Use of CART and Logistic Regression Analysis to Identify Key Determinants of Pregnancy Wastage

Austin Luki Mueke

(W62/60012/2010)

Institute of Tropical and Infectious Diseases (UNITID)

College of Health Sciences

University of Nairobi

Presented in partial fulfilment for the requirements

For the degree of

Master of Science in Medical Statistics

University of Nairobi,

November 2013
DECLARATION

This project is my original work and has not been presented for a degree in any other University.

Signature______________________________     Date__________________________

AUSTIN LUKI MUEKE

This project has been submitted for examination with my approval as the University Supervisor:

Signature______________________________     Date__________________________

ANNE WANG’OMBE
ABSTRACT

This master research project compares the performance of Classification and Regression Trees (CART) and Logistic Regression in studying determinants of pregnancy wastage using pregnancy information from a population-based sample survey, The Kenya Demographic and Health Survey 2008/2009. The project report also describes in detail the fundamental principles of tree construction, splitting algorithms and pruning procedures. It also briefly introduces the logistic regression and then shows the comparisons of the analysis results from the two statistical methods using Receiver Operating Curve, Variable Importance and Hosmer-LemeshowModel Goodness of Fit Tests. Logistic regression performed slightly better than CART using AUC with both agreeing on age of the woman as the most important determinant of pregnancy wastage. CART found that the age of the woman, highest level of educational attainment, age at first birth, Type of place of residence being either ourban or rural and birth order to be the most important determinants of pregnancy wastage. Logistic regression analysis found out that Age of the woman, marriage to first birth interval, usage of anti-malarial during pregnancy, type of place of residence and usage of iron supplementation during pregnancy to be the most important determinants. The Hosmer-Lemeshow Goodness of Fit Test showed that CART didn’t fit the well the data while the Hosmer-Lemeshow Goodness of Fit Test for logistic regression showed that did fit the data well. The lack of close fit for the data could be explained by the nature of data and this needs further investigation comparing fits both population based data and obstetric data. However, CART results could be used for selection of key variables to be used in logistic regression analysis. When applied prudently, both CART and logistic regression are suitable for the analysis of the determinants of pregnancy wastage.
ACKNOWLEDGEMENTS

I owe a lot of thanks to many who supported during the entire masters course and research project.

My utmost gratitude goes to my research project supervisor, Anne Wang’ombe for her supervision, advice, guidance and humane kindness.

My sincere appreciation goes to the Director, Deputy Director and Staff of the Institute of Tropical and Infectious Diseases, UNITID. Thank you for always accepting my requests for help in the duration of the course.

Special thanks also to my classmates who encouraged me in many ways.
DEDICATION

I dedicate this master’s degree research project to my beloved wife, SusanVataKisamwa for her unwavering support and encouragement all through. I specifically acknowledge the time you spend at the HMIS department of The National Spinal Injury Hospital at Hurligham, Nairobi, as you went through dusty patient files as we unsuccessfully tried to assemble a dataset for the Masters Research Project. I appreciate all the sacrifice you made in many ways. May God, The Holy One Of Israel, Jehovah God, Bless You. Thank You Always.
# TABLE OF CONTENTS

## CHAPTER ONE

### INTRODUCTION

1.1 Background .............................................................................................................. 1  
1.2 Statement of Hypothesis and Research Questions .............................................. 2  
1.3 Significant of the Study ..................................................................................... 2  
1.4 Justification of the study ................................................................................... 3  
1.5 Definitions .......................................................................................................... 3  
1.6 Scope, Limitations and Assumptions ................................................................... 4

## CHAPTER TWO

### LITERATURE REVIEW

2.1 Introduction .......................................................................................................... 5  
2.2 Pregnancy Wastage ............................................................................................. 5  
2.3 Statistical analysis methods and pregnancy wastage ......................................... 7  
2.4 Summary of Literature Review .......................................................................... 8

## CHAPTER THREE

### METHODS

3.1 CART Model ......................................................................................................... 9  
3.2 Regression Tree ................................................................................................. 14  
3.3 Logistic Regression ............................................................................................. 24
CHAPTER FOUR
RESULTS AND DISCUSSIONS

4.1 Bivariate Analysis………………………………………………………………25
4.2 Classification and Regression Tree…………………………………………25
4.3 Logistic Regression……………………………………………………………..31
4.4 CART and Logistic Regression……………………………………………..35
4.4.1 Receiver Operating Characteristic Analysis……………………………35
4.4.2 Hosmer-Lemeshow Analysis……………………………………………..38

CHAPTER FIVE
CONCLUSION

Conclusions……………………………………………………………………………40
References …………………………………………………………………………….42
List of Figures

Table 4.2.1  Detail of Trees Generated using CART with Relative Error Complexities….26
Table 4.2.2  CART variable importance of risk factors for pregnancy wastage……………27
Figure 4.2.1  Splitters from the tree generated using CART…………………………….....28
Figure 4.2.2  Classification Tree for Pregnancy Wastage……………………………………29
Figure 4.2.3  The tree sequence of lowest complexity which yield 13 nodes with cross
validation error rate…………………………………………………………………..30
Table 4.3.1  Variable importance in multivariable logistic regression…………………..31
Table 4.3.2  Results of multivariate logistic regression……………………………………32
Figure 4.4.1  CART Navigator 1(7 nodes) – Summary results Gains CHART – ROC.,…..35
Figure 4.4.2  ROC curve for Logistic Regression…………………………………………36
Table 4.4.1  Comparison of CART and Logistic Regression Variable Importance……….37
Table 4.4.2  Logistic Regression Hosmer-Lemeshow Model Goodness of Fit Test………38
Table 4.4.3  CART Hosmer-Lemeshow Model Goodness of Fit Test…………………..38
### ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CART</td>
<td>Classification and Regression Tree</td>
</tr>
<tr>
<td>ROC</td>
<td>Receive Operating Characteristic</td>
</tr>
<tr>
<td>AUC</td>
<td>Area Under Curve</td>
</tr>
</tbody>
</table>
CHAPTER ONE: INTRODUCTION

1.1 Background

Pregnancy is a female state that is produced due to the implantation of the fertilised ovum in
the uterine endometrium and ultimately gives rise to a foetus; and pregnancy wastage is the
loss of product of conception normally or therapeutically (Jeffcoate, 1975). Pregnancy
wastage is one of the adverse birth outcomes among others including preterm births and low
birth weights. Adverse birth outcomes are far more frequent in the developing world because
of the many unplanned pregnancies, pregnancy and birth complications whose repercussions
are generally unfavorable both for the mother and the baby. They remain significant
contributors to perinatal mortality and developmental disabilities globally. Abu-Saad and
Fraser, (2010) concluded that adverse birth outcomes carry lifelong consequences for
development, life quality and health care costs.

Pregnancy wastage can be classified as intra-uterine foetal death, abortion and menstrual
regulation (Jeffcoate 1195, and Shaw, Scouther and Stanton 2003).

Every year about eight million women suffer from pregnancy related complications and over
half a million die (Atikur et al). About 99% of them are in developing countries (WHO,
2004).

Generally pregnancy wastage into one of the four categories: miscarriages, stillbirth, birth
loss and medically based termination.

