DOTS strategy recommended by WHO and 85% treatment success of those detected [2]. Studies reported that WHO's target of treatment success is achievable in people with smear-positive TB even in under-resourced developing countries.

In 1993, the World Health Organization (WHO) adopted a national case management strategy (DOTS) to reduce the increasing global burden of tuberculosis (TB), especially in developing countries. The five elements of the DOTS strategy are:

- (1) Sustainable government commitment.
- (2) Quality assurance of sputum microscopy.
- (3) Standardized short-course treatment (including direct observation of therapy).
- (4) Regular supply of drugs and,
- (5) Establishment of reporting and recording systems [3].

The DOTS policy which is a key policy in TB management, has been highly successful in terms of national alignment: in 2008, 202 countries had reported implementing the strategy. This policy evolved into the STOP TB policy in 2006, in response to indications that the DOTS strategy alone was not sufficient to achieve the 2015 TB-related MDGs. However, the Stop TB policy re-emphasises the importance and central position of DOTS. The general aim of the Stop TB strategy is to dramatically reduce the global burden of TB which, it suggests, can be achieved through the following six initiatives [4]:

- (1) Pursue high-quality DOTS expansion and enhancement;
- (2) Address TB /HIV, MDR TB, and the needs of poor and vulnerable populations;
- (3) Contribute to health system strengthening based on primary health care;

- (4) Engage all care providers;
- (5) Empower people with TB, and communities through partnership; and
- (6) Enable and promote research.

As from the foregoing, the Stop TB policy has expanded its focus from the DOTS policy by including the wider drivers of the TB epidemic. However, the Stop TB policy retains DOTS expansion as one of its goals.

Despite optimal efforts invested in curbing the menace in Kenya, TB remains a major cause of morbidity and mortality affecting all age groups, but has its greatest toll in the most productive age group of 15 to 44 years. The major factor responsible for the large TB disease burden in Kenya is the concurrent HIV epidemic. Other factors that have contributed to this large TB disease burden include poverty and social deprivation that has led to a mushrooming of peri-urban slums, congestion in prisons and limited access to general health care services.

With a case notification rate of 338 per 100, 000 population in 2006 to 28 per 100, 000 in 2009, Kenya now ranks at position five in Africa according to facts released by WHO as at March 2012. Through the DLTLD, Kenya adopted the Directly Observed Therapy Short Course (DOTS) strategy for the control of TB in 1993 and achieved countrywide geographic DOTS coverage by 1997. The Division of Leprosy, Tuberculosis and Lung Disease (DLTLD) is implementing initiatives towards achieving internationally agreed TB control targets including the TB relevant Millennium Development Goals (MDGs). The TB MDGs are, to have halted and begin to reverse the incidence and mortality due to TB by 2015. Therefore, DLTLD, in line with international trends, has launched several new approaches to increase access to DOTS and truly expand population DOTS coverage. These approaches include community based DOTS (CB-DOTS), Public-Private Mix for DOTS (PPMDOTS), collaboration between TB and HIV control programs and the development of an elaborate advocacy, communication and social mobilization strategy aimed at influencing communities to seek care early when TB symptoms occur and to remain on treatment until this is completed when treatment is initiated. The initiatives have further been complicated by the growing resistance to TB medications i.e. the emergence of "extensively drug-resistant" tuberculosis (XDR-TB). In Kenya, So far, there are more than 334 cases of

MDR-TB cases on treatment in 101 treatment sites scattered in the country. It was reported that two patients lost their lives due to XDR-TB, in 2010 and 2011 respectively. In addition, Kenya is currently treating its third case of extensively drug resistance TB having lost the previous two cases. The 57% of MDR cases have the potential risk of infecting 2,107 people in one year at the rate of 7 infections per year per one MDR untreated patient. However, the Ministry of Public Health & Sanitation with support from donor funding, total health budget, private (including households) and public sector resources, have put efforts to ensure that they reduce the incidence and mortality due to TB by 2015 and work towards eradicating TB by 2050 as outlined in the Millennium Development Goals (MDGs). The Ministry of Health in partnership with the World Bank is also decentralising TB culture services from the only laboratory in Nairobi based at the Kenya Medical Research Institute for the Multi Drug Resistant (MDR) Tuberculosis to ease congestion at the central TB testing unit to five additional laboratories scattered throughout the country.

Approximately one-third of people in the world today are infected with tuberculosis (latent TB), and nine million people develop tuberculosis disease (active TB) every year, with 90% of all active TB cases occurring in developing countries. This implies that more than 8 million people become sick with TB each year. Other facts alludes that someone gets sick from TB every 4 seconds and Someone also dies of TB every 10 seconds, TB worldwide kills more youth and adults than any other infectious disease. The world health organization's prioritized plan for tuberculosis control in 2006 through patient supervision and support requires prediction of patient treatment course destination to determine how intensive should be the level of supplying services and supports in DOTS (directly-observed treatment, short course), this seeks to improve the DOTS quality. Consequently, there needs to be developed a tool to predict the patients' treatment course destination to spot high risk cases for non-compliance which can assist DOTS more progressively. In this regard, a valid regression and or classification model to predict the treatment outcome in TB treatment course will be usable to determine the level of patients' supervision and support.

Advancements in technology have led to a wide range of genetic and biological markers that hold great potential in improving the prediction of treatment destination outcomes. Although

such new markers promise better disease prognosis, the accuracy in identifying short term and long term outcomes remains unsatisfactory for most complex diseases. It has often been argued that short term clinical outcomes may have potential in predicting long term outcomes. Ensuring that the TB patient completes therapy to cure in order to prevent drug resistance cases and developing TB in the community is one of the crucial objectives of DOTS.

Some controversies have been raised in the past about the application of DOTS to control TB in practice. This came about as a result of the rising incidences of TB starting from mid-1980s thus it seemed that the strategy was failing practically in many countries. It has also been recorded that DOTS imposes extra burdens on the patient and health care system in lengthened admission and frequent attendance at clinics at the expense of self-administration with suitable cure rate in some other cases. On the other hand, other empirical investigations have confirmed the DOTS' role in treatment success rather than case detection. In overall, it seems that DOTS as one of the most widely-implemented and longest running global health intervention in health history is going to continue as a foundation strategy for TB control. However, because of pointed imperfections in practice, it needs some additional change and support to promote the quality of treatment and gain the defined objectives. Hence, WHO in "Stop TB Strategy" has focused on pursuing high-quality DOTS expansion and enhancement; one of the most crucial components of this worldwide plan is standardized treatment, with supervision and patient support. It has been emphasized that services for TB care should identify and address factors that may make patients interrupt or stop treatment. Moreover, supervision must be carried out in a context-specific and patient-sensitive manner, and is designed to ensure adherence on the part both of providers (in giving proper care and support) and of patients (in taking regular treatment). Also, it has been brought to light that preferred patient groups, for example prisoners, drug users, and affected people by mental health disorders may need intensive support including DOTS [5].

PROBLEM STATEMENT

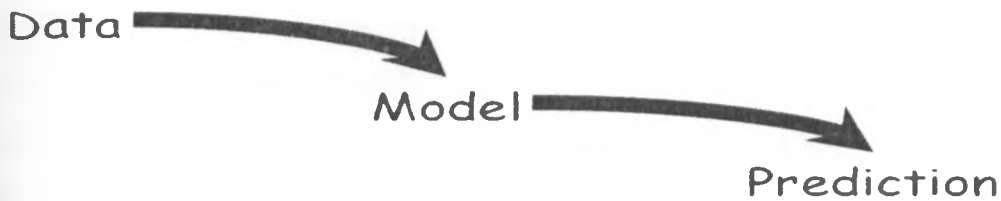
Many health institution information systems are designed to support services like patient billing, inventory management and generation of basic statistics. Despite the fact that some hospitals use decision support systems, they are to a large extent limited. These systems are designed to provide simple pieces of information like “What is the average age of patients who have a certain disease?”, “How many patients were admitted in the hospital more than once in a year or for a particular period?”, “Identify the number of patients from a particular gender, their marital status and age, and summaries of patients treated for a particular ailment.” However, situations may demand the need for complex queries like to identify the most important causal agents that prolong the patient’s stay in hospital. Which treatment combination suits a particular patient and to some extent predict the probability of patients getting a disease or of a disease that will advance to another stage.

The all-important question that this study seeks to answer is how to help healthcare practitioners in combating the TB menace to make better informed clinical decisions on how to turn knowledge-rich data into some useful information in forecasting the treatment course destination of tuberculosis (TB) patients. This will be achieved by making use of historical patients’ attributes to build a predictive model using MARS and validate the same using real data for patients in Kenya. By doing this, health institutions can make informed decisions relating to the optimal use of health resources and implementation of disease-specific intervention strategies.

OBJECTIVES OF THE STUDY

OVERALL OBJECTIVE

The overall objective of this study is to construct and validate a comprehensive clinical tool using multivariate adaptive regression splines to forecast TB cases treatment course destination by a vector of covariates.



SPECIFIC OBJECTIVES

The specific objectives of this study include:

- i. To measure TB treatment outcomes defined as getting cured, completion treatment courses, quit the treatment course, fail in treatment, and death among patients receiving TB treatment.
- ii. To identify tuberculosis (TB) risk predictive factors that contribute to the treatment outcomes defined in (i) above, that is, to generate an 'optimal' input variable predictor set by utilizing the goodness of fit measurement indices for the proposed model
- i. To investigate the feasibility of applying multivariate adaptive regression splines (MARS) algorithm in predicting tuberculosis (TB) course destination under the DOTS strategy.

JUSTIFICATION

Many healthcare organizations including hospitals and medical centers are challenged majorly by inadequate information in the provision of quality healthcare services at affordable costs. While at the same time the industry collects huge amounts of data which needs to be well utilized to discover vital information for effective decision making. The discovery of hidden patterns and relationships needs to be fully exploited by use of appropriate data mining and analysis techniques. Poor clinical decisions can lead to disastrous consequences which can have adverse unethical and legal effects. Health institutions must also minimize the cost of clinical tests. They can achieve these results by employing appropriate computer-based information and/or decision support systems.

Most hospitals today employ some sort of hospital information systems to manage their healthcare or patient data. The systems typically hold huge amounts of data in the form of numbers, text, charts and images. Unfortunately, these data are rarely used to support clinical decision making. There is a wealth of hidden information in these data that is largely underutilized.

The world health organization's prioritized plan for tuberculosis control in 2006 through patient supervision and support requires prediction of patient treatment course destination to determine how intensive should be the level of supplying services and supports in DOTS (directly-observed treatment, short course), this seeks to improve the DOTS quality. Correct treatment of tuberculosis (TB) aims at curing the individual patient, interrupting transmission of TB to other persons, and preventing bacilli from becoming drug resistant. Monitoring the treatment outcome of TB is essential in order to evaluate the effectiveness and efficiency of TB intervention programs such as treatment method, procedure, protocol, including diagnosis and follow up. However, these aims have not been achieved in many regions of the world even when anti-tuberculosis drugs are available. The main reasons are death of the patients during treatment, default before the scheduled end of treatment or resistance to the drugs prescribed. Patient non-adherence to treatment is interpreted as a failure of the health care system to cope with the natural tendency of humans to quit treatment as soon as they feel subjectively better, or better without treatment if adverse drug events supervene [6].

The foregoing situation necessitates the need to develop a tool to predict the patients' treatment course destination to spot high risk cases for non-compliance which can assist healthcare professionals in achieving DOTS more progressively. In this regard, a valid regression model to predict the treatment outcome in TB treatment course might be usable to determine the level of patients' supervision and support. The applied methodology Mars adapts to its data in a splines-like way by fitting piecewise linear basis functions of the explanatory variables that best predict the dependent variable. Together, the many fitted pieces often resemble a polynomial curve. An initial forward pass in the Mars algorithm chooses the predictors and accompanying basis or hinge functions so as maximize the reduction in the sum-of-squares residual error. The process continues until a given number of terms is reached or the residual reduction is not meaningful. A major advantage of Mars is the routinized and tunable handling of curvilinearity and interactions among predictors, both of which are challenges with parametric linear regression. At the same time, Mars is flexible in providing modelers the option to enter predictors linearly, in which case the results look a lot like least squares regression.

RESEARCH QUESTIONS

- (1) Can a multiple statistical technique such as multivariate adaptive regression splines analysis be successfully integrated and applied to problem such as the prediction of a treatment course destination degree?

This question will be examined by fitting the model and determining if the predictive results of the model achieve acceptable accuracy.

- (2) What are the TB treatment outcomes and factors associated with TB "negative" treatment outcomes in persons receiving TB treatment in various parts of Kenya?

To be explored by determining the set of variables that were measured for a group of patients as they progressed in treatment that affect the probability that a given patient fails to complete the treatment therapy or any of the other "negative" treatment outcomes.

RESEARCH SIGNIFICANCE AND CONTRIBUTIONS

This research is the first attempt to use MARS to forecast the tuberculosis (TB) course destination or treatment outcome to a Kenyan database and to a large extent globally at least as per the comprehensive literature review done. The main contributions of this work are both conceptual and practical. The conceptual contribution refers to the development of an adaptive algorithm for the prediction of multiple response variable based on MARS technique. This adapted algorithm is suitable to be used for small and large sample sizes and for variables governed by ambiguous relationship. The practical contribution of this research work refers to the possibility of using the model as a prediction tool that can be further developed into an automated technique for TB detection and screening purposes. The models may be utilized to predict the treatment course destination with modest discriminatory accuracy, similar to that of other tuberculosis prediction models. The statistically significant variables in the models can be assessed by patient interview and probably clinical examination.