It is estimated that more than 3.3 million babies are stillborn every year. Worldwide 1.2
million stillbirths occur during labour (intrapartum) according to The Lancet. Before labour
(antepartum) stillbirths account for more than half of stillbirths (1.4 million). The five big
causes of stillbirth are childbirth complications, maternal infections in pregnancy, maternal
disorders, especially hypertension and diabetes, fetal growth restriction and congenital
abnormalies. The priority programme investments include family planning, care at birth,
antenatal care with hypertension and advanced antenatal care. In 2005 WHO gave a global
estimate of 3.3 million stillbirth deaths.

In Kenya a study on poor pregnancy outcomes by Magadi2006 among adolescents in South
Nyanza region of Kenya found significant associations between socio-economic and
demographic characteristics with pregnancy outcomes with place being rural or urban and the level of education attainment being strongly associated with adverse pregnancy outcomes.

The purpose of this study is to determine risk factors associated with the pregnancy wastage as an adverse birth outcome, the normal pregnancy wastage which is not therapeutic.

1.2 Statement of the Hypothesis and Research Questions

Based on 2008/09 KDHS, there were 6,079 pregnancies reported in Kenya out of which 1,370 ended preterm resulting in pregnancy wastage. The question is what are the major risk factors determining the occurrence of pregnancy wastage?

1.2.1 Hypotheses

- To determine if maternal age at first pregnancy is a determinant of pregnancy wastage.
- To establish whether a woman’s household wealth index is a determinant of pregnancy wastage.
- To examine whether the place of residence is a determinant of pregnancy wastage.
- To compare CART and Logistic Regression Analysis model performances for identifying key determinants of pregnancy wastage.

1.2.2 Research Questions

- Is maternal age at first pregnancy a risk factor in pregnancy wastage?
- Is Woman’s household wealth index a risk factor for pregnancy wastage?
- Is the place of residence a risk factor for pregnancy wastage?

1.3 Significant of the Study

Risk factors for pregnancy wastage have not been studied extensively in Kenya. One documented study was done poor pregnancy outcomes among adolescents in South Nyanza region of Kenya(Magadi, 2006).

Pregnancy outcome is influenced by hereditary and environmental factors including those which affect stature in early life, current health and nutritional status, inter-pregnancy interval, maternal age, genitourinary or general diseases in women and socioeconomic and educational status.
This study found that adolescents living in rural areas, not enrolled in school or with low educational attainment had higher proportion of experiencing pregnancy wastage compared to their counterparts who were living in urban areas, had higher educational attainment or enrolled in school at the time of index pregnancy among other associations. Overall the patterns of associations between socio-economic and demographic characteristics with pregnancy outcomes observed in the bivariate and multivariate analyses conform to what is expected but the issue is that these associations are not statistically significant possibly due to the relatively small number of cases analysed and hence insufficient power to detect statistical significance.

Also the study only covered a region and such there is no study to at national level to look at the trends of pregnancy wastage or the issues which aggravate the situation.

This study will use two classification techniques, one parametric-discriminant function analysis and the other non-parametric-classification and regression trees in order to provide a non-subjective risk analysis of pregnancy wastage.

1.4 Justifications Of the Study

The findings of the study shall also enable program developers and implementers to know profiles of women with enormous need for reproductive health services at population level, including treatment of reproductive tract infections and malaria prevention so as to check on adverse birth outcomes which may not be captured at health facility level because program data may not. The risk factors for pregnancy wastage in Kenya have therefore not been studied extensively and there is no national study that has been done to establish socio-demographic and health risk factors for pregnancy wastage using population based data. As such therefore, there is need to have information the effect of various socio-demographic and health factors on the risk of pregnancy outcome in Kenya concerned with family planning services to trace areas of need and advise people on the need to check high-risk pregnancies and unsafe abortions so as to reduce the number of adverse pregnancy outcomes.

1.5 Definitions

Miscarriages

Any pregnancy that ends unintentionally before the foetus is viable.
Preterm births

A Preterm birth is defined as birth of less than 37 weeks' gestation pregnancy.

Still births

WHO promotes the definition of stillbirth or late fetal death as death occurring at least 28 weeks of gestation or at least 1000g birth weight(3). This means its babies born within the last trimester of pregnancy.

Pregnancy Wastage

Pregnancy is a female that is produced due to implantation of the fertilized ovum in the uterine endometrium and ultimately giving rise to a foetus. Pregnancy wastage is the loss of product of conception normally or therapeutically (Jeffcoate, 1975). Pregnancy wastage in this case means stillbirths, abortions and miscarriages.

1.6 Scope, Limitations and Assumptions

The study will utilise the 2008-09 Kenya Demographic and Health Survey (KDHS 2008/09) which is a nationally representative sample survey of 8,444 women aged 15 to 49 and men aged 15 to 54 selected from 400 sample points(clusters) throughout Kenya. It is designed to provide data to monitor the population and health situation.

The survey utilised a two-stage sample based on the 1999 Population and Housing Census and was designed to produce separate estimates for key indicators for each of the eight provinces in Kenya. Data collection took place over a three month period from 13th November 2008 to late February 2009.

Obstetric data is not available for this study, however, the population based approach gives a real picture of what is happening country wide. Furthermore, the KDHS data is statistically representative of the country as opposed to specific facility data which may be biased based on the profile of the people coming for service based on the geographical and spatial distribution, socio-economic and other factors, human and natural.

The authorization to use the dataset was sought from both MEASURE DHS and The Kenya National Data Archive (KeNADA).
CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

This chapter reviews various literatures on factors associated with pregnancy wastage. It especially points socio-economic, demographic and health factors found out to determine occurrence of pregnancy wastage. It will also do a review of literature on the methods for studies done on methods Presentation of conceptual framework and operational framework will follow, and then hypotheses and definition of key variables will be discussed. Reviews of this literature will cite what is already known about pregnancy wastage as well as spell out the gaps from different studies.

2.2 Pregnancy Wastage

Sidhu and Sidhu(1) examined pregnancy wastage and found that the poor living conditions contributed to higher pregnancy wastage in the scheduled caste women of Punjab. They could not attribute the very high rate if pregnancy wastage to a particular cause but envisaged that illiteracy, lower socio-economic status and non-availability of medical care may be contributing factors. They did note an increasing rate of foetal deaths with increasing age just as was observed by other investigators including Yerushalmy et al. 1956, Potter et al. 1965 and Nortman, 1974.

Atikur et al (2) studying the velocity and elasticity of pregnancy wastage and caesarian deliveries in Bangladesh on the contrary found out that pregnancy wastage decreases over ages where caesarian delivery increases. In looking at the two, they found out that increased age increase the risk of caesarian delivery but decrease the risk of pregnancy wastage. They however did agree that in the extreme age groups, pregnancy wastages are observed substantially larger.

Prakash et al (3) using data from the third wave of National Family Health Survey (NFHS, 2005-2006) for India to examine the effects of early marriage on the reproductive health status of women and on the well-being of their children found that early marriage had detrimental effects on the reproductive health status of women so that women who married at an early age were exposed to frequent childbearing, unplanned motherhood and abortions. Also in relation to marriage, Sureender et al (4) in their analysis of the the same data for Tamil Nadu, 1992 revealed that women marrying their close relatives had low age at
marriage and experienced a higher percent of pregnancy wastage and child loss (first child) as compared to those women marrying their distant relatives or nonrelatives.