The findings of this research will also provide some initial parameters for forecasting TB disease susceptibility at individual and group levels. These findings will help clinician device better screening procedures by differentiating between the high-risk and the low-risk group. Similarly, knowledge on the impact of risk factors on disease treatment outcome as revealed by the proposed MARS prediction model will provide guidelines for patient counselling and health quality improvement efforts. It is hoped that the proposed MARS prediction model may serve as an aid in screening procedure applicable in reducing morbidity and mortality from TB by early detection.

CHAPTER 2

LITERATURE REVIEW

Most data found in the real-world problems are finite or discrete, in such a situation, the regression models used for analyzing such data sets to disclose the relationship between the predictors and response variable(s) are also called discrete or Gaussian approximation. Regression analysis is the most widely used statistical technique, in investigating and modeling the relationship between variables. There are many regression models used for several purposes such as data description, parameter estimation for learning, prediction and control.

A multivariate modeling is undertaken when an individual sets to relate a set of independent variables X 's to a dependent variable Y by using some form of mathematical model with an expectation to uncover the complex inter-relationships among the many variables. Currently, there are many available statistical techniques for solving multivariate problems. However, these statistical methods are restricted by assumptions about the distribution of the dependent variable and the distribution of the residuals (Miles and Shevlin, 2003). When these assumptions are not satisfied, the conclusion derived from the statistical analysis will be incorrect.

Soft computing and data mining provides an alternative to statistical multivariate modeling when variables distributions problems arise. Literature is rich and has shown that there is an increasing trend in the use of soft computing in diseases diagnostic and prognostic prediction modeling (Abdul Kareem, 2002; Garibaldi and Ozen, 2007; Dreiseitl and Ohno-Machado, 2002; Speight and Hammond, 1998; Catto et al, 2006; Liu et al, 2005; Matheny and Ohno-Machado; 2007).

The goal of predictive modeling in clinical medicine is to derive models that can be used to predict the outcome of interest thus supporting clinical decision making in prognosis, diagnosis or treatment planning based on patient-specific information (Bellazi and Zupan, 2008). Commonly used predictive techniques include decision trees, logistic regression, artificial neural networks, k-nearest neighbors and support vector machine.

TUBERCULOSIS TREATMENT OUTCOME PREDICTION AND RELATED WORKS

There are areas related to the above techniques in which significant progress has been made and this includes: The developing of mathematical models of disease transmission within human populations that has been acknowledged in helping policy makers and epidemiologists to interpret epidemiological trends, understand the dynamics of disease spread and to measure the efficiency of disease prevention and control, such as measles, HIV and other emerging infections [7].

In risk prediction, the context of cardio-vascular diseases is most well developed, for which prediction models use a combination of variables (blood pressure, smoking history, lipid levels, and family history of heart disease) to assess an individual's risk of heart disease [8]. Risk data from the Framingham Heart Study have been used to construct the Framingham Coronary Risk Prediction Model and to formulate guidelines for cholesterol-lowering therapy As Grundy et al. note, the Framingham risk scores can both motivate and reassure the patient; they also illustrate the cumulative nature of multiple risk factors [9].

Developing a model to predict between a positive outcome and any another outcome for a given patient involves using data to discriminate between the two potential results. A valuable tool in assessing the accuracy of the discrimination is Receiver Operating Characteristics (ROC) curve [10], analysis. ROC curve analysis was developed as a concept in signal detection theory during World War II where radar operators examined radar signals to detect oncoming Japanese aircraft and distinguish such readings from "noise" in the signal. The goal was to increase the accuracy of predictions and decrease the likelihood of false alarms or missed detections. The prediction accuracy is a trade-off between sensitivity and specificity. Sensitivity is the probability of correctly identifying a signal while specificity is the probability of correctly identifying system "noise." This type of analysis has been widely adapted in other fields to evaluate how well models discriminate between potential outcomes. In the medical research, the concept has been used to evaluate the diagnostic value of medical tests as well as to determine the therapeutic value of treatments [11]. A medical test may result in concluding that a disease is present i.e. a "positive" test result, or that it is not present i.e. a "negative" test result. Ideally, a diagnostic test should accurately detect when a disease is present and accurately indicate when it is not. False positive

test results lead to unwarranted concern and potentially unnecessary treatment while false negative test results may lead to adverse health results as a condition goes untreated.

Discriminating between two outcomes leads to four possible results. The test could classify a result as positive or negative. The classification could be correct or incorrect. The four possible results are correct disease detection (true positive), incorrect disease detection (false positive), correct healthy status (true negative), and incorrect healthy status (false negative). The sensitivity of a diagnostic test measures its ability to identify the presence of disease and the specificity of the test measures its ability to identify the absence of disease. In this context, sensitivity is the probability of a true positive while specificity is the probability of a true negative. The cutoff point used in the outcome discrimination determines the sensitivity vs. (1 - specificity) for the diagnostic test. The two measures are directly associated for a given cut off point value. This association means there is a tradeoff between achieving good sensitivity and good specificity in outcome discrimination.

ROC curve analysis is used in predictive modeling to evaluate models by examining the resulting ROC curve for a wide range of cut off points. The models that result in an ideally shaped ROC curve have a better ability to discriminate between two potential outcomes than those with a flatter ROC curve. A shallow ROC curve implies that the model has negligible discrimination power. Such a model is as likely to predict a true positive as a false positive and has no useful predictive ability. The area under the ROC curve ("AUC") provides an estimate of the model's predictive ability. A model with a high value for AUC is judged to better discriminate between the potential outcomes. Hosmer and Lemeshow [12], indicate that a value for AUC of 0.5 indicates that the model is of little use since it is as likely to correctly predict a binary outcome as flipping a fair coin. A result of $0.7 \leq \text{AUC} < 0.8$ represents "acceptable" ability to discriminate between potential outcomes. A result of $0.8 \leq \text{AUC} < 0.9$ represents "excellent" ability to discriminate between potential outcomes. A result of $\text{AUC} \geq 0.9$ represents "outstanding" ability to discriminate between potential outcomes.

Some modeling studies enrich the basic model framework with heterogeneities by subdividing subgroups, so as to generate greater and more realistic structure, [Grenfell et al (2001), K. M. Hassmiller (2010), Osgood et al (2011)]. Such disaggregation which uses the technique of

attribute-based disaggregation [N.D. Osgood (2004)], can be used to stratify a model to reflect more complex hierarchy of population or to integrate personal characteristics to generate rich dynamical behaviors. The Tuberculosis (TB) model by [Osgood et al (2011)] integrated transmission preference to express the mixing preferences of different age groups and ethnic categories. In addition, Hassmiller's TB model [K. M. Hassmiller (2007)] with smoking impact evaluation also stratified people into subgroups regarding smoking status and apply specific mixing patterns.

In 2001, in order to reduce the risk of developing active TB especially among those within their first two years of infection and who are at particular risk of progression to Active TB, a mathematical model of the TB epidemic was developed to quantify the effectiveness of treatment for early latent TB infection [13]. Positive effects have been observed in lowering the incidence of TB and eliminating the disease especially in latent tuberculosis, which suggests that targeted preventive therapy for newly infected contacts through may ultimately offer great contribution in TB control. Another TB model including preventive treatment for Latent TB infection produces similar outcome and confirms the effectiveness of contact tracing in decreasing the incidence rate of TB [14].

Recently, Osgood and Mohamoud et al.[15] have extended their aggregate TB models by incorporating age as well as ethnic stratification to fit TB data from the Canadian province of Saskatchewan and to investigate targeted intervention strategies for high risk subgroups and their impact on lifetime TB outcome. It is observed that a temporary elevation in incidence rate can bring notable influence on individuals' lifelong risk of TB, and it indicates the presence of system memory in the form of latently infected population.

In a study done in a rural South Africa to determine tuberculosis (TB) treatment outcomes in adult patients attending a rural HIV clinic, it was found out that the outcomes of TB treatment are improved in HIV-infected persons taking TB treatment under the support of DOT supporter. It was further shown that a twice-weekly directly observed therapy (DOT) for TB in HIV-infected and non-infected persons was effective. After six months of follow-up; 71% of the par-

ticipants were cured, 3% completed treatment without being cured, 2% transferred out and only 2% reported treatment failure. The study concluded that a twice-weekly rifampicin- containing drug regimen given under DOT cures most adherent patients irrespective of HIV status and previous TB treatment history [16]. There is generally no disparity on the outcomes of TB treatment between males and females, except on people at 25 years of age and older [Nsubuga P et al 2002]. Age was further reported to be a predictor of TB mortality among with other factors such as residence in rural area, sputum smear-negative disease, defaulting TB treatment and prolonged symptom duration prior to initial diagnosis [17].

In Uganda a study enrolled a cohort of 105 male and 109 female HIV-infected adults receiving treatment for initial episodes of culture confirmed TB between March 1993 and March 1995. In this study favorable outcomes were defined as cured or alive while unfavorable outcomes were not being cured or dead. At the end of one year of follow-up there was no difference in the likelihood of experiencing a favorable outcome (RR 1.02, 95 CI 0.89-1.17). While differences existed between males and females with HIV-associated TB at baseline, the outcomes at one year after initiation of TB treatment were similar [18].

Modeling studies have employed different statistical techniques to unravel the complexity of interactions between distributions and environmental factors. Those include Generalized Linear Models (GLM; Guisan et al. 1998), especially Logistic Multiple Regression (LMR; Narumalani et al. 1997; Felicísimo et al. 2002); Generalized Additive Models (GAM; Yee & Mitchell 1991) and Classification and Regression Trees. In addition, Guisan & Zimmermann (2000) made a comprehensive review of predictive modelling and noticed the lack of comparative studies in which more than two statistical techniques were applied to the same data set.

Most studies on predictive modelling are based on methodologies that assume a Gaussian relation between response variable and the predictors, and also that the contribution to the response from the interactions among predictors is uniform across their range of values. Both assumptions are unwarranted in most cases (Austin & Cunningham 1981; Austin et al. 1990, 1994). Nevertheless, logistic multiple regressions (LMR) with a quadratic function to represent Gaussian responses have often implied high predictive success. Further problems associated with classical

regression analysis arise when many predictors are used. Such increase in the number of predictors implies an increase greater than exponential in the number of possible regression structures, and the almost inevitable problem with multicollinearity.

To avoid these problems, analysts impose strong model assumptions, forcing the variables to act globally over the response by limiting or eliminating local changes in response or interactions. This strategy can no longer be justified if suitable predictors are likely to act and interact differently on the response variable across their range of values. The search for a model that handles the above problems brings us to explore and seek to improve a relatively new statistical technique employed in data-mining strategies in fields such as chemical engineering, marketing campaigns or weather forecasting: multivariate adaptive regression splines (MARS; Friedman 1991).

SUMMARY

The reviews of literature for this research has been developed based on the importance of accurate, transparent and reliable prediction modeling on systems relying on unclear relationships among variables within a multi faceted databases. Thus based on this literature review I propose a data mining based prediction tool that can be incorporated as a screening aid for predicting a polytomous outcome in disease treatment outcome prediction.

CHAPTER 3

METHODOLOGY

INTRODUCTION

The purpose of the model is to **identify** the significant quantitative and qualitative factors of the patients and to utilize them to accurately project the treatment outcome status. The model uses this set of variables as inputs to multivariate adaptive regression splines. The model's output is a predicted probability that a patient's treatment outcome is with a given multivariate vector is positive. This permits the identification of those patients for whom an intervention could be beneficial in terms of prompting them to comprehensive support program. The model's value is assessed by comparing how accurately the predicted outcomes for a group of patients match the actual results.

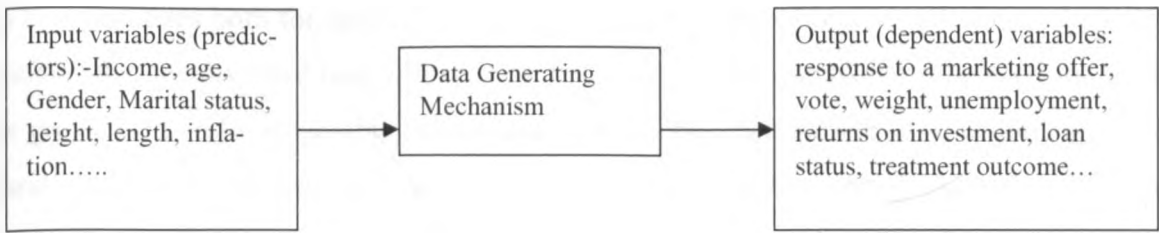
Statistical analysis and modelling involves the application of appropriate statistical analysis techniques that each requires certain assumptions be met so as to perform hypothesis tests, interpret the data, and reach valid conclusions.

Basically, there are two goals when analyzing data:

- (i) Prediction: By analyzing the past, one assumes that conclusions drawn can be used to predict the future.
- (ii) Inferential: In this case, one may be interested to investigate the nature of the relationship between different sides of a complex phenomenon.

We pursue the first goal.

Irrespective of the goals of the analysis, we always have a set of output variables, a set of input variables and an unknown mechanism that relates the output with the input. This mechanism can be called the Data Generating Mechanism:



Discussing briefly about the middle stage i.e. data generating mechanism, statisticians are divided into two main groups:

- Stochastic modelling. One assumes that the relationship between the input and output is driven from a stochastic process: linear regression, logistic, regression, Cox model,...
- The algorithmic approach. According to this approach, the relationship is too complex and unknown. Instead of a concrete equation, this approach looks for an algorithm that can help predict the future of the output from the input: Decision Trees, Neural nets, Genetic models,...