In a longitudinal study for the period 1988-92, Agarwal et al (5) looking at the relationship between pregnancy wastage and maternal under-nutrition and other socio-demographic factors in rural Indian women found no differences in abortion and stillbirth rates during the study years. However, an increase in haemoglobin showed consistent reduction in abortion ratios. Stillbirths showed significant relationship with maternal weight and height. In this study, risk factors for increased perinatal mortality (stillbirths and neonatal deaths in the first week) were illiteracy, birth interval, previous stillbirth, previous preterm, untrained birth attendant and birth weight.

In Kenya a study on poor pregnancy outcomes by Magadi (2006) among adolescents in South Nyanza region of Kenya found significant associations between socio-economic and demographic characteristics with pregnancy outcomes with place being rural or urban and the level of education attainment being strongly associated with adverse pregnancy outcomes. A similar study among the adolescents in Bangladesh by Rahman et al 2010 on adolescent pregnancy compilation and wastage found out that young adolescents aged under 20 years were observed to have highest proportions of delivery complications and pregnancy wastage.

A study done in Cameroon using 2004 CDHS on spousal violence on potentially preventable single and recurrent spontaneous fetal loss found out that spousal violence increases the likelihood of single or repeated fetal loss. A large proportion of the risk for recurrent fetal mortality is attributable to spousal violence, and therefore is potentially preventable. The findings found out that Cameroonian women exposed to spousal violence are 50% more likely to experience single or repeated episodes of spontaneous fetal loss. Intimate partner violence within the household among women was associated with a third of reported fetal mortality in women.

A prospective study of 84 pregnant women done by Sundari (1993) looking at the effect of socio-demographic factors found out that the highest proportion of negative outcomes was to women who were pregnant for the first time with 12 percent of all pregnancies ending up wastage. In this study, it was found out that negative pregnancy outcome is lower for lower parities. However, in the same study, it was found out that the history of difficult and complicated deliveries were ages 20-24 and over 35 years suffered the highest proportion of negative outcome. On the contrary, in a study done by Ibrahim on a community-based
prospective study of 6275 women studying the occurrence of stillbirths, it was found out that the correlation of the outcome of the last pregnancy with that of the current pregnancy was highly significant.

Data collected from 750 women of the reproductive ages (15-49) in the rural Rajshahi district of Bangladesh by Mahfuzar et al in a purposive sample found out that increased age increases the risk of pregnancy wastage. The pregnancy wastage in the age group 25-29 was lower among the women, followed by the age-group 40-44 which was the highest. The proportion of pregnancy wastage to live births was high in the age group 15-19.

According to Priyali and Umesh, various studies document the relationships between lowered zinc concentrations during pregnancy and low birth weight so that there is a threshold for serum zinc concentration below which adverse pregnancy outcome increases significantly.

Ibrahim et al in a community-based prospective study on stillbirths found out that the outcome of last pregnancy with that of the current pregnancy was highly significant.


2.3 Statistical Analysis Methods Review and Pregnancy Wastage

In review of statistical analysis methods review for pregnancy wastage, Amina et al (2009) using Cameroon DHS, studied single and recurrent fetal loss as the dependent variable with physical, sexual and emotional as independent variable with individual level of analysis used Chi-square test to measure the differences in maternal and socio-demographic characteristics between the 2 groups(violence and no violence). Logistic regression model was used to generate adjusted odds ratio and 95% C.Is. Generalised estimating equation to account for intraclass correlation.

Rahman et al (2010) studied reproductive complications leading to pregnancy wastage using Micro level survey of 400 adolescents (10-19) and indepth interview with 37 adolescents who had experienced pregnancy wastage. Micro level survey of 400 adolescents (10-19) and in-depth interview with 37 adolescents who had experienced pregnancy wastage. Logistic regression was used to estimate the relative risks of the predictors of higher proportions of women suffer pregnancy problems especially in cases of early conception.

Ibrahim et al used Linear Logistic regression were fitted to calculate odds ratios of Outcome of last pregnancy with that of the current pregnancy. Stratification was used based on the mid
wife. Mahfuzar Rahman used Logistic regression model to study pregnancy wastage using Data collected from 750 women of reproductive ages (15-49) in the rural Rajshahi district, Bangladesh.

Using 2001 Census data for Madya Pradesh, India, Diamond-Smith et al studying fetal losses and maternal deaths using Monte Carlo Sensitivity analysis to assess the plausibility of the estimates of Fetal losses and maternal deaths.

Monica Magadi (2004) used Bivariate and multinomial regression analysis to study pregnancy outcome including pregnancy wastage, premature live birth and full time live births.

Z proportion test, Relative Risk and Multiple Regression were performed on a Follow up study between January 1988 to December 1992 by Agarwal et al to study pregnancy wastage.

2.4 Summary of Literature review

These studies provide some evidence that a relationship exists between pregnancy wastage and various factors including poor living conditions, lower socio-economic status, non-availability of medical care, extreme ages, early marriage and maternal under nutrition. Other factors including maternal under nutrition, birth interval, spousal violence and parity are also associated with higher incidences of pregnancy wastage.

Various methods have also been used to study pregnancy wastage including logistic regression, generalised estimating equation, linear logistic regression, bivariate and multinomial regression and various other methods.

An investigation is needed to establish in depth the risk profiles of pregnancy wastage in the Kenyan situation.
CHAPTER THREE: METHODS

3.1 CART Model

Classification and Regression Tree is a type of decision tree introduced by Leo Breiman et al in 1984. In mathematical terms, a decision tree is defined as a directed, acyclic and connected graph having one distinguishable vertex called a root node. The tree structure consists of nodes and branches connecting these nodes. If a node has branches leading to other nodes, it is called a parent node, and the nodes to which these branches lead to are called children of this node. The terminal nodes are called leaves.

As indicated CART methodology was developed in 80s by Breiman, Freidman, Olshen, Stone in their paper “classification and Regression Trees” (1984). For building decision trees, CART use so-called learning sample- a set of historical data with pre-assigned classes for all observations. For example, learning sample for credit scoring system would be fundamental information about previous borrowes (variables) matched with actual payoff results (classes).

Decision trees are represented by a set of questions which splits the learning sample into smaller and smaller parts. CART asks only yes/no questions. A possible question could be: “Is age greater than 50?” or “is sex male?” CART algorithm will search for all possible variables and all possible values in order to find the best split-the question that splits the data into two parts with maximum homogeneity. The process is then repeated for each of the resulting data fragments. Here is an example of simple classification tree, used by San Diego Medical Center for classification of their patients to different levels of risk:
Classification Tree

The characteristic feature of CART is that the tree constructed by CART algorithm is strictly binary. The decision tree is constructed recursively where an attribute is selected at first to be placed at the root of the node to make one branch for each possible value. This splits up the example set into subsets, one for every value of the attribute (Witten, Frank; 2000). The cases from the training set are recursively portioned into subsets with similar values of the target variable and the tree is built through the thorough search of all available variables and all possible divisions for each decision node and the selection of the optimal divisions according to a given criterion.