Regardless of what standpoint of the above that one chooses to adopt, statistical analysis and modelling requires careful selection of analytic techniques, verification of assumptions, and verification of data. Descriptive statistics, graphs, and relational plots of the data should first be examined to evaluate the legitimacy of the data, identify possible outliers and assumption violations, and form preliminary ideas on variable relationships for modelling. This project will adhere to the above requirements in this section to the letter but briefly let's look at the origin of MARS.

In recursive partitioning, the main goal is to use the data to simultaneously estimate a good set of sub-regions and the parameters associated with the separate functions in each sub-region. Continuity at sub-region boundaries is not enforced. The partitioning is accomplished through the recursive splitting of previous sub-regions. The starting region is the entire domain. At each stage of the partitioning all existing sub-regions are each optimally split into two sub-regions. The recursive subdivision is continued until a large number of sub-regions are generated. The sub-regions are then recombined in a reverse manner until an optimal set is reached, based on a crite-

tion that penalizes both for lack-of-fit and increasing number of regions (Breiman et al., 1984). Variables that locally have less influence on the response are less likely to be used for splitting. This gives rise to a local variable subset selection. Global variable subset selection emerges as a natural consequence. Recursive partitioning based on linear functions lacks this local variable subset selection feature which tends to limit its power and interpretability. Also, recursive partitioning regression exploits the marginal consequences of interaction effects whereby local intrinsic dependence on several variables, when best approximated by an additive function, does not lead to a constant model.

Unlike regression that returns a subset of variables, classification trees can rank order the factors that affect the retention rate. When recursive partitioning models use piecewise constant approximations they are fairly easily interpretable owing to the fact that they are very simple and can be represented by a binary tree [Breiman et al. (1984)]. They are also fairly especially rapid to evaluate.

However, recursive partitioning as a multivariate function approximation suffers from some severe restrictions that limit its effectiveness. Firstly, the approximating function is discontinuous at the sub-region boundaries. This problem limits the accuracy of the approximation, especially when the true underlying function is continuous. Another problem with recursive partitioning is that certain types of simple functions are difficult to approximate. These include linear functions with more than a few nonzero coefficients. More generally, it has difficulty when the dominant interactions involve a small fraction of the total number of variables. In addition, one cannot discern from the representation of the model whether the approximating function is close to a simple one, such as linear or additive, or whether it involves complex interactions among the variables. To overcome some of the above limitations, we describe the multivariate adaptive regression spline (MARS) approach to multivariate non-parametric regression. Multivariate adaptive regression spline (MARS) denotes a tool from statistics, important in classification and regression, with applicability in many areas of finance, science and technology. It is very useful in high dimensional problems and shows a great promise for fitting nonlinear multivariate functions [19].

LINEAR REGRESSION MODEL

If a regression model is linear in fitted parameters, it is called as linear regression model (LRM). In general, the following equation represents an LRM.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon, \quad (3.1)$$

In the above equation, y is called the response variable or the dependent variable and x_j ($j = 1, 2, \dots, k$) are called the regressor variables (predictor or independent variables). Furthermore, ε is a random error component. The errors are assumed to have a normal distribution with a mean of zero and unknown constant variance, σ^2 . It is also assumed that the errors are uncorrelated. In other words, the value of one error is independent from the value of any other error. The parameter β_0 means the intercept and the other parameters β_j ($j = 1; 2; \dots; k$) are the regression coefficients. The parameter β_1 represents the expected change in the response y per unit change in x_1 when all of the remaining regressor variables x_j ($j = 1; 2; \dots; k; j \neq 1$) are held constant.

To solve most real-world problems, we need to find the values of the regression coefficients β_j and the error variance σ^2 which are always not known. These parameters and the error variance must be estimated from a sample data set. The fitted regression equation or the model enables us to predict future observations of the response variable y . Least squares estimation (LSE) or maximum likelihood estimation (MLE) are two widely used optimization methods applied on the regression model for estimating the unknown regression parameters.

GENERALIZED LINEAR MODEL

We combine both linear and nonlinear regression models under the framework of generalized linear models (GLMs). This approach is used when the assumptions of normality and constant variance are not satisfied. It enables the incorporation of non-normal response distributions. It allows the mean of a dependent variable, y , to depend on a linear predictor through a nonlinear link function and also allows the probability distribution of y , to be any member of an exponential family of distributions. Many widely used statistical models belong to GLMs. These include classical linear models with normal errors, logistic and probit models for binary data, and log-linear models for multinomial data and many other useful statistical models such as the Poisson,

binomial, Gamma, and normal distribution have been formulated as GLMs by the selection of an appropriate link function and response probability distribution.

A GLM has the following basic structure:

$$h(\mu_i) = x_i^T \beta, \quad (3.2)$$

where $\mu_i = E(Y_i)$, h is a smooth monotonic "link function", x_i is the input variable of predictors, and β is a vector of an unknown parameters. A GLM usually imposes the distributional assumption that the response variables Y_i are independent and can have any distribution from exponential family density of the form:

$$Y_i \sim f_{Y_i}(y_i, \theta_i, \phi) = \exp \left\{ \left[\frac{\theta_i y_i - b_i(\theta_i)}{a_i(\phi)} \right] + c_i(y_i; \phi) \right\} \quad (i = 1, 2, \dots, N), \quad (3.3)$$

Where a_i , b_i , and c_i are arbitrary functions, ϕ is an arbitrary "scale" parameter and θ_i is called a natural parameter. We can also obtain a general expression for the mean and variance of dependent variable Y_i using log likelihood of θ_i , $\mu_i = E(Y_i) = b_i'(\theta_i)$ and $\text{Var}(Y_i) = b_i''(\theta_i) a_i(\phi)$. The symbol $'$ is used for differentiation [20].

NONPARAMETRIC REGRESSION MODELS AND ACCURACY MEASURES

A general nonparametric regression model is of the form;

$$y = f(x) + \varepsilon; \quad (3.4)$$

Where $x = (x_1, x_2, \dots, x_k)^T$. The aim of traditional regression analysis is to estimate the parameters of the model, while the aim of nonparametric regression is to estimate the regression function f directly, this function is implicitly assumed to be a generally smooth and a continuous function and in the model the error term ε has zero mean and constant variance σ^2 . However, in some cases it can be non-smooth.

The additive regression model is of the form;

$$y = \beta_0 + f_1(x_1) + f_2(x_2) + \dots + f_k(x_k) + \varepsilon, \quad (3.5)$$

Where β_0 is the unknown bias (intercept) and the partial regression functions f_j ($j=1, 2, \dots, k$) are assumed to be smooth. Both β_0 and the functions f_j ($j=1, 2, \dots, k$) are to be estimated from the data. Variations of the additive regression models are the *semiparametric regression model*, in which predictor variables are "additively" separated by the unknown functions like:

$$y = \beta_0 + \beta_1 x_1 + f_1(x_1) + f_2(x_2) + \dots + f_k(x_k) + \varepsilon, \quad (3.6)$$

or interactions of some predictor variables are expressed in unknown functions that appear as higher-dimensional terms such as:

$$y = \beta_0 + f_{12}(x_1, x_2) + f_1(x_1) + \dots + f_k(x_k) + \varepsilon. \quad (3.7)$$

These models have also been extended to generalized nonparametric regression.

MULTIVARIATE ADAPTIVE REGRESSION SPLINES (MARS)

The Multivariate adaptive regression splines (MARS) is a data mining technique (Friedman, 1991; Hastie, Tibshirani and Friedman, 2001) that can be used for solving regression-type problems. It is a non-parametric procedure, for fitting adaptive regressions that uses piecewise basis functions to define relationships between a dependent variable and a set of predictors and thus no functional relationship between the dependent and independent variables is assumed prior to the analysis. The 1988 brain child of Jerome Friedman, MARS combines properties of regression and tree techniques. Like regression, MARS attempts to optimize a fit of a dependent variable using the least squares method. Unlike regression, MARS allows for the specification of more complex terms than linear and additive ones in the model. Like trees, MARS partitions data, but unlike trees, MARS allows for the capture of linear and additive relationships and for the splitting over all nodes at each step, rather than just the currently terminal ones. Either Categorical or continuous outcomes can be modeled using categorical or continuous predictors with this technique. Therefore, the MARS model splits the data into several splines on an equivalent interval

basis (Friedman, 1991). In every spline, MARS splits the data further into many subgroups. Several knots are created by MARS. These knots can be located between different input variables or different intervals in the same input variable, to separate the subgroups. The data of each subgroup are represented by a basis function (BF). The model takes the form of an expansion in product spline basis functions, where the number of basis functions as well as the parameters associated with each one (product degree and knot locations) are automatically determined by the data. Splines are curves which are required to be continuous and smooth. Splines are generally n -degree piecewise polynomials whose function values and first $n-1$ derivatives agree at the points where they join (the abscissa values of the join points are called "knots"). MARS replaces the step function used in trees with a truncated power spline in order to produce a continuous model.

This procedure is motivated by the recursive partitioning approach to regression and shares its attractive properties. Unlike recursive partitioning, however, this method produces continuous models with continuous derivatives. It has more power and flexibility to model relationships that are nearly additive or involve interactions in at most a few variables. In addition, the model can be represented in a form that separately identifies the additive contributions and those associated with the different multivariable interactions. The modeling procedure is inspired by the recursive partitioning technique governing CART and generalized additive modeling (Hastie and Tibshirani, 1990), resulting in a model that is continuous with continuous derivatives. MARS excels at finding optimal variable transformations and interactions, the complex data structure that often hides in high-dimensional data. And hence it can effectively uncover important data patterns and relationships that are difficult, if not impossible, for other methods to reveal.

MARS essentially builds flexible models by fitting piecewise linear regressions; that is, the nonlinearity of a model is approximated through the use of separate regression slopes in distinct intervals of the independent variable space. Therefore the slope of the regression line is allowed to change from one interval to the other as the two "knot" points are crossed. The variables to use and the end points of the intervals for each variable are found via a fast but intensive search procedure. In addition to searching variables one by one, MARS also searches for interactions between variables, allowing any degree of interaction to be considered.

Basis Functions

In multivariate and adaptive regression splines, basis functions are the tools used for generalizing the search for knots. Basis functions are a set of functions used to represent the information contained in one or more variables. The functions used to re-express the relations between dependent and independent variables. Multivariate and Adaptive Regression Splines model almost always creates the basis functions in pairs. For example, basis function (BF1) on the variable elevation is defined by MARS as:

$$BF_1 = \max(0, \text{elevation} - 219) \quad (3.8)$$

Data for elevation variables are grouped into two sets: the first set is assigned 0 for all elevation values that are below a threshold (e.g., $c = 219$ m), and the second set contains the elevation values that are more than 219 m. Elevation has no relation to the probability of presence (i.e., slope = 0) for values below the threshold of 219 m, but has a negative relationship (slope < 0) above this threshold.

The MARS model is a spline regression model that uses a specific class of basis functions as predictors in place of the original data [21]. The MARS basis function transform makes it possible to selectively blank out certain regions of a variable by making them zero, allowing MARS to focus on specific sub-regions of the data. MARS excels at finding optimal variable transformations and interactions, as well as the complex data structure that often hides in high-dimensional data.

MARS uses two-sided truncated functions of the form shown below as basis functions for linear or nonlinear expansion, which approximates the relationships between the response and predictor variables.

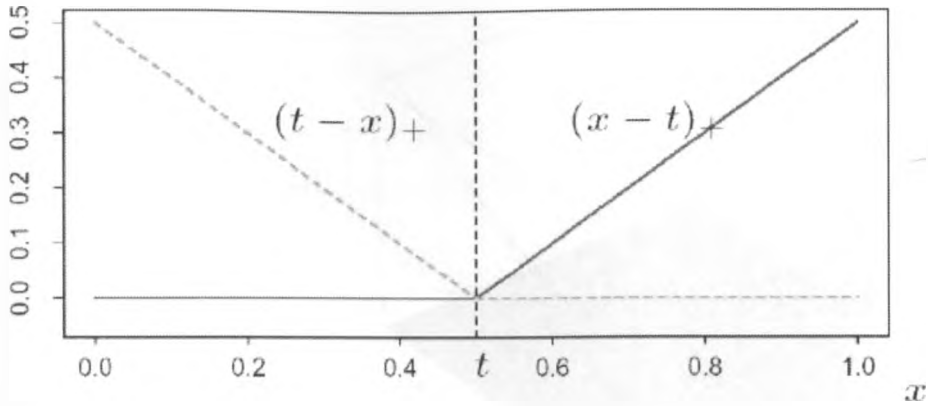


Figure 3.1: The BFs used by MARS $(x-t)_+$ and $(t-x)_+$.

Shown above is a simple example of two basis functions $(t-x)_+$ and $(x-t)_+$ (Hastie, et al., 2001). Parameter t is the knot of the basis functions that defines the "pieces" of the piecewise linear regression; these knots or parameters are determined from the data. The "+" signs next to the terms $(t-x)$ and $(x-t)$ simply denotes that only positive results of the respective equations are considered; otherwise the respective functions evaluate to zero.

The collection of basis functions is:

$$\begin{aligned}
 C &= \{(X_j - t)_+, (t - X_j)_+\} & (3.9) \\
 t &\in \{x_{1j}, x_{2j}, \dots, x_{Nj}\} \\
 j &= 1, 2, \dots, p.
 \end{aligned}$$

If the input values are distinct: $2Np$ basis functions.