The basic principle of the tree model is to partition space spanned by the input variables to maximise a score of class purity and that the majority of the points in each cell of the partition belong to one class. They are mappings of observations to conclusions (target values). Each inner node responds to a variable; an arc to a child represents a possible value of that variable. A leaf represents the predicted value of target variable given the values of the variables represented by the path from the root (T. Menzies, Y. Hu; 2003).

Splitting Criterion

The splitting criterions have always the following form. The case is moved to the left child if the condition is met, and goes to the right child if otherwise.

For continuous variables, the condition is defined as “explanatory variable $x_j \leq C$”. For nominal variables, the condition expresses the fact the variables takes on specific values.

The splitting criterion measures the decrease in the distribution of the target variable between the root node and the first subsequent node and between each pair of subsequent nodes. The splitting criterion has two objectives: to determine the best split for each input variable and to choose the best split among a multitude of possible splits from input variables. For classification trees, there are two common splitting criterions, which are Gini index and entropy reduction.
Gini Index (Total Leaf Impurity)

Gini index is a measure of purity. If the target values are the same within the node then the Gini index is one. The best split is selected by the largest Gini index. The Gini index can be calculated as follows:

\[
\text{Gini index} = E_G(m) = \sum_{\text{leaves}} p_{\text{leaves}} \left( \sum_{\text{classes}} p_i^2 \right)
\]

- \( E_G(m) \) is the Gini index at node \( m \).
- \( p_i \) is the proportion of each target class in the \( i^{\text{th}} \) node.
- \( p_{\text{leaves}} = \frac{\text{number of individuals in the leaf node}}{\text{total number of individuals in node } m} \)
  (Matiganon, 2007)

Entropy reduction

Entropy reduction is a measure of variability in categorical data. If the target values are the same within the node then the entropy is zero. Thus, the best split is selected by smallest entropy reduction. The Entropy reduction is defined as:

\[
\text{Entropy reduction} = E_E(m)
\]

\[
= -1 \cdot \sum_{\text{leaves}} p_{\text{leaves}} \left( \sum_{\text{classes}} p_i \cdot \log_2 p_i \right)
\]

- \( p_i \) is the proportion of each target class in corresponding node \( m \). (Matignon, 2007).
**Construction of maximum tree**

Let \( t_p \) be a parent node and \( t_l, t_r \) – respectively left and tight child nodes of parent node \( t_p \).

Consider the learning sample with variable matrix \( X \) with \( M \) number of variables \( x_j \) and \( N \) observations. Let class vector \( Y \) consist of \( N \) observations with total amount of \( K \) classes.

Classification tree is built in accordance with splitting rule-the rule that performs the splitting of learning sample into smaller parts. Each time data have to be divided into two parts with maximum homogeneity:

\[
\begin{align*}
\text{if } x_j \leq x_j^R & \\
\text{then } & \\
& t_{Left} & t_{Right} \\
& P_{Left} & P_{Right}
\end{align*}
\]

\[ X_j \leq x_j^R \]

Figure 2.1: Splitting algorithm of CART

where \( t_p, t_l, t_r \) – parent, left and right nodes; \( x_j \) – variable \( j \); \( x_j^R \) – best splitting value of variable \( x_j \).

Maximum homogeneity of child nodes is defined by so-called impurity function \( i(t) \). Since the impurity of parent node \( t_p \) is constant for any of the possible splits \( x_j \leq x_j^R, j = 1, \ldots, m \), the maximum homogeneity of left and right child nodes will be equivalent to the maximization of change of impurity function \( \Delta i(t) \):

\[
\Delta i(t) = i(t_p) - E\{i(t_c)\}
\]

where \( t_c \) – left and right child nodes of the parent node \( t_p \). Assuming that the \( P_l, P_r \)-probabilities of right and left nodes, we get:
\[ \Delta i(t) = i(t_p) - P_1 i(t_1) - P_1 i(t_r) \]

Therefore, at each node CART solves the following maximization problem:

\[ \arg\max \left[ i(t_p) - P_1 i(t_1) - P_1 i(t_r) \right] \]

Equation 2.1 implies that CART will search through all possible values of all variables in matrix X for the best split question \( x_j < x_j^R \) which will maximize the change of impurity measure \( \Delta i(t) \).

The next important question is how to define the impurity function \( I(t) \). In theory there are several impurity functions, but only two of them are widely used in practice: Gini splitting rule and Twoing splitting rule.

**Gini splitting rule**

Gini splitting rule (or Gini index) is most broadly used rule. It uses the following impurity function \( I(t) \):

\[ i(t) = \sum_{k \neq 1} P \left( \frac{K}{t_p} \right) p (l/t) \]

where \( k, l, \ldots \ldots \ldots \), \( k \), -index of the class; \( p(k/t) \)-conditional probability of class \( k \) provided we are in node \( t \).

Applying the Gini impurity function 2.2 to maximization problem 2.1 we will get the following change of impurity measure \( \Delta i(t) \):

\[ \Delta i(t) = - \sum_{k=1}^{K} p^2 \left( \frac{K}{t_p} \right) + P_1 \sum_{k=1}^{K} p^2 \left( \frac{K}{t_1} \right) + P_r \sum_{k=1}^{K} p^2 \left( \frac{K}{t_r} \right) \]

Therefore, Gini algorithm will solve the following problem:

\[ \arg\max \]

\[ x_j \leq x_j^R, \quad j = 1, \ldots, m \left[ - \sum_{k=1}^{K} p^2(k/t_p) + P_1 \sum_{k=1}^{K} p^2 \left( \frac{K}{t_1} \right) + P_r \sum_{k=1}^{K} p^2 \left( \frac{K}{t_r} \right) \right] \]
Gini algorithm will search in learning sample for the largest class and isolate it from the rest of the data. Gini works well for noisy data.

3.2 Regression tree

Regression trees do not have classes. Instead there are response vector \( Y \) which represents the response values for each observation in variable matrix \( X \). Since regression trees do not have pre-assigned classes, classification splitting rules like Gini 2.3 or Twoing 2.4 cannot be applied.

Splitting in regression trees is made in accordance with squared residuals minimization algorithm which implies that expected sum variances for two resulting nodes should be minimized. 

\[
\text{arg min} \left[ P_l \text{var}(Y_l) + P_r \text{var}(Y_r) \right]
\]

\[
x_j \leq x^R_j, j = 1, \ldots, m \left( P_l \text{var}(Y_l) + P_r \text{var}(Y_r) \right)
\]

where \( \text{Var}(Y_l), \text{Var}(Y_r) \) — response vectors for corresponding left and right child nodes; 

\( x_j \leq x^R_j, j = 1, \ldots, m \) — optimal splitting question which satisfies the condition 2.5

Squared residuals minimization algorithm is identical to Gini splitting rule. Gini impurity function 2.2 simple to interpret through variances notation. If we assign to objects of class \( k \) the value, 1, and value 0 to objects of other classes, then sample variance of these values would be equal to \( p(k/t)(1-p(k/t)) \). Summarizing by number of classes \( k \), we will get the following impurity measure \( I(t) \):

\[
i(t) = 1 - \sum_{k=1}^{k} p^2 \left( \frac{k}{t} \right)
\]

Up to this point so-called maximum tree was constructed which means that splitting was made up to the last observations in learning sample. Maximum tree may turn out to be very big, especially in the case of regression trees, when each response value may result in a separate node.
**Recursive partitioning**

Up to this point the classification problem has been used to define and motivate our formulae. However, the partitioning procedure is quite general and can be extended by specifying 5 “ingredients”.