The example below is a function $h(X_1, X_2) = (X_1 - x_{51})_+ \cdot (x_{72} - X_2)_+$, resulting from multiplication of two piecewise linear MARS basis functions.

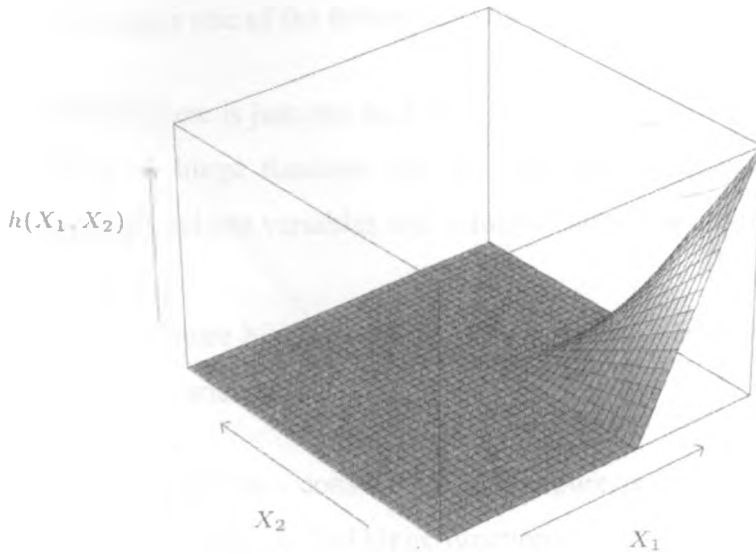


Figure 3.2: Two-way interaction basis functions.

The result is non zero only over the small part of the feature space where both component functions are nonzero.

The MARS Model

The MARS model invented by Friedman (1991) is a flexible nonparametric regression model for high dimensional data. For the presumed system that generated the data:

$$y = f(x_1, \dots, x_p) + \epsilon \tag{3.10}$$

The general MARS model equation (Hastie et al., 2001,) is given as:

$$y = f(x) = \beta_o + \sum_{m=1}^M \beta_m h_m(X) + \epsilon \tag{3.11}$$

Where, y is predicted as a function of the predictor variables X and in some cases together with their interactions; this function consists of an intercept parameter (β_o) and the weighted (β_m) sum of one or more basis functions $h_m(X)$. The summation is over the M non-constant terms in the model. The basis functions together with the model parameters which are estimated through least squares estimation are combined to produce the predictions given the inputs.

Each basis function $\beta_i(x)$ takes one of the following three forms:

- 1) A constant 1. When there is just one such term, the intercept.
- 2) A *hinge function*. A hinge function has the form; $\max(0, x - \text{const})$ or $\max(0, \text{const} - x)$. MARS automatically selects variables and values of those variables for knots of the hinge functions.
- 3) A product of two or more hinge functions. These basis functions can model interaction between two or more variables.

A basis function can be as simple as a constant, or as complex as one or the product of multiple sub-functions that add non-linearities called hinge functions.

The MARS model “selects” a weighted sum of basis functions from the set of a large number of basis functions that span all values of each predictor. The MARS algorithm then searches over the space of all inputs and predictor values i.e. knot locations t as well as interactions between variables. During this search, an increasingly larger number of basis functions are added to the model (selected from the set of possible basis functions), to maximize an overall least squares goodness-of-fit criterion. As a result of these operations, MARS automatically determines the most important independent variables as well as the most significant interactions among them.

The basic underlying assumption of MARS is that the function f is locally smooth. Friedman (1993) extended the MARS methodology to the model having nominal categorical explanatory variables to which the usual definition of smoothness s does not apply. For the case of a simple categorical variable x such that $x \in \{c_1, \dots, c_K\}$, the function estimate is:

$$f(x) = \sum_{m=1}^M a_m I(x \in A_m), M \leq K, \tag{3.12}$$

where I is the indicator function and (A_1, \dots, A_M) are subsets of $\{c_1, \dots, c_K\}$. The estimate with smaller M is said to be smoother. Friedman (1993) developed the MARS algorithm which accommodates mixed (i.e. categorical and continuous) explanatory variables.

Hinge Functions

Hinge functions are a key part of MARS models. A hinge function takes the form $\max(0, x-c)$ or $\max(0, c-x)$ where 0 is the minimum value for the function, x is the variable, and c is a constant that provides a kink or sharp turn in one-dimension called the knot. A Knot is where one local regression model gives way to another and thus is the point of intersection between two splines. An example is shown below of a mirrored pair of hinge functions with a knot at 3.1.

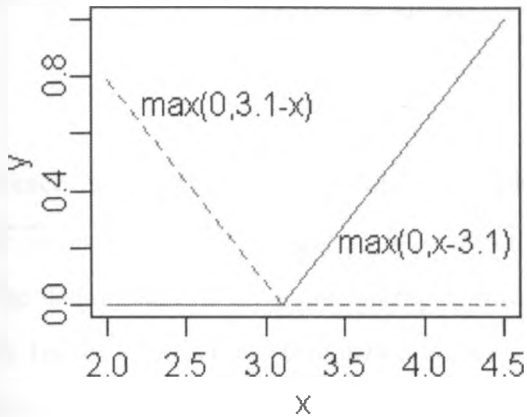


Figure 3.3: A mirrored pair of hinge functions with a knot at $x=3.1$

Apart from forming piecewise linear functions from hinge functions, the hinge functions can also be multiplied together to form non-linear functions.

The Model Building Process: MARS Approach

There are two phases of building a model in MARS, namely the forward and the backward pass. The two stage approach is synonymous to that used by recursive partitioning trees.

The MARS algorithm uses expansions in piecewise linear basis functions to build models from two sided truncated functions of the predictors (x) of the form $(x-t)_+$ and $(t-x)_+$ where “+” means positive part, with a knotting value at t . so:

The following two functions are truncated where $x \in \mathbb{R}$:

$$\text{and } (x-t)_+ = \begin{cases} x-t & \text{if } x > t \\ 0 & \text{otherwise} \end{cases}$$

(3.13)

$$(t-x)_+ = \begin{cases} t-x & \text{if } x < t \\ 0 & \text{otherwise} \end{cases}$$

In Figure 3.1, each function is piecewise linear whose knot is on the point of value t . The equation $(\cdot)_+$ means that only the positive parts are used, if not it is given a zero value. These two functions are named as a reflected pair. Here, as to every input x_{ij} , each input x_j with knot constitutes a reflected pair. So the collection of BFs is:

$$C = \left\{ (x_j - t)_+, (t - x_j)_+ \mid t \in \{x_{1j}, x_{2j}, \dots, x_{Nj}\}, j \in \{1, 2, \dots, p\} \right\} \quad (3.14)$$

If each input value is not equal to any other one, there will be $2Np$ BFs totally. Although every BF is only related to single x_j , we can still consider it a function over the whole input space \mathbb{R}^p . The usual method for generalizing spline fitting in higher dimensions is to employ BF that are the tensor products of univariate spline functions. Therefore, multivariate splines BFs take the following form:

$$B_m(x) = \prod_{k=1}^{K_m} [s_{km} \cdot (x_{v(km)} - t_{km})]_+, \quad (3.15)$$

where K_m is the total number of truncated linear functions in the m th BF, $x_{v(km)}$ is the input variable corresponding to the k th truncated linear function in the m th basis function, t_{km} is the corresponding knot value and $s_{km} \in \{\mp\}$ [22]

The way to construct model is analogous to forward stepwise linear regression, but it allows the use of functions and their products from the set C , not the initial value. The form of the model is as follows:

$$f(x) = \beta_0 + \sum_{m=1}^M \beta_m B_m(x) + \varepsilon, \quad (3.16)$$

Where each $B_m(x)$ is a function from the set C , or the product of two or more functions from C . Given B_m a choice, coefficient β_m could be estimated by minimizing the sum of squares of residual differences, that is, by standard linear regression. However, the important issue is how to

structure the function $B_m(\mathbf{x})$. Firstly, there is only a constant function $B_o(\mathbf{x}) = 1$, and the other functions in set C are candidate functions.

At each phase, we consider the product of the function B_m in model collection M and the function of its reflective pair in C a new basis function pair. Then putting the following formula as an item into the model M , we get:

$$\hat{\beta}_{M+1} B_l(x)(x_j - t)_+ + \hat{\beta}_{M+2} B_l(x)(t - x_j)_+, B_l \in M \quad (3.17)$$

This procedure minimizes training errors. Here, $\hat{\beta}_{M+1}$ and $\hat{\beta}_{M+2}$ are the coefficients which can be estimated by least squares along with other $M + 1$ coefficients. The better product can be put into the model. This process will restart until the number of items in the collection M reaches the given maximum number.

The total model normally overfits the data, in order to avoid this; we start the backward elimination process. In each step we delete the term which causes the smallest increase in the residual squared error. We continue until we attain the best model \hat{f}_α with the optimal value α .

In order to estimate the optimal value of α reducing computational cost, *generalized cross-validation* (GCV) is used. The GCV parameter is an adjusted residual sum of squares, in which a penalty is imposed on the model complexity. Based on the GCV criterion, MARS balances between overfitting and underfitting, in order to return an optimal model. GCV gives the amount of degradation in the model when a variable is deleted. Because a global model tends to be biased but have low variance while local models are more likely to have less bias but suffer from high variance, the MARS approach could be conceptualized as a way to balance between bias and variance. This formulation, also known as lack of fit criterion, is defined in detail in page 37.

These serve as basis functions for linear or nonlinear expansion that approximates some true underlying function $f(x)$.

The MARS model for a dependent variable y , and M terms, can be summarized in the following equation:

$$y = f(x) = \beta_0 + \sum_{m=1}^M \beta_m + H_{km}(x_{v(k,m)}) \quad (3.18)$$

where the summation is over the M terms in the model, and β_0 and β_m are parameters of the model (along with the knots t for each basis function, which are also estimated from the data). Function H is defined as:

$$H_{km}(x_{v(k,m)}) = \prod_{k=1}^K h_{km} \quad (3.19)$$

where $x_{v(k,m)}$ is the predictor in the k 'th of the m 'th product. For order of interactions $K=1$, the model is additive and for $K=2$ the model is pairwise interactive.

The procedure starts with the constant function: $h_0(X) = 1$ and all functions in the set C are candidate functions.

We consider as a new basis function pair all products of a function h_m in the model set M with one of the reflected pairs in C . We add to the model M the term of the form:

$$\tilde{\beta}_{M+1} h_l(X) * (X_j - t)_+ + \tilde{\beta}_{M+2} h_l(X) * (t - X_j)_+, h_l \in M \quad (3.20)$$

that produces the largest decreases in training error.

The winning products are added to the model and the process is continued until the model set M contains some preset maximum number of terms.

The Forward Pass

In this phase, MARS starts with a model which consists of only just the intercept term which is the mean of the response values. MARS then repeatedly adds basis function in pairs to the model. At each step it finds the pair of basis functions that gives the maximum reduction in sum-of-squares of the residual error. The two basis functions in the pair are identical except that a different side of a mirrored hinge function is used for each function. Each new basis function consists of a term already in the model and that could perhaps be the intercept i.e. a constant 1 multiplied

by a new hinge function. A hinge function is defined by a variable and a knot, so to add a new basis function; MARS must search over all combinations of the following:

- 1) Existing terms - called parent terms.
- 2) All variables - to select one for the new basis function.
- 3) All values of each variable - for the knot of the new hinge function.

This process of adding terms continues until the change in residual error is too small to continue or until the maximum number of terms is reached. The maximum number of terms is specified by the user before model building starts. Because of the nature of the hinge functions, the search can be done relatively quickly by a heuristic that reduces the number of parent terms to consider at each step [23].

The Backward Pass

After the forward pass, the usually large model made up of many predictors model that remains is almost always overfit to its training data. Overfitting is undesirable since the “model” that emerges from training will generally not project to new data i.e. an overfit model has a good fit to the data used to build the model but will not generalize well to new data. To build a model with better generalization ability, the backward pass prunes the model much like it is done with CART (Classification and Regressions Trees). It removes terms one by one, deleting the least effective term at each step until it finds the best sub-model. Model subsets are compared using the Generalized cross validation (GCV) criterion described below. Models can then be further cross-validated with test data to assess their fidelity. Analysts can also address overfitting by proactively limiting the number of terms in the model and “penalizing” new entrants on the forward pass.

The backward pass has an advantage over the forward pass: at any step it can choose any term to delete, whereas the forward pass at each step can only see the next pair of terms.

The forward pass adds terms in pairs, but the backward pass typically discards one side of the pair and so terms are often not seen in pairs in the final model.

Generalized Cross Validation (GCV)

The backward pass uses GCV to compare the performance of model subsets in order to choose the best subset: lower values of GCV are better. The GCV is a form of regularization that trades off goodness-of-fit against model complexity. In most cases, we want to estimate how well a model performs on *new* data, not on the training data. Such new data is usually not available at the time of model building, so instead we use GCV to estimate what performance would be on new data. The raw residual sum-of-squares (RSS) on the training data is inadequate for comparing models, because the RSS always increases as MARS terms are dropped. In other words, if the RSS were used to compare models, the backward pass would always choose the largest model - but the largest model typically does not have the best generalization performance.

The formula for the GCV is:

$$\text{GCV} = \text{RSS} / (N * (1 - \text{Effective Number Of Parameters} / N)^2)$$

Where RSS is the residual sum-of-squares measured on the training data and N is the number of observations or the number of rows in the x matrix.

The Effective Number of Parameters is defined in the MARS context as:

Effective Number of Parameters = Number of Mars Terms + Penalty * (Number of Mars Terms - 1) / 2, where Penalty is about 2 or 3.

Note that (Number of Mars Terms - 1) / 2 is the number of hinge-function knots, so the formula penalizes the addition of knots. Thus the GCV formula adjusts (i.e. increases) the training RSS to take into account the flexibility of the model. We penalize flexibility because models that are too flexible will model the specific realization of noise in the data instead of just the systematic structure of the data.

The Generalized Cross Validation error is a measure of the goodness of fit that takes into account not only the residual error but also the model complexity as well. It is given by:

$$GCV = \frac{\sum_{i=1}^N (y_i - f(x_i))^2}{\left(1 - \frac{c}{N}\right)^2} \quad (3.21)$$

with

$$C = 1 + cd$$

where N is the number of cases in the data set, d is the effective degrees of freedom, which is equal to the number of independent basis functions. The quantity c is the penalty for adding a basis function. Typically experiments have shown that the best value for C can be found somewhere in the range $2 < d < 3$ (Hastie et al., 2001).

In order to evaluate the appropriateness of the model, MARS uses generalized cross-validation (GCV) which is residual square errors penalized by a function related to complexity of the model [21]. The numerator in GCV is the average residual squared error and the denominator is a penalty term that reflects model complexity. The use of the denominator is to prohibit selection of a model with many terms that decreases only slightly the residual errors. The GCV statistic is an estimate of the variance for error in a regression model that includes a penalty term for the number of parameters used in the regression. The GCV R-squared statistic is the ordinary R-squared statistic calculated with the variance for error replaced with the GCV statistic [21].

Generalized Cross Validation is so named because it uses a formula to approximate the error that would be determined by leave-one-out validation. It is just an approximation but works well in practice. GCVs were introduced by Craven and Wahba (1979) and extended by Friedman for MARS.

ADVANTAGES AND DISADVANTAGES OF MARS.

No single adaptive regression modeling technique can perform uniformly best for all situations.

Advantages:

MARS has a lot to offer as a predictive modeling mainstay.

- Useful tool for simplifying high-dimensional problems where there are many explanatory variables.
- MARS can handle continuous and categorical independent and dependent variables making it a powerful general-purpose tool. In addition, like CART and stepwise regression, MARS computations can be automated, modelers having to choose only input variables and tuning parameters.
- MARS uses piecewise linear functions which produce continuous models and provides a more effective way to model nonlinearities.
- MARS is not computationally intensive and is straightforward to implement in order to look for suitable interactions between independent variables, which make it in particular preferable whenever there is a large number of interacting variables.
- Both the additive and the interactive effects of the predictors are allowed to determine the response variable.
- Though it's a non-parametric technique that makes no assumptions on how dependent variables relate to predictors, MARS feels a lot like traditional least squares regression, albeit with much more flexibility, and is easier to interpret than "black box" machine learners like neural nets and random forests.
- MARS can handle complex (nonlinear) relationships and interactions providing an interpretable model.
- MARS identifies interactions and also produces graphs that help visualize and understand interactions.
- MARS has automated capabilities for handling missing data, a common feature of large databases.

- MARS is pretty efficient and able to handle large models in a reasonable amount of time and computer resources in as much as it is dependent on execution options.

Disadvantages:

- The MARS methodology has a risk of over-fitting because of very exhaustive search that is conducted to identify nonlinearities and interactions. However, there are protections against over-fitting such as setting a lower maximum number of BFs and a higher "cost" per knot [22].
- The dataset has to be large enough to make the use of MARS possible and worthwhile.
- While MARS has offered a reasonable compromise to the bias/variance challenge and has developed a reasonable reputation for predictive accuracy, it is not as good as the more computationally-intensive bagging or boosting.

CHAPTER 4

DATA ANALYSIS AND RESULTS

This chapter describes the data that was used in the project, its collection and the results obtained.

DATA

Six thousands and twenty four (6024) TB-infected patients receiving TB treatment under the DOTS strategy in various public and private medical institutions in Kenya were eligible for inclusion in the study. A total of 1419 participants had more than 50% missing observations in the variables of interest to the study and were excluded from the analysis. Therefore, only 4605 participants were included in the analysis.

To check the model validity, whole data set was divided into training set including 67% of the instances to build the model and testing sets with 33% of the instances to check the developed model validity respectively. (Sharma, 1996) indicated that classification of the same data used in model estimation is biased since it only reflects model fit and not necessarily its predictive ability. The optimal strategy is to develop models using the training samples then apply the models to testing samples to evaluate the predictive performance of the models. Applying and refitting models using the validation data set protects against model over-fit that can result from aggressive use of data mining procedures. 16 patients' attributions were applied to be explored and after variable selection process via an automatic process in MARS the model was developed through 9 identified components out of 40 predictors.

DATA COLLECTION

TB is a notifiable disease under the Public Health Act Cap 242 and therefore all TB Cases (diagnosed by the public or private sector) must be notified to the Ministry of Health through National TB control program. The Kenya national TB treatment guidelines of 2009, states that all patients diagnosed in health care facilities implementing the DOTS must be registered at the start of treatment. The TB patients are *line listed* in TB treatment facility register which is maintained at each health facility where tuberculosis treatment provided. These patients are then registered by district TB and leprosy coordinator (DTLC) into one combined district register and in which patient is given unique district registration number. The register takes the form of either manual system or an electronic TB register. The district register forms the basis of notifying the country of diagnosed and registered TB patients every calendar year.

In order to standardized management, recording and reporting in TB control, Tuberculosis patients are classified into three categories namely:-

1. *Category I*: New Smear positive PTB patients who have never been treated before or used anti-TB drugs for less than one month.
2. *Category II*: Previously treated TB patients
 - i. Relapse
 - ii. Failures
 - iii. Returnees after default
3. *Category III*: New tuberculosis patients with less severe form of TB namely
 - i. New Sputum smear negative PTB
 - ii. Extra-pulmonary (EPTB).

All the patients diagnosed with TB at one of these Health Facilities (HFs) are given DOT (Directly observed therapy) in accordance with Division of Leprosy TB and Lung Disease policy guidelines. Every dose of treatment is directly observed during intensive phase (IP) and continuation phase (CP) by treatment supporter. The patients make weekly visit during intensive phase (2 months) and fortnightly visits in continuation phase (4 months), where the patient is reviewed, follow up test done and drug fill up undertaken. In all, the patients, makes minimum of 16 visits during the 6 months of treatment. At the end of the treatment period, patients are evaluated for treatment outcomes. The treatment outcomes of the TB cohort are notified to the na-

tional program every 13-15 months from the date of registration. These cohort treatment outcomes include;

1. **TB Cured**, which refers to a patient who was initially TB smear-positive who then becomes smear-negative in the last month of treatment and on at least one previous occasion.
2. **TB Completed treatment, which is define as a patient** who completed treatment but did not meet the criteria for cure or failure. This definition applies to pulmonary smear-positive and smear-negative patients and to patients with extra pulmonary disease.
3. **Died**, which refers to a patient who died from any cause during treatment.
4. **Treatment failure**, which is defined as a patient who was initially smear-positive and who remained smear-positive at month 5 or later during treatment.
5. **Defaulted/Out of Control**, this is defined as a patient whose treatment program was interrupted for 2 consecutive months or more.

Successful treatment therefore refers to a patient who was cured or who completed treatment. Based on the proportions of the treatment outcomes, bootstrapping was done on four treatment outcomes namely: - Dead, Failed, Out of Control and Transfer out outcomes to generate 2000 cases for each treatment outcome. From this newly created dataset of 11897 instances, 7972 records were used to build the model and testing sets with 3925 instances.

VARIABLES

THE OUTCOME VARIABLES

1. Completed treatment- Treatment Success
2. Interrupted treatment – Out of Control
3. TB mortality – Dead ~
4. Transfer out
5. Treatment failure

The proportions of the treatment outcomes in the original data were as follows:

Table 4.1: Treatment outcomes proportion in the original data

Treatment outcome	No of cases	Proportion %
DEAD	212	4.5
FAILED	27	0.5
OUT OF CONTROL	253	5
TRANSFER OUT	216	5
TREATMENT SUCCESS	3897	85

EXPOSURE VARIABLES

1. Age
2. ART
3. CotrimPrevTherapy
4. DOTIntPhase
5. Height
6. HIVTest
7. NutriSupport
8. PatientType
9. Region
10. PTBSubType
11. Sex

12. SputumMo0

13. TBType

14. Weight

15. X_Ray

16. BMI

The distributions of the 16 predictor variables were as follows:

Table 4.2: A summary of continuous predictor variables

variable	Min	1 st Quartile	Median	Mean	3 rd Quartile	Max
Age	1	24	32	32.8	40	99
Height	0.34	1.33	1.6	1.51	1.7	3
Weight	4.7	47	55	54	62	117
BMI	2.8	17.3	21.8	31.65	29.96	51

Table 4.3: A summary of categorical predictors

Variable	Levels	Number of cases
ART	N – No	1367
	Unknown	2106
	Y –Yes	1132
CotrimPrevTherapy	N	772
	Unknown	2083
	T	1750
DOTIntPhase	CV	147
	H	3799
	HCW	659

HIVTest	Declined	10
	Negative	2292
	Not Done	505
	Positive	1798
NutriSupport	N	2452
	U	993
	Y	762
	ND	233
	MN	108
	FS	31
	VITAMIN	19
	FM	7
PatientType	F	55
	N	3709
	R-	307
	R+	381
	RAD	32
	REP	97
	TI	24
Region	CENTRAL	333
	COAST	154
	NAIROBI NORTH	1046
	NAIROBI SOUTH	1827
	NORTH EASTERN	71
	NYANZA NORTH	637

	WESTERN	537
PTBSubType	ND – Not Done	251
	Neg -Negative	1580
	Pos - Positive	2774
Sex	Female	1831
	Male	2774
SputumMo0	ND – Not Done	1530
	Neg - Negative	1486
	Pos - Positive	1589
TBType	EP- Extra Pulmonary	1304
	P- Pulmonary	3301
X_Ray	No	1460
	Unknown	175
	Yes	2970

DATA PROCESSING METHODS AND DATA ANALYSIS

Data management was conducted by the use of Microsoft Office Access by DLTLTD. For this secondary data analysis study, data was received in Microsoft Office Excel format. R version 2.15.1 (R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.) was used for data cleaning and statistical analysis. All variables for the study were imported into R version 2.15.1 and important variables for model building were automatically selected in R via the *earth* package. Data cleaning in R version 2.15.1 included assessing quality of data in terms of missing values, and internal consistencies of responses.

VARIABLE SELECTION AND IMPORTANCE

Variable's importance is a measure of the effect that observed changes to the variable have on the observed response or the expectation of that effect over the population. It can be measured by changing the variable's value and measuring how the response changes.

THREE CRITERIA FOR ESTIMATING VARIABLE IMPORTANCE IN MARS

- (i) The *nsubsets* criterion counts the number of model subsets that include the variable. Variables that are included in more subsets are considered more important. "*Subsets*" mean the subsets of terms generated by the pruning pass.

- (ii) The *rss* criterion first calculates the decrease in the RSS for each subset relative to the previous subset. For multiple response models, RSS's are calculated over all responses. Then for each variable it sums these decreases over all subsets that include the variable. Variables which cause larger net decreases in the RSS are considered more important.

- (iii) The *gcv* criterion is the same, but uses the GCV instead of the RSS. Adding a variable can increase the GCV, i.e., adding the variable has a deleterious effect on the model.

From the above, 9 of 40 predictors were selected as shown in the table below:

Table 4.4: A table of variable importance

variable	Nsubsets	Gcv	Rss
SputumMo0Pos	12	100	100
HIVTestPos	11	88.6	89
ARTUnknown	10	80.4	81
X_RayYes	9	70.8	71.8
Age	8	63.0	64.2
RegionNAIROBINORTH	7	55.2	56.5
NutriSupportND	6	46.0	47.7
Weight	5	36.3	38.5
BMI	4	31.9	33.9

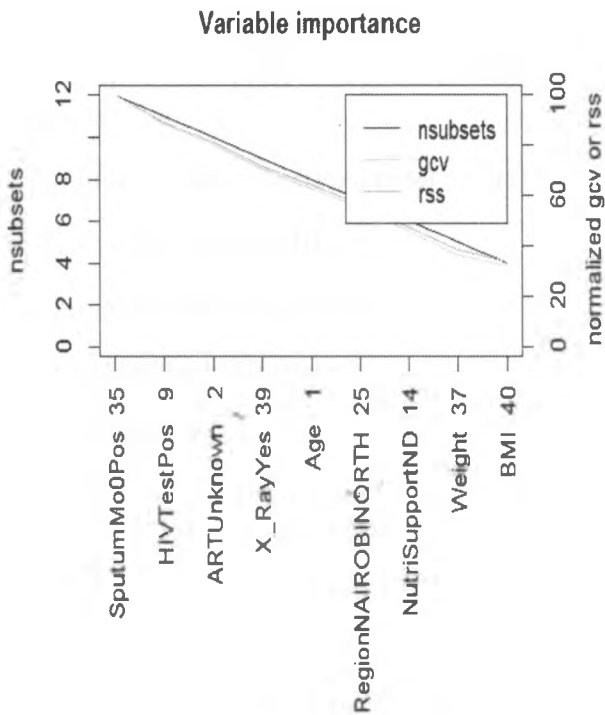


Figure 4.1: A plot of variable importance

MULTIVARIATE MODEL BUILDING AND ANALYSIS

If the response y has k columns then MARS via *earth* builds k simultaneous models. Each model has the same set of basis functions i.e. the same selected terms, and cuts but different coefficients. The returned coefficients will have k columns. The models are built and pruned as usual but with the GCVs summed across all k responses.