- A splitting criterion, which is used to decide which variable, gives the best split. For classification this was either the Gini or log–likelihood function. In the anova method the splitting criteria are $SS_T = (SS_L + SS_R)$, where $SS_T = \sum(y_i - \bar{y})^2$ the sum of squares for the node is, and $SS_R, SS_L$ are the sums of squares for the right and left son, respectively. This is equivalent to choosing the split to maximize the between-groups sum-of-squares in a simple analysis of variance. This rule is identical to the regression option for tree.
- A summary statistics or vector, which is used to describe a node. The first element of the vector is considered to be fitted value. For the anova method this is the mean of the node; for classification the response is the predicted class followed by the vector of class probabilities.
- The error of a node. This will be the variance of $y$ for anova, and the predicted loss for classification.
- The prediction error for a new observation, assigned to the node. For anova this is $(Y_{new} - y)$.
- Any necessary initialization.

The anova method leads to regression trees; its is the default method if $y$ a simple numeric vector, i.e, not a factor, matrix, or survival object.

**Notation**

The partitioning method can be applied to many different kinds of data. We will start by looking at the classification problem, which is one of the more instructive cases (but also has the most complex equations). The sample population consists of $n$ observations from $C$ classes. A given model will break these observations into $k$ terminal groups; to each of these groups is assigned a predicted class (this will be the response variable). In an actual application, most parameters will be estimated from the data, such estimates are given by $\approx$ formulae.
\[ \pi_i = 1, 2, \ldots, C \]  
Prior probabilities of each class.

\[ L(i, j) = 1, 2, \ldots, C \]  
Loss matrix for incorrectly classifying an I as a j. \( L(i, i) = 0 \)

A  
Some node of the tree.

Note that A represents both a set of individuals in the sample data, and, via the tree that produced it, a classification rule for future data.

\[ r(x) \]  
True class of an observation \( x \), where \( x \) is the vector of predictor variables.

\[ r(A) \]  
The class assigned to A, if A were to be taken as a final node.

\[ n_i, n_A \]  
Number of observations in the sample that are class I, number of obs in node A.

\[ n_{iA} \]  
Number of observations in the sample that are class i and node A.

\[ P(A) \]  
Probability of A (for future observations).

\[ = \sum_{i=1}^{C} \pi_i P(x \in A \mid r(x) = i) \]

\[ \approx \sum_{i=1}^{C} \pi_i n_{iA} / n_i \]

\[ R(A) \]  
Risk of A

\[ = \sum_{i=1}^{C} P(i)A L(i, r(A)) \]

where \( r(A) \) is chosen to minimize this risk.

\[ R(T) \]  
Risk of a model (or tree) T

\[ = \sum_{j=1}^{k} P(A_j)R(A_j) \]

where \( A_j \) are the terminal nodes of the tree.

If \( L(i, j) = 1 \) for all \( i \neq j \), and we set the prior probabilities \( \pi_i \) equal to the observed class frequencies in the sample then \( P(i/A) = n_{iA} / n_A \) and \( R(T) \) is the proportion misclassified.
Building the tree

Splitting criteria

If we split a node $A$ into two sons $A_L$ and $A_R$ into two sons, we will have:

$$P(A_L)R(A_L) + P(A_R)R(A_R) \leq P(A)R(A)$$

(this is proven in (1). Using this, one obvious way to build a tree is to choose that split which maximizes $\Delta R$, the decrease in risk. There are defects with this, however, as the following example shows.

Suppose losses are equal and that the data is 80% class I’s and that some trial split results in $A_L$ being 54% class 1’s and $A_R$ being 100% class 1’s. Class 1’s versus class 0’s are the outcome variable in this example. Since the minimum risk prediction for both the left and right son is $r(A_L) = r(A_R) = 1$, this split will have $\Delta R = 0$, yet scientifically this is a very informative division of the sample. In real data with such a majority, the first few splits very often can do no better than this.

A more serious defect with maximizing $\Delta R$ is that the risk reduction is essentially linear. If there were two competing splits, one separating the data into groups of 85% and 50% purity respectively, and the other into 70%-70%, we would usually prefer the former, if for no other reason than because it better sets things up for the next splits.

One way around both of these problems is to use look ahead rules; but these are computationally very expensive. Instead rpart uses one of several measurers of impurity, of diversity, of a node. Let $f$ be some impurity function and define the impurity of a node $A$ as.

$$I(A) = \sum_{i=1}^{C} f(P_i A)$$

where $P_i A$ is the proportion of those in $A$ that belong to class I for future samples. Since we would like $I(A) = 0$ when $A$ is pure, $f$ must be concave with $f(0) = f(1) = 0$.

Two candidates for $f$ are the information index $f(p) = -p \log(p)$ and the Gini index $f(p) = p(1-p)$. We then use that split with maximal impurity reduction.

$$\Delta I = P(A)I(A) - P(A_L)I(A_L) - P(A_R)I(A_R)$$
The two impurity functions are plotted in figure (2), with the second plot scaled so that the maximum for both measurers is at 1. For the two class problem the measurers differ only slightly, and will nearly always choose the same split point.

Another convex criteria not quite of the above class is towing for which

\[ 1(A) = \min (f(P_{C_1}) + f(P_{C_2})) \]

where \( C_1, C_2 \) is some partition of the C classes into two disjoint sets. If \( C=2 \) towing is equivalent to the usual impurity index for \( f \). Surprisingly, towing can be calculated almost as efficiently as the usual impurity index. One potential advantage of towing is that the output may give the user additional insight concerning the structure of the data. It can be viewed as the partition of \( C \) into two super classes which are in some sense the most dissimilar for those observations in \( A \). For certain problems there may be a natural ordering of the response categories (e.g. level of education), in which case ordered towing can be naturally defined, by restricting \( C_1 \) to be an interval \((1, 2, ..., k)\) of classes. Twoing is not part of rpart.

**Incorporating losses**

One salutatory aspect of the risk reduction criteria not found in the impurity measurers is inclusion of the loss function. Two different ways of extending the impurity criteria to also include losses are implemented in CART, the generalized Gini index and altered priors. The rpart software implements only the altered priors method.

**Generalized Gini index**

The Gini index has the following interesting interpretation. Suppose an object is selected at random from one of \( C \) classes according to the probabilities \( (P_1, P_2, ..., P_C) \) and is randomly assigned to a class using the same distribution. The probability of misclassification is

\[ \sum_i \sum_{j \neq i} P_i P_j = \sum_i \sum_j P_i P_j - \sum_i P_i^2 = \sum_i 1 - P_i^2 = \text{Gini index for } p \]
Let \( L(i,j) \) be the loss of assigning class \( j \) to an object which actually belongs to class \( i \). The expected cost of misclassification is \( \sum \sum L(i,j)P_iP_j \). This suggests defining a generalized Gini index of impurity by:

\[
G(P) = \left( \frac{1}{2} \right) \sum \sum L(i,j)P_iP_j
\]

In particular, for two-class problems, \( G \) in effect ignores the loss matrix.