Factors are treated in the response in a non-standard way that makes use of *earth's* ability to handle multiple responses, e.g. a two level factor or logical is converted to a single indicator column of 1s and 0s. A factor with three or more levels is converted into k indicator columns of 1s and 0s, where k is the number of levels.

Basis Functions

8 basis functions for predicting the treatment outcomes were generated as follows:

Table 4.5: A table of basis functions

Basis function	Definition	P-value
bf1	h(Age-28)	0.99E-15
bf2	h(28-Age)	0.86E-15
bf3	ARTUnknownHIVTestPosRegionNAIROBINORTHSputum-Mo0Posh(Age-28)	0.1276
bf4	NutriSupportNDHIVTestPosh(BMI-39.54)	0.2134
bf5	HIVTestPosh(39.54-BMI)	0.1257
bf6	SputumMo0Posh(Weight-70)	0.1923
bf7	SputumMo0Posh(70-Weight)	0.0987
bf8	SputumMo0PosX_RayYes	0.3247

Treatment Outcome Models

Five (5) simultaneous models for the five treatments outcomes were constructed and they are as follows:

$$\begin{aligned} \text{DEAD} = & 0.3689167 + 0.06825529 * \text{bf1} + 0.01999946 * \text{bf2} + 0.06876191 * \text{ART Unknown} + \\ & 0.1853321 * \text{HIVTestPos} - 0.07115787 * \text{RegionNAIROBINORTH} - 0.1199515 * \text{SputumMo0Pos} - \\ & 0.09186554 * \text{bf1} * \text{NutriSupportND} - 0.01830668 * \text{HIVTestPos} * \text{h(BMI-39.54)} + 0.02430828 * \\ & \text{HIVTestPos} * \text{h(39.54-BMI)} - 0.05304674 * \text{SputumMo0Pos} * \text{h(Weight-70)} + 0.08682914 * \text{Sputum-} \\ & \text{Mo0Pos} * \text{h(70-Weight)} + 0.1304237 * \text{SputumMo0Pos} * \text{X_RayYes} \end{aligned}$$

$$\begin{aligned} \text{FAILED} = & 0.1745858 - 0.02980281 * \text{bf1} - 0.0667545 * \text{bf2} - 0.06677359 * \text{ARTUnknown} - \\ & 0.1761477 * \text{HIVTestPos} - 0.01083627 * \text{RegionNAIROBINORTH} + 0.4142206 * \text{SputumMo0Pos} + \\ & 0.01899715 * \text{bf1} * \text{NutriSupportND} + 0.01778107 * \text{HIVTestPos} * \text{h(BMI-39.54)} + 0.08410968 * \\ & \text{HIVTestPos} * \text{h(39.54-BMI)} - 0.01575885 * \text{SputumMo0Pos} * \text{h(Weight-70)} - 0.08676202 * \text{Sputum-} \\ & \text{Mo0Pos} * \text{h(70-Weight)} - 0.2733959 * \text{SputumMo0Pos} * \text{X_RayYes} \end{aligned}$$

$$\begin{aligned} \text{OUTOFCONTROL} = & 0.07799735 + 0.001405914 * \text{bf1} + 0.06948043 * \text{bf2} + 0.08782741 * \text{AR-} \\ & \text{TUnknown} + 0.05168776 * \text{HIVTestPos} + 0.1364471 * \text{RegionNAIROBINORTH} - 0.01313107 * \text{Sp-} \\ & \text{utumMo0Pos} + 0.01870312 * \text{bf1} * \text{NutriSupportND} + 0.01959584 * \text{HIVTestPos} * \text{h(BMI-39.54)} - \\ & 0.01106825 * \text{HIVTestPos} * \text{h(39.54-BMI)} - 0.03631802 * \text{SputumMo0Pos} * \text{h(Weight-70)} - \\ & 0.05495595 * \text{SputumMo0Pos} * \text{h(70-Weight)} - 0.02850668 * \text{SputumMo0Pos} * \text{X_RayYes} \end{aligned}$$

$$\begin{aligned} \text{TRANSFEROUT} = & 0.06120845 - 0.01838049 * \text{bf1} + 0.0141061 * \text{bf2} + 0.1917132 * \text{ARTUnk-} \\ & \text{nown} + 0.2715252 * \text{HIVTestPos} - 0.103655 * \text{RegionNAIROBINORTH} - 0.1143405 * \text{Sputum-} \\ & \text{Mo0Pos} - 0.06740918 * \text{bf1} * \text{NutriSupportND} - 0.02110266 * \text{HIVTestPos} * \text{h(BMI-39.54)} - \\ & 0.06307808 * \text{HIVTestPos} * \text{h(39.54-BMI)} + 0.01070733 * \text{SputumMo0Pos} * \text{h(Weight-70)} + \\ & 0.04358813 * \text{SputumMo0Pos} * \text{h(70-Weight)} + 0.02295 * \text{SputumMo0Pos} * \text{X_RayYes} \end{aligned}$$

$$\begin{aligned} \text{TREATMENTSUCCESS} = & 0.6493167 - 0.03413113 * \text{bf1} + 0.03839638 * \text{bf2} - 0.281529 * \text{AR-} \\ & \text{TUnknown} - 0.3323974 * \text{HIVTestPos} + 0.04920199 * \text{RegionNAIROBINORTH} - 0.1667975 * \text{Sp-} \\ & \text{utumMo0Pos} - 0.04939993 * \text{bf1} * \text{NutriSupportND} + 0.0180354 * \text{HIVTestPos} * \text{h(BMI-39.54)} - \\ & 0.01239418 * \text{HIVTestPos} * \text{h(39.54-BMI)} + 0.01071938 * \text{SputumMo0Pos} * \text{h(Weight-70)} + \\ & 0.04780119 * \text{SputumMo0Pos} * \text{h(70-Weight)} + 0.1485289 * \text{SputumMo0Pos} * \text{X_RayYes} \end{aligned}$$

MODEL SUMMARY

Table 4.6: A table of model summary

Treatment Outcome	GCV	RSS	RSq	ClassRate	Sd
DEAD	0.124	98.69	0.97	0.78	0.013
FAILED	0.114	90.49	0.91	0.80	0.018
OUTOFCONTROL	0.134	106.34	0.51	0.83	0.005
TRANSFEROUT	0.133	105.61	0.48	0.83	0.005
TREATMENTSUCCESS	0.204	161.30	0.84	0.66	0.018
All	0.171	562.44	0.79	0.70	0.013

R-squared - 0.79

The GCV and RSq are measures of the generalization ability of the model, i.e., how well the model would predict using data not in the training set. The effective number of model parameters is a just an estimate in MARS models. With an R-squared value of 0.79, the predictive model has a satisfactory discriminative ability.

MODEL DIAGNOSTICS

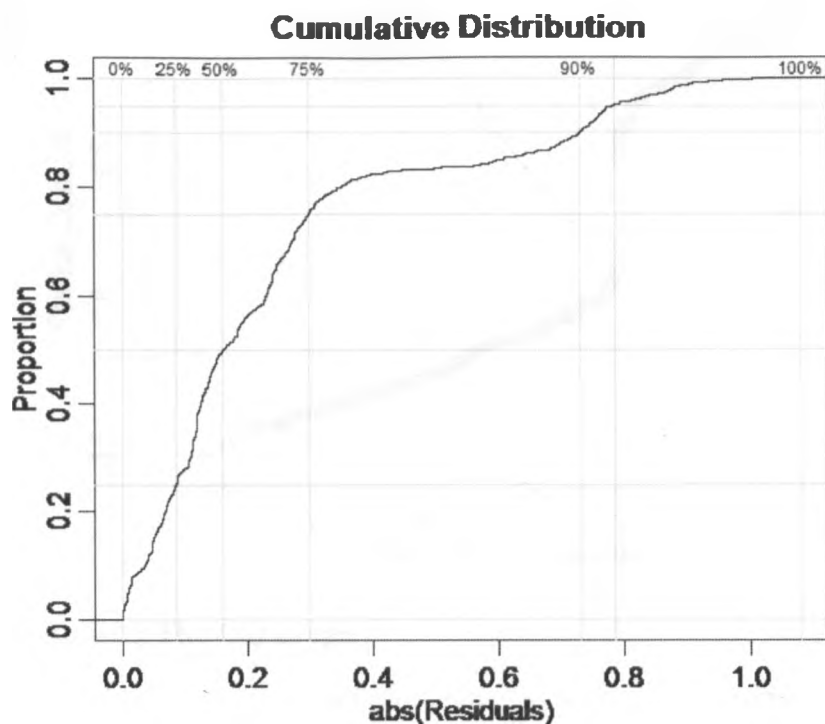


Figure 4.2 The Cumulative Distribution graph

The Cumulative Distribution graph above shows the cumulative distribution of the absolute values of residuals. An ideal situation in a cumulative distribution curve would be a graph that starts at 0 and shoots up quickly to 1. In Figure 4.2, the median absolute residual is about 0.18 (by looking at the vertical gray line for 50%). We also see that 90% of the absolute values of residuals are less than about 0.7. So in the training data, 90% of the time the predicted value is within 0.7 units of the observed value.

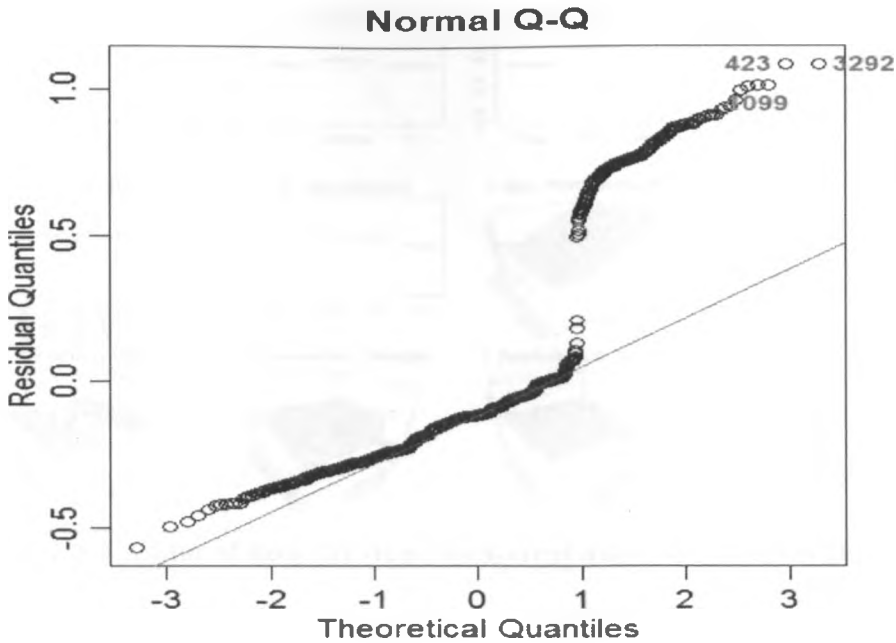


Figure 4.3 The QQ graph

The QQ (quantile-quantile) plot above compares the distribution of the residuals to a normal distribution. If the residuals are distributed normally they will lie on the line. Normality of the residuals isn't too important for MARS models build via *earth* package, but the graph is useful for discovering outlying residuals and other anomalies. We see that cases 423, 1099, and 3292 have the largest residuals.

MODEL'S PARTIAL RESPONSE

We plot a degree1 main effect plot to predict the response when changing one variable while holding all other variables at their median values. For degree2 or interaction plots, two variables are changed while holding others at their medians. The first level is used instead of the median for factors.

The following are plots of a model's response when varying one or two predictors while holding the other predictors constant.

DEAD earth(formula=TreatOutcome~.,data=train,...

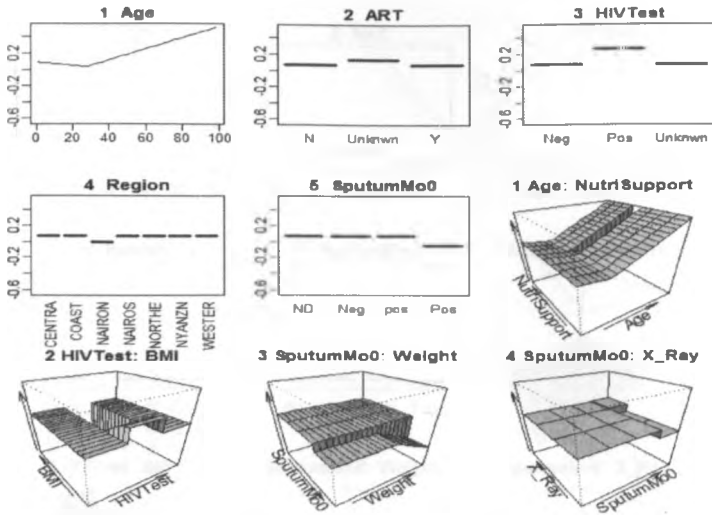


Figure 4.4 A plot of how the dead treatment outcome varies with the predictors

FAILED earth(formula=TreatOutcome~.,data=trai...