### 3.2.2 Altered priors

Remember the definition of \( R(A) \)

\[
R(A) = \sum_{i=1}^{c} P_i A L(i, r(A))
\]

\[
\sum_{i=L}^{c} \pi_i L(i, r(A))(ni_A/n_Ai/n_i)(n/\eta A)
\]

Assume there exists \( \pi \) and \( L \) be such that

\[
\pi_i L(i,j) = \pi_i L(i,j) \forall i, j \in C
\]

Then \( R(A) \) is unchanged under the new losses and priors. If \( L \) is proportional to the zero-one loss matrix then the priors \( \pi \) should be used in the splitting criteria. This is possible only if \( L \) is of the form.

\[
L(i,j) = \begin{pmatrix} L_i & i \neq j \\ 0 & i = j \end{pmatrix}
\]

in which case.

This is always possible when \( C=2 \), and hence altered priors are exact for the two class problem. For arbitrary loss matrix of dimension \( C>2 \), \( rpart \) uses the above formula with \( L_i \sum_j L(?i,j) \).

A second justification for altered priors is this. An impurity index \( I(A) = \sum f(P_i) \) has its maximum at \( P_1 = P_2 = \cdots = P_c = \frac{1}{c} \). If a problem had, for instance, a misclassification loss
for class 1 which was twice the loss for a class 2 or 3 observation, one would wish $P_1 = 1/5, P_2 = P_3 = 2/5$, since this is the worst possible set of proportions on which to decide a node’s class.

The altered priors technique does exactly this, by shifting the $P_i$

To final notes

When altered priors are used, they affect only the choice of split. The ordinary losses and priors are used to compute the risk of the node. The altered priors simply help the impurity rule choose splits that are likely to be “good” in terms of the risk.

The argument for altered priors is valid for both the gini and information splitting rules.

**Pruning the tree**

We have built a complete tree, possibly quite large and/or complex, and must now decide how much of that model to retain. In forward stepwise regression, for instance, this issue is addressed sequentially and no additional variables are added when the F-test for the remaining variables fails to achieve some level $a$.

Let $T_1, T_2, \ldots, T_k$ be the terminal nodes of a tree $T$. Define

$$|T| = \text{number of terminal nodes}$$

$$\text{risk of } T = R(T) = \sum_{i=1}^{k} P(T_i)R(T_i)$$

In comparison to regression, $|T|$ is analogous to the model degrees of freedom and $R(T)$ to the residual sum of squares.

Now let $a$ be some number between 0 and $\infty$ which measures the ‘cost’ of adding another variable to the model; $a$ will be called a complexity parameter. Let $R(T_0)$ be the risk for the zero split tree. Define

$$R_a(T) = R(T) + a|T|$$

to be the cost for the tree, and define $T_a$ to be that sub-tree of the full model which has minimal cost. Obviously $T_0$ =the full model and $T_\infty =$the model with no splits at all. The following results are shown in [1].
1. If $T_1$ and $T_2$ are subtrees of $T$ with $R_a(T_1) = R_a(T_2)$, then either $T_1$ is a subtree of $T_2$ or $T_2$ is a subtree of $T$; hence either $|T_1| < |T_2|$ or $|T_2| < |T_1|$.

2. If $a > \beta$ then either $T_a = T_\beta$ or $T_a$ is a strict subtree of $T_\beta$.

3. Given some set of numbers $a_1, a_2, \ldots, a_m$; both $T_{a_1}, T_{a_2}, \ldots, T_{a_m}$ and $R(T_{a_1}), R(T_{a_2}), \ldots, R(T_{a_m})$ can be computed efficiently.

Using the first result, we can uniquely define $T_a$ as the smallest tree $T$ for which $R_a(T)$ is minimized.

2 implies that all possible values of $a$ can be grouped into $m$ intervals, $m \leq |T|$.

$$I_1 = [0, a_1]$$
$$I_2 = (a_1, a_2)$$
$$\vdots$$
$$I_m = (a_m - 1, \infty)$$

where all $a \in I_i$ share the same minimizing subtree.

**Cross-validation**

The procedure of cross validation is based on optimal proportion between the complexity of the tree and misclassification error. With the increase in size of the tree, misclassification error is decreasing and in case of maximum tree, misclassification error is equal to 0. But on the other hand, complex decision trees poorly perform on independent data. Performance of decision tree on independent data is called true predictive power of the tree. Therefore, the primary task is to find the optimal proportion between the tree complexity and misclassification error. This task is achieved through cost-complexity function:

$$R_a(T) = R(T) + a(T) - \min_T$$
where R (T)-misclassification error of the tree T; a (T)-complexity measure which depends
on T-total sum of terminal nodes in the tree a-parameter is found through the sequence of in-
sample testing when a part of learning sample is used to build the tree, the other part of the
data is taken as a testing sample. The process repeated several times for randomly selected
learning and testing samples.

Although cross-validation does not require adjustment of any parameters, this process is time
consuming since the sequence of trees is constructed because the testing and learning sample
are chosen randomly, the final tree may differ from time to time.

Cross-validation is used to choose a best value for $\alpha$ by the following steps:

1. Fit the full model on the data set

   computer $1_1, 1_2, \ldots, 1_m$

   set $\beta_1 = 0$

   $\beta_2 = \sqrt{a_1 a_2}$

   $\beta_3 = \sqrt{a_2 a_3}$

   .

   .

   $\beta_m - 1 = \sqrt{a_m - 2a_m} - 1$

   $\beta_m = \infty$

   each $\beta_i$ is a ‘typical value’ for its $L_i$

2. Divide the data set into s groups $G_1, G_2, \ldots, G_s$ each of size $s/n$. and for each group
   separately:
   
   - Fit a full model on the data set ‘everyone except $G_i$, and determine $T_{\beta_1}, T_{\beta_2}, \ldots, T_{\beta_m}$
     for this reduced data set.
   - Compute the predicted class for each observation in $G_i$, under each of the models
   - from this compute the risk for each subject.
3. Sum over the $G_i$ to get an estimate of risk for each $\beta_j$. For that $\beta$ (complexity parameter) with smallest risk compute $T_\beta$ for the full data set, this is chosen as the best trimmed tree.

In actual practice, we may use instead the I-SE rule. A plot of $\beta$ versus risk often has an initial sharp drop followed by a relatively flat plateau and then a slow rise. The choice of $\beta$ among those models on the plateau can be essentially random. To avoid this, both an estimate of the risk within one standard error of the achieved minimum is marked as being equivalent to the minimum (i.e. considered to be part of the flat plateau). The simplest model, among all those “tied” on the plateau, is chosen.

In the usual definition of cross-validation we would have taken $s=n$ above, i.e. each of the $G_i$ would contain exactly one observation, but for moderate $n$ this is computationally prohibitive. A value of $s=10$ has been found to be sufficient.
3.3 Logistic Regression

Logistic regression is used in this study to show whether an event will occur or not using a set of independent variables. Furthermore, it will be used to explain the percent of variance in the dependent variable which is explained by a specific predictor variable. This will be explained in terms of odds ratio. The logistic equation may be written as follows:

An explanation of logistic regression begins with an explanation of the logistic function, which always takes on values between zero and one:

\[ f(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}, \]

and viewing \( t \) as a linear function of an explanatory variable, we have

\[ \pi(x) = \frac{e^{(\beta_0 + \beta_1 x)}}{e^{(\beta_0 + \beta_1 x)} + 1} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}, \]

Where \( \pi(x) \) is the probability that the response \( y = 1 \)

- \( \beta_0 \) is the equation constant and
- \( \beta_1 \) is the coefficient of the predictor \( X_1 \)

The advantage of a logistic regression model is that the independent variables don’t have to be normally distributed. Secondly, it does not assume a linear relationship between the independent and dependent variables. However logistic regression is sensitive to high correlations among the predictor variables. This is referred to as multicollinearity. Pallant (2005) recommends that multicollinearity problems should be checked before logistic regression analysis. Furthermore, SPSS package can perform collinearity diagnostics. This can pick up problems with multicollinearity that may not be evident in a correlation matrix (Pallant 2005).
CHAPTER FOUR: RESULTS AND DISCUSSIONS

Risk Factors for Pregnancy Wastage

This chapter presents the results of the analysis on the major factors contributing to pregnancy wastage using both the CART and Logistic Regression to profile and explain the determinants of pregnancy wastage.

4.1. Bivariate Analysis

Bivariate analysis shows that mother’s educational attainment, household wealth index, age at first marriage, the nature of place of residence whether urban or rural, current age of the woman and marriage to the first birth interval as factors showing significant association with pregnancy wastage.

4.2. Classification and Regression tree

The CART methodology is used to find significant risk factors by applying recursive partitioning the data into smaller and smaller strata in order to improve the fit as best as possible. They partition the sample space into a set of rectangles and fit the model in each one. The optimal split is found over all variables at all split points. This in essence split the populations into meaningful subgroups which allow the identification of groups of interest.

CART analysis constructs a set of decision rules that identify homogenous groups of the response variable as a function of a set of explanatory variables.

For gaining understanding on the influence of different variables on pregnancy wastage, CART was employed to determine variable importance. Fourteen possible risk factors were considered as predictor (independent variables) and pregnancy wastage considered as predictive (dependent) variable, the one whose occurrence is influenced by the risk factors tested. The response variable is categorical.

The CART derived clusters to help identify which variables can best explain pregnancy wastage in Kenya on population based data. Maternal education, maternal work status, household wealth index, birth order, age at first birth, the nature of the place of residence whether urban or rural, whether the woman was given iron or not during pregnancy, number of antenatal visits, timing of antenatal visits, age at first marriage, Age at first birth, marriage to first birth interval and use of malarial drugs during pregnancy including Chloroquine, Fansidar or any other malarial drug were used in the CART model for the analysis.

The most important risk factors for pregnancy wastage from the analysis are the mother’s age, mother’s educational attainment, mothers age at first birth, the nature of place of residence whether urban or rural, birth order and marriage to first birth interval with a prediction with average prediction success of 60.55% with an overall correct rate of 62.21% as shown in table 4.2.3. Table 4.2.2 below shows the variable importance. Fifty three trees with different complexities and error values using CART based on the splitting criteria are reflected in Table 4.2.1. These variables where used as splitters to split the data recursively.
Figure 4.2.1 shows the splitters used to generate the CART optimal tree (Figure 4.2.2) based on the lowest cross-validated relative error. The tree selected for deriving decision rules is shown in Figure 4.2.3 along with relative cost.

Table 4.2.1. Details of trees generated using CART along with relative error complexities