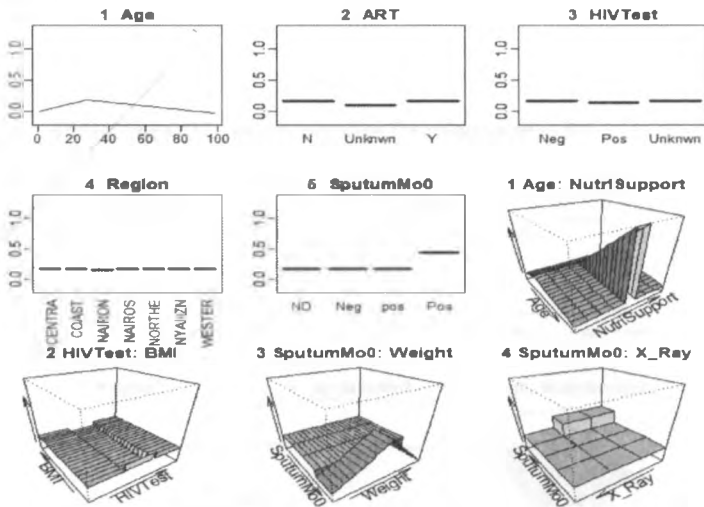


Figure 4.5 A plot of how the failed treatment outcome varies with the predictors

OUTOFCONTROL earth(formula=TreatOutcome~.,da...

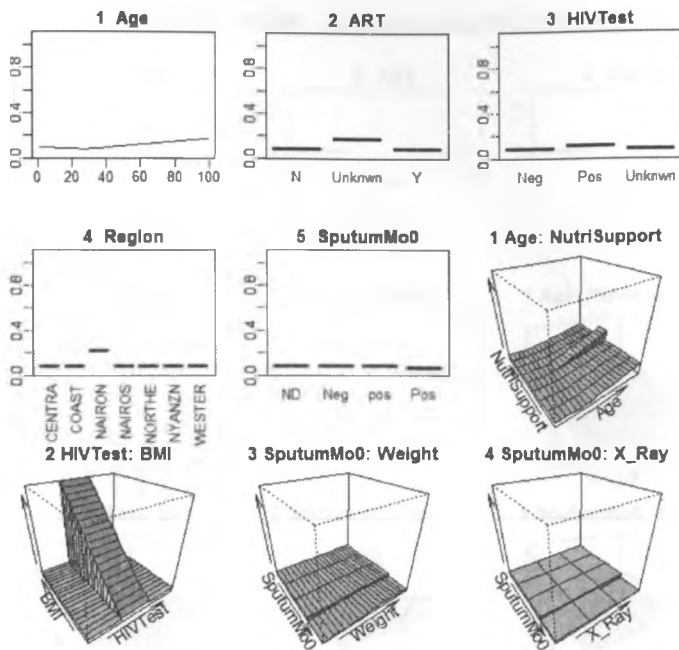


Figure 4.6 A plot of how the out of control treatment outcome varies with the predictors

TRANSFEROUT earth(formula=TreatOutcome~.,dat...

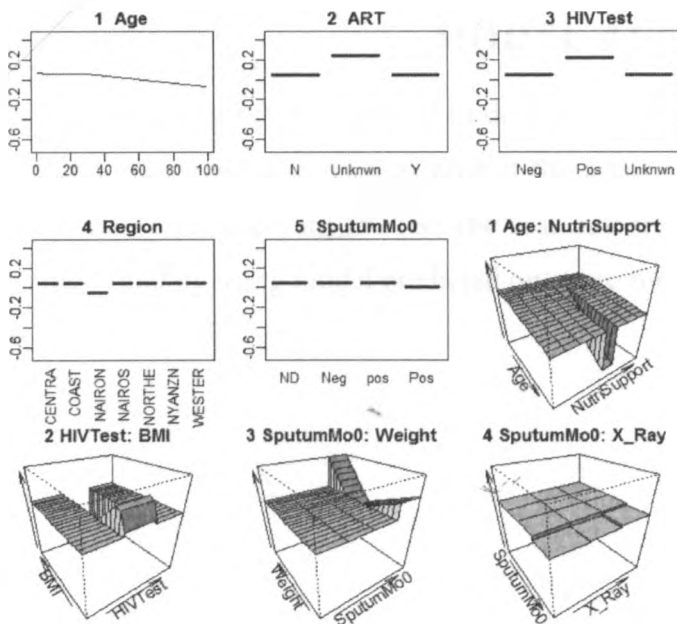


Figure 4.7 A plot of how the transfer out treatment outcome varies with the predictors

TREATMENTSUCCESS earth(formula=TreatOutcome~...

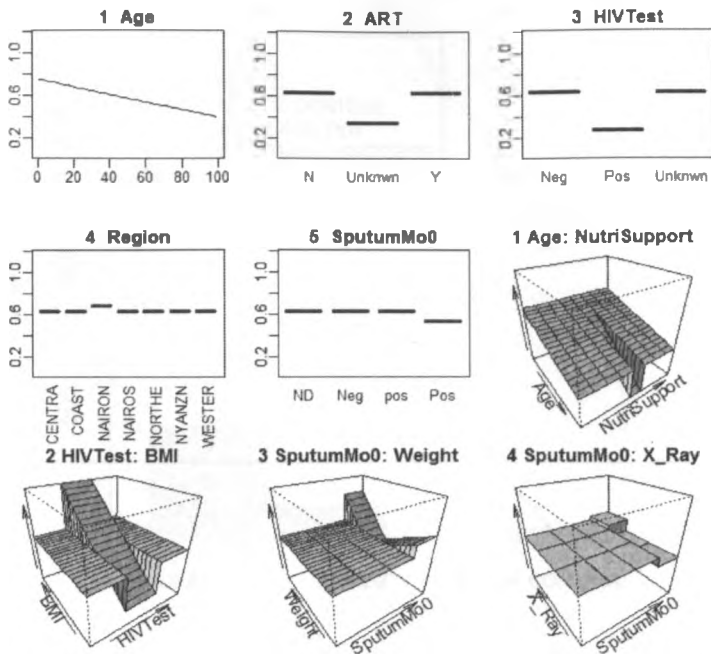


Figure 4.8 A plot of how the treatment success treatment outcome varies with the predictors

MODEL ACCURACY

First, the training set was applied to develop the model; afterwards to check the model's accuracy the testing set was used to predict the given outcome which was the TB patients' course destination after applying DOTS. The real outcome for each patient in testing set was already available and by using model predicted outcome for each case that were defined.

model class nresp=1

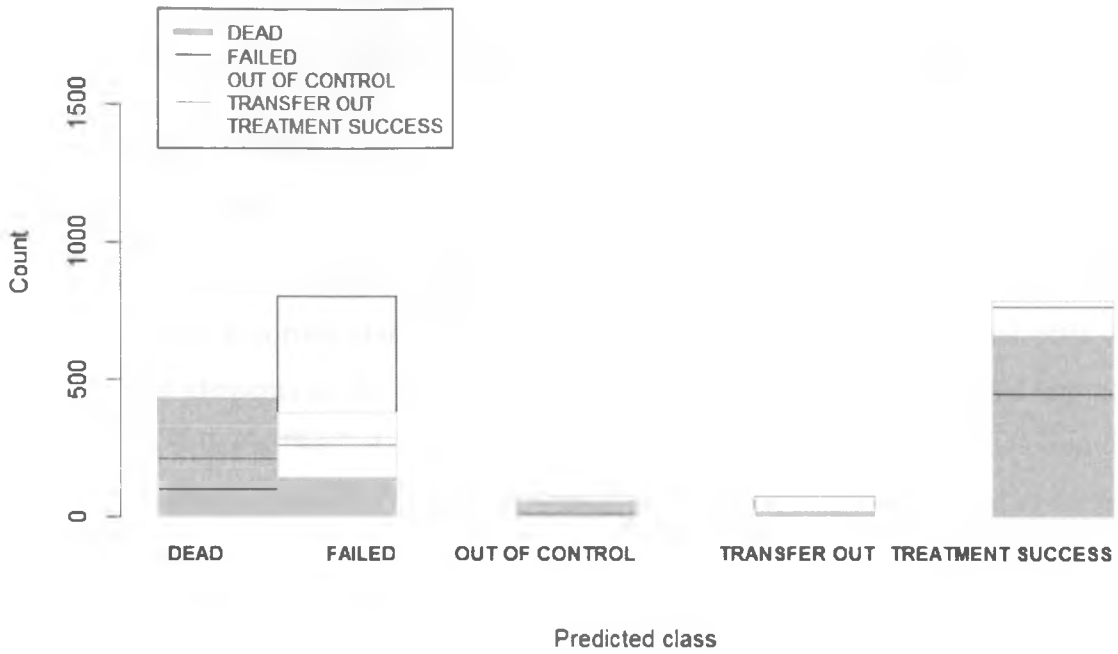


Figure 4.9 A plot of the distribution of the predicted values for each class

MISCLASSIFICATION MATRIX

Table 4.7 Classification Accuracy Matrix

TREATMENT OUTCOME	DEAD	FAILED	OUT OF CONTROL	TRANSFER OUT	TREATMENT SUCCESS
DEAD	0.63	0.07	0.10	0.05	0.15
FAILED	0.05	0.80	0.04	0.04	0.07
OUT OF CONTROL	0.08	0.00	0.66	0.14	0.12
TRANSFER OUT	0.05	0.07	0.05	0.65	0.18
TREATMENT SUCCESS	0.05	0.06	0.05	0.06	0.78

Table 4.8 Misclassification matrix

TREATMENT OUT- COME	DEAD	FAILED	OUT OF CON- TROL	TRANSFER OUT	TREATMENT SUCCESS
MISCALSSIFICA- TION RATE	0.371	0.202	0.338	0.350	0.220

Risk estimate = 0.1725

Standard error= 0.079

The misclassification matrix counts up the predicted and actual category values and displays them in a table. A correct classification is added to the counts in the diagonal cells of the table. The diagonal elements of the table represent agreement between the predicted and actual value, this is often called a "hit." An incorrect classification, called a "miss", means that there is disagreement between predicted and actual value. Misclassifications are counted in the off-diagonal elements of the matrix.

The risk estimate and standard error of risk estimate are values that indicate how well the model classification is performing. In this case, the risk estimate for the five-level treatment outcome MARS model is 0.1725, and the standard error for the risk estimate is 0.079. In other words, we are missing 17.25% of the time. This estimate implies that more parameters needs to be factored into the model to account for the missing variation, nevertheless the statistic is predictive enough for the model to be used as a decision support tool.

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

SUMMARY

The objective of this thesis is to construct and validate a representation of TB Treatment outcome dynamics using a Multivariate Adaptive regression Splines (MARS), so that we can enhance our understanding of the dynamics of TB in terms of inter relationship between factors ranging from demographic, clinical and social factors. The accuracy of MARS model is evaluated so that confidence can be built around the applied statistical technique. MARS has proven to be a powerful nonparametric regression and classification technique that can unearth hidden patterns in data especially where the underlying dynamics are not generally well known, the automatic variable importance selection and interaction capability of MARS is awesome and should be explored further as a statistical and data mining tool to build complex models. Such models like TB course destination determining can assist in sorting alternatives and optimizing TB prevention and control programs.

Pursuing the idea of providing the DOTS in different levels to TB patients based on their status is a necessary purpose requiring a tool to determine the patient destination after getting DOTS. This study was aimed to develop this tool as a valid MARS model. This prediction would be carried out at commence of patient treatment in frame of DOTS. This valid MARS model can determine the level of patient support and supervision assisting the health workers to understand how intensive should be their care for each specific patient.

Practical use of studies such as this, should be made use of in order to ensure that every bit of information that could help the individuals struggling with this disease find comfort and happiness in their lives.

All the R codes use to generate the various statistics are available in the Appendix.

LIMITATIONS OF THE STUDY

The data was not having information on drug-resistance TB. Drug-resistance TB cannot be cured with standard treatment or first-line drugs and is often associated with high mortality. One of the other limitations in this study concerned the application of the earth package in R for modelling purposes. There are aspects of MARS mentioned in Friedman's papers but not implemented in earth and this includes:

- (i) Piecewise cubic models (to smooth out sharpness at the hinges).
- (ii) Model slicing.
- (iii) Handling missing values.
- (iv) Automatic grouping of categorical predictors into subsets.

Also, although several risk factors were examined, there is probably considerable residual confounding. CD4 counts, which would be a better measure of the degree of immune suppression in HIV-TB co-infection, were available with considerable missing data and were not included in the analysis. This study as a secondary data analysis study does not have control over data quality, collection methods and missing information. The problem of under-reporting of TB mortality may arise where a proportion of patients with "missing outcome" have in fact died with the reported cause of death as TB.

PLAN FOR UTILIZATION AND DISSEMINATION OF RESULTS

Copies of the final report will be submitted to Department of Mathematics, University of Nairobi. Results of the study will be presented to the University of Nairobi school of Mathematics. Other copies of the final report will be available to the Department of Tuberculosis, Leprosy and Lung Disease of the Ministry of Health, Kenya on request. The results will be ready for publishing in a leading journal.

FUTURE RESEARCH

A further project in this area especially concerning Tb control could involve investigation of the effects of disease features (e.g. length of latent period) upon prevalence level and persistence time, by numerical evaluation of quasi-stationary distributions and analytical study of approximation methods. This includes studying the behavior of an infection which has become established in the population, i.e. an endemic disease. Quantities of interest include the endemic level of prevalence i.e. proportion of the population infected, and the persistence time i.e. the time until the disease eventually dies out.

Much infectious disease modelling is concerned with studying the initial stages of an epidemic outbreak. In order to achieve a major objective of mathematical modelling of disease transmission and to understand how best to intervene to reduce spread of infection, one approach to this is through mathematical control theory, whereby each infection generates a cost, any form of intervention generates a cost, and the aim is to minimize the combined total cost of infection and of intervention. A project in this area will involve numerical evaluation of optimal disease of the policy.

The next step would be to use the model to predict the practical cases for specific patients. The cases will need to take the form of a treatment schedule to make a real difference. This implies that a controller for this model should be designed and a method devised to convert the equations to physical medical scheme inputs to the system.

Co-infection with other disease could also be modelled. Many co-infections exist which can be modelled, verified and simulated.

REFERENCES:

1. Stedman's Medical Dictionary, 26th ed. (Baltimore: Williams and Wilkins. 1996).
2. World Health Organization. Tuberculosis facts. Geneva: WHO report, 2008.
3. World Health Organization, *Global Tuberculosis Control: Epidemiology, Strategy, Financing*. 2009, World Health Organization: Geneva.
4. Raviglione MC, Snider DE, Kochi A. Global epidemiology of tuberculosis: morbidity and mortality of a worldwide epidemic. *JAMA* 1995; 273:220—6, World Health Organization. Treatment of tuberculosis: guidelines for national programs. Geneva, Switzerland: World Health Organization; 2003 (WHO/CDS/TB/2003.313).
5. World Health Organization. The Stop TB Strategy, Document WHO/HTM/TB/2006.35. Geneva: WHO.
6. Sumartojo E. When tuberculosis treatment fails. A social behavioural account of patient adherence. *Am Rev Respir Dis* 1993; 147:1311—20.
7. R. M. Anderson and R. M. May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, 1991.
8. Chambless LE, Dobson AJ, Patterson CC, Raines B. On the use of a logistic risk score in predicting risk of coronary heart disease. *Stat Med* 1990; 9:385—396.
9. Grundy SM, Balady GJ, Criqui MH, Fletcher G, Greenland P, Hiratzka LF, et al. Primary prevention of coronary heart disease: guidance from Framingham. *Circulation* 1998; 97:1876—87.
10. Fawcett, Tom, "An Introduction to ROC Analysis," Elsevier B.V. 2005.
11. Pepe, M.S., "The Statistical Evaluations of Medical Tests for Classification and Prediction," Oxford, UK: Oxford University Press, 2003. Lasko, Thomas A., Jui G. Bhagwat,

- Kelly H. Zou, and Lucila Ohno-Machado , “The use of receiver operating characteristic curves in biomedical informatics,” *Journal of Biomedical Informatics*, Vol. 38, Issue 5 *Clinical Machine Learning*, October 2005, pp. 404-415, PMID: 16198999.
12. Hosmer, David W. and Stanley Lemeshow, “Applied Logistic Regression 2nd edition.” John Wiley & Sons, Inc., New York, NY, 2000, pgs. 160-164.
 13. E. Ziv, C.L. Daley and S.M. Blower. Early therapy for latent tuberculosis infection. *Amer. J. of Epidemiology*. 153(4), 381-385, 2001.
 14. J. P. Aparicio and J. C. Hernandez. Preventive treatment of tuberculosis through contact tracing. *Contemporary Mathematics*, Volume 410, 2006.
 15. N. Osgood, A. Mahamoud, K. Hassmiller, Y. Tian, A. Al-Azem and V. Hoepfner. Estimating the Relative Impact of Early-Life Infection Exposure on Later-Life Tuberculosis Outcomes in a Canadian Sample. *Research in Human Development*. 8(1). pp.26-47, 2011.
 16. Davies GR, Connolly C, Sturm AW, McAdam KPWJ, Wilkinson D. Twice-weekly, directly observed treatment for HIV-infected and uninfected tuberculosis patients: cohort study in rural South Africa. *AIDS* 1999;13:7., Gandhi NR, Moll AP, Pawinski R, Zeller K, Moodley P et al. Successful integration of Tuberculosis and HIV Treatment in Rural South Africa: The Siyanq’oba study. *Acquired Immuno Deficiency Syndrome* 2009; 50:1.
 17. Lawn SD, Acheampong JW. Pulmonary tuberculosis in adults: factors associated with mortality at a Ghanaian teaching hospital. *West Afr J Med* 1999; 18(4):270-4.

18. Nsubuga P, Johnson JL, Okwera A, Mugarwa RD, Ellner JJ, Whalen CC. Gender and HIV- associated pulmonary tuberculosis: presentation and outcome at one year after beginning antituberculosis treatment in Uganda. *BMC Pulmonary Medicine* 2002; 2:4.
19. Taylan, P., Weber, G.W., and Yerlikaya, F., Continuous optimization applied in MARS for modern applications in finance, science and technology, in the ISI Proceedings of 20th Mini-EURO Conference Continuous Optimization and Knowledge- Based Technologies, Neringa, Lithuania (2008) 317-322.
20. Wood, S.N., *Generalized Additive Models, An Introduction with R*, Chapman and Hall, New-York, 2006.
21. Friedman, J.H., Multivariate adaptive regression splines, *The Annals of Statistics*, 1991, 19:1-141.
22. Yerlikaya, F., *A New Contribution to Nonlinear Robust Regression and Classification with Mars and Its Applications to Data Mining for Quality Control in Manufacturing*, 2008, Master Thesis, METU, Ankara.

APPENDIX R CODE

```
###Reading the data into R.
setwd("C:/Users/cheruiyot.erick/Desktop")
data1=read.csv(file="data2.csv",head=TRUE,sep=",")
fix(data1)
summary=summary(data1)
write.csv(summary, file='summary_original.csv')

###Dividing the data into different treatment outcome data frames.
DEAD<-subset(data1,data1$TreatOutcome=="DEAD")
fix(DEAD)
write.csv(DEAD,file='Dead.csv')
FAILED<-subset(data1,data1$TreatOutcome=="FAILED")
fix(FAILED)
write.csv(FAILED,file='FAIL.csv')
OOC<-subset(data1,data1$TreatOutcome=="OOC")
fix(OOC)
write.csv(OOC,file='OOC.csv')
TRANSFEROUT<-subset(data1,data1$TreatOutcome=="TRANSFER OUT")
fix(TRANSFEROUT)
write.csv(TRANSFEROUT,file='Transfer out.csv')
TREATMENTSUCCESS<-subset(data1,data1$TreatOutcome=="TREATMENT SUCCESS")
fix(TREATMENTSUCCESS)
write.csv(TREATMENTSUCCESS,file='Treat_Success.csv')

###Creating a factor to get the 2000 records of FAILED
bb <- (2000/27)

lb <-length(data1$TreatOutcome[data1$TreatOutcome=="FAILED"])
lb

###2000 Failed
pb=round(bb*lb)
pb

### Take a random sample of fails
set.seed(1)
sample.rows.rep<-sample(x=nrow(FAILED), size=pb, replace=TRUE)
FAIL_sample <- FAILED[ sample.rows.rep , ]
fix(FAIL_sample)
write.csv(FAIL_sample,file='Failed.csv')

### Creating a factor to get the 2000 records of DEAD
bb <- (2000/212)
```

```

lb <-length(data1$TreatOutcome[data1$TreatOutcome=="DEAD"])
lb

###2000 deads
pb=round(bb*lb)
pb

### Take a random sample of deads
set.seed(1)
sample.rows.rep<-sample(x=nrow(DEAD), size=pb, replace=TRUE)
Dead_sample <- DEAD[ sample.rows.rep , ]
fix(Dead_sample)
write.csv(Dead_sample,file='Dead.csv')

### Creating a factor to get the 2000 records of OOC
bb <- (2000/253)

lb <-length(data1$TreatOutcome[data1$TreatOutcome=="OOC"])
lb

### 2000 oocs
pb=round(bb*lb)
pb

### Take a random sample of OOCs
set.seed(1)
sample.rows.rep<-sample(x=nrow(OOC), size=pb, replace=TRUE)
ooc_sample <- OOC[ sample.rows.rep , ]
fix(ooc_sample)
write.csv(ooc_sample,file='OUT OF CONTROL.csv')

### Creating a factor to get the 2000 records of TRANSFER OUT
bb <- (2000/216)

lb <-length(data1$TreatOutcome[data1$TreatOutcome=="TRANSFER OUT"])
lb

### 2000 transfers
pb=round(bb*lb)
pb

### Take a random sample of fails
set.seed(1)
sample.rows.rep<-sample(x=nrow(TRANSFEROUT), size=pb, replace=TRUE)
transfer_sample <- TRANSFEROUT[ sample.rows.rep , ]

```

```

fix(transfer_sample)
write.csv(transfer_sample,file='TRANSFER OUT.csv')

#####
setwd("C:/Users/cheruiyot.erick/Desktop")
regression_dataset=read.csv(file="regression_data.csv",head=TRUE,sep=",")
colnames(regression_dataset)
fix(regression_dataset)
s=summary(regression_dataset)
write.csv(s, file='summary_regression data.csv')
#data <- regression_dataset()
training_set <- sample(11897,7972)
train <- regression_dataset[training_set,]
write.csv(train,file='data.csv')
test <- regression_dataset[!(1:11897)[-training_set],]
write.csv(test,file='test.csv')
train1=read.csv(file="data.csv")
Summary(train1)
Summary(test)
#####
library(earth)
model <- earth(TreatOutcome~.,# the formular for prediction
  train, # the training dataset
  trace=1, # provide overview information during model building
  nk=20, # the maximum number of terms
  nfold=10,
  ncross=10,
  stratify=TRUE,
  keepxy=T,
  degree=2, # the maximum number of interaction (degrees of freedom)
  penalty=2, # penalty per knot for GCV during pruning
  thresh=0.001, # minimum change in SSR in forward stage
  minspan=2, # minimum distance between knots in the model
  fast.k=0, # disable Fast MARS adding multiple terms per forward step
  fast.beta=0, # aging coefficient used in Fast MARS
  nprune=NULL,
  pmethod="backward") # pruning method during backward pass
head(training_set)
#summarize the model
summary(model)
summary(model, decomp = "anova",style = "bf") #c("h", "pmax", "max", "bf", di-
digits=model$digits, fixed.point=TRUE)
#summary(model, decomp = "anova",style = "pmax") #c("h", "pmax", "max", "bf", di-
digits=model$digits, fixed.point=TRUE)
summary(model, decomp = "anova",style = "h") #c("h", "pmax", "max", "bf", di-
digits=model$digits, fixed.point=TRUE)
summary(model, decomp = "anova",style = "max") #c("h", "pmax", "max", "bf", di-
digits=model$digits, fixed.point=TRUE)

```

```

summary(model, decomp = "none")
cat(format(model,digits = getOption("digits"), use.names = TRUE,
decomp = "anova", style = "bf", colon.char = "*"))
#summarize the importance of input variables
ev=evimp(model)
plot(ev)
print(ev)

evimp(model,trim=FALSE)
# plot diagnostics of the model
plot(model) # plots all the four graphs
plot(model, which=1, col.rsq=0)
plot(model, which=2, col.rsq=1)
plot(model, which=3, col.rsq=0)
plot(model, which=4, col.rsq=2)
# plot the line of best fit for the training data
plotmo(model, nresponse=1, clip=FALSE)
plotmo(model, nresponse=2, ylim=NULL, clip=FALSE)
plotmo(model, nresponse=3, ylim=NULL, clip=FALSE)
plotmo(model, nresponse=4, ylim=NULL, clip=FALSE)
plotmo(model, nresponse=5, ylim=NULL, clip=FALSE)
#case.names(model)
extractAIC(model, warn=FALSE)
model.matrix(model)
update(model)

#####
# make predictions for the test data
predictions <- predict(model, newdata=test, type="class")
write.csv(predictions, file='predictions.csv')
# plot the line of best fit for the training data
plotmo(model, nresponse=1, clip=FALSE)
plotmo(model, nresponse=2, ylim=NULL, clip=FALSE)
plotmo(model, nresponse=3, ylim=NULL, clip=FALSE)
plotmo(model, nresponse=4, ylim=NULL, clip=FALSE)
plotmo(model, nresponse=5, ylim=NULL, clip=FALSE)

#case.names(model)
extractAIC(model, warn=FALSE)
model.matrix(model)
update(model)

###update(model)
head(resid(model, warn=TRUE)) # earth residuals, a column for each response
head(resid(model, type="earth")) # same
head(resid(model, type="deviance")) # GLM deviance residuals, a column for each response
variable.names(model)
resid(model)

```

```
print(model)
deviance(model)
model.matrix(model)
```

```
#####
plotd(model, hist = FALSE, nresponse = NULL, dichot = FALSE,
trace = 2, xlim = NULL, ylim = NULL, jitter = FALSE, main=NULL,
xlab = "Predicted Value", ylab = "Count",
lty = 1, col = c("gray70", 1, "lightblue", "brown", "pink", 2, 3, 4),
fill =1,breaks = "Sturges", labels = TRUE,
kernel = "gaussian", legend = TRUE, legend.names = NULL, legend.pos = NULL,
cex.legend = .8, legend.bg = "white", legend.extra = TRUE,
vline.col = 0, vline.thresh = .5, vline.lty = 1, vline.lwd = 1,
err.thresh = vline.thresh, err.col = c(2,3,4,5,6), err.border = 0, err.lwd = 1,
xaxt = "s", yaxt = "s", xaxis.cex = 1, sd.thresh = 0.01)
```

```
#####Misclassification matrix#####
setwd("C:/Users/cheruiyot.erick/Desktop")
data1=read.csv(file="pred.csv",head=TRUE,sep=",")
fix(data1)
w=table(data1$observed,data1$predicted)
s=diag(1-prop.table(table(data1$observed, data1$predicted), 1))
prop.table(s)
prop.table(w, 1)
e=diag(1-prop.table(w, 1))
write.csv (e, file='misclassification.csv')
```