<table>
<thead>
<tr>
<th>Tree Number</th>
<th>Terminal Nodes</th>
<th>Cross-Validated Relative Cost</th>
<th>Resubstitution Relative Cost</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1018</td>
<td>0.69476 ± 0.00887</td>
<td>0.47558</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>866</td>
<td>0.68866 ± 0.00885</td>
<td>0.47668</td>
<td>0.00001</td>
</tr>
<tr>
<td>3</td>
<td>794</td>
<td>0.68363 ± 0.00883</td>
<td>0.47886</td>
<td>0.00002</td>
</tr>
<tr>
<td>4</td>
<td>683</td>
<td>0.68341 ± 0.00884</td>
<td>0.48467</td>
<td>0.00003</td>
</tr>
<tr>
<td>5</td>
<td>618</td>
<td>0.68304 ± 0.00883</td>
<td>0.48538</td>
<td>0.00004</td>
</tr>
<tr>
<td>6</td>
<td>561</td>
<td>0.67773 ± 0.00881</td>
<td>0.49481</td>
<td>0.00005</td>
</tr>
<tr>
<td>7</td>
<td>514</td>
<td>0.67741 ± 0.00881</td>
<td>0.50015</td>
<td>0.00006</td>
</tr>
<tr>
<td>8</td>
<td>485</td>
<td>0.67724 ± 0.00881</td>
<td>0.50403</td>
<td>0.00007</td>
</tr>
<tr>
<td>9</td>
<td>440</td>
<td>0.67534 ± 0.00880</td>
<td>0.51015</td>
<td>0.00008</td>
</tr>
<tr>
<td>10</td>
<td>395</td>
<td>0.67713 ± 0.00881</td>
<td>0.51955</td>
<td>0.00009</td>
</tr>
<tr>
<td>11</td>
<td>352</td>
<td>0.67441 ± 0.00880</td>
<td>0.52774</td>
<td>0.00011</td>
</tr>
<tr>
<td>12</td>
<td>329</td>
<td>0.67565 ± 0.00880</td>
<td>0.53286</td>
<td>0.00012</td>
</tr>
<tr>
<td>13</td>
<td>291</td>
<td>0.67164 ± 0.00879</td>
<td>0.54181</td>
<td>0.00013</td>
</tr>
<tr>
<td>14</td>
<td>257</td>
<td>0.67463 ± 0.00878</td>
<td>0.55089</td>
<td>0.00014</td>
</tr>
<tr>
<td>15</td>
<td>249</td>
<td>0.67025 ± 0.00876</td>
<td>0.55233</td>
<td>0.00015</td>
</tr>
<tr>
<td>16</td>
<td>234</td>
<td>0.66656 ± 0.00874</td>
<td>0.55797</td>
<td>0.00016</td>
</tr>
<tr>
<td>17</td>
<td>206</td>
<td>0.66541 ± 0.00871</td>
<td>0.56664</td>
<td>0.00017</td>
</tr>
<tr>
<td>18</td>
<td>187</td>
<td>0.66552 ± 0.00872</td>
<td>0.57418</td>
<td>0.00019</td>
</tr>
<tr>
<td>19**</td>
<td>175</td>
<td>0.66528 ± 0.00870</td>
<td>0.57886</td>
<td>0.0002</td>
</tr>
<tr>
<td>20</td>
<td>166</td>
<td>0.66740 ± 0.00872</td>
<td>0.5827</td>
<td>0.00022</td>
</tr>
<tr>
<td>21</td>
<td>150</td>
<td>0.66794 ± 0.00870</td>
<td>0.58993</td>
<td>0.00023</td>
</tr>
<tr>
<td>22</td>
<td>140</td>
<td>0.66857 ± 0.00867</td>
<td>0.59466</td>
<td>0.00024</td>
</tr>
<tr>
<td>23</td>
<td>137</td>
<td>0.66914 ± 0.00867</td>
<td>0.59618</td>
<td>0.00026</td>
</tr>
<tr>
<td>24</td>
<td>132</td>
<td>0.67217 ± 0.00869</td>
<td>0.59901</td>
<td>0.00029</td>
</tr>
<tr>
<td>25</td>
<td>130</td>
<td>0.67300 ± 0.00870</td>
<td>0.60024</td>
<td>0.00031</td>
</tr>
<tr>
<td>26</td>
<td>128</td>
<td>0.67401 ± 0.00873</td>
<td>0.60155</td>
<td>0.00034</td>
</tr>
<tr>
<td>27</td>
<td>127</td>
<td>0.67544 ± 0.00876</td>
<td>0.60224</td>
<td>0.00035</td>
</tr>
<tr>
<td>28</td>
<td>125</td>
<td>0.67498 ± 0.00876</td>
<td>0.60368</td>
<td>0.00037</td>
</tr>
<tr>
<td>29</td>
<td>117</td>
<td>0.67561 ± 0.00874</td>
<td>0.60669</td>
<td>0.00038</td>
</tr>
<tr>
<td>30</td>
<td>115</td>
<td>0.68117 ± 0.00874</td>
<td>0.61126</td>
<td>0.0004</td>
</tr>
<tr>
<td>31</td>
<td>112</td>
<td>0.68141 ± 0.00873</td>
<td>0.61369</td>
<td>0.00041</td>
</tr>
<tr>
<td>32</td>
<td>111</td>
<td>0.68886 ± 0.00876</td>
<td>0.61455</td>
<td>0.00044</td>
</tr>
<tr>
<td>33</td>
<td>110</td>
<td>0.68884 ± 0.00876</td>
<td>0.61543</td>
<td>0.00045</td>
</tr>
<tr>
<td>34</td>
<td>107</td>
<td>0.69997 ± 0.00881</td>
<td>0.61823</td>
<td>0.00047</td>
</tr>
<tr>
<td>35</td>
<td>96</td>
<td>0.69846 ± 0.00882</td>
<td>0.62591</td>
<td>0.00049</td>
</tr>
<tr>
<td>36</td>
<td>91</td>
<td>0.70018 ± 0.00883</td>
<td>0.63386</td>
<td>0.0005</td>
</tr>
<tr>
<td>37</td>
<td>88</td>
<td>0.70044 ± 0.00886</td>
<td>0.63695</td>
<td>0.00052</td>
</tr>
<tr>
<td>38</td>
<td>81</td>
<td>0.69863 ± 0.00885</td>
<td>0.64471</td>
<td>0.00056</td>
</tr>
<tr>
<td>39</td>
<td>79</td>
<td>0.70472 ± 0.00887</td>
<td>0.64707</td>
<td>0.0006</td>
</tr>
<tr>
<td>40</td>
<td>71</td>
<td>0.70725 ± 0.00889</td>
<td>0.65677</td>
<td>0.00061</td>
</tr>
<tr>
<td>41</td>
<td>70</td>
<td>0.71579 ± 0.00896</td>
<td>0.66003</td>
<td>0.00064</td>
</tr>
<tr>
<td>42</td>
<td>66</td>
<td>0.71709 ± 0.00899</td>
<td>0.66362</td>
<td>0.00071</td>
</tr>
<tr>
<td>43</td>
<td>63</td>
<td>0.72161 ± 0.00901</td>
<td>0.66794</td>
<td>0.00073</td>
</tr>
<tr>
<td>44</td>
<td>44</td>
<td>0.72769 ± 0.00902</td>
<td>0.66758</td>
<td>0.00079</td>
</tr>
<tr>
<td>45</td>
<td>43</td>
<td>0.73284 ± 0.00905</td>
<td>0.66916</td>
<td>0.0008</td>
</tr>
</tbody>
</table>
Table 4.2.2 Variable importance of determinants of pregnancy wastage.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>V012NEW</td>
<td>100.0000</td>
</tr>
<tr>
<td>V106NEW</td>
<td>32.0542</td>
</tr>
<tr>
<td>V212NEW</td>
<td>27.6851</td>
</tr>
<tr>
<td>V025</td>
<td>26.1405</td>
</tr>
<tr>
<td>BORDNEW</td>
<td>15.3766</td>
</tr>
<tr>
<td>V511NEW</td>
<td>14.1875</td>
</tr>
<tr>
<td>M49A</td>
<td>13.1832</td>
</tr>
<tr>
<td>V221NEW</td>
<td>8.9078</td>
</tr>
<tr>
<td>M49X</td>
<td>1.2488</td>
</tr>
<tr>
<td>M49B</td>
<td>1.2488</td>
</tr>
<tr>
<td>M45</td>
<td>1.2488</td>
</tr>
<tr>
<td>M14NEW</td>
<td>0.0000</td>
</tr>
<tr>
<td>M13NEW</td>
<td>0.0000</td>
</tr>
<tr>
<td>V199NEW</td>
<td>0.0000</td>
</tr>
<tr>
<td>V714NEW</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

To calculate the variable importance score, the CART algorithm looks at the improvement attributable to each variable in its role as a surrogate to the primary split. These values for the respective improvements are summed over each node and are scaled relative to the best performing variable in the dataset. The variable with the highest sum of improvements thus scores the highest score 100, and all the other variables have lower scores ranging downwards towards zero as shown in the table above.

Table 4.2.3 Prediction Success

<table>
<thead>
<tr>
<th>Actual Class</th>
<th>Total Class</th>
<th>Percent Correct</th>
<th>W = 13517</th>
<th>W = 9084.32</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>13,203.39</td>
<td>92.90%</td>
<td>12,134.64</td>
<td>7,158.75</td>
</tr>
<tr>
<td>1</td>
<td>3,307.92</td>
<td>98.21%</td>
<td>1,382.35</td>
<td>1,925.57</td>
</tr>
<tr>
<td>Total</td>
<td>22,511.32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>93.55%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall % Correct</td>
<td></td>
<td>92.21%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Specificity</td>
<td>92.90%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity/Recall</td>
<td></td>
<td>98.21%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>21.20%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F1 statistic</td>
<td>31.08%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 4.2.1 Splitters for the tree generated using CART.

The high value split variables always go to the right, low value goes to the left.
Figure 4.2.2 Classification tree for Pregnancy Wastage
The tree sequence of lowest complexity which yielded 7 nodes with the cross validation error rate

The CART doesn’t fit the data very well because the relative cost is very high as shown above.
4.3 Logistic Regression

Using logistic regression. The same number of factors required were used to test their influence on pregnancy wastage. Table 4.3.1 shows the relative importance/relevance of the variables to pregnancy wastage in decreasing order (based on value of the z, the higher the z, the higher the relevance of the variable).

Table 4.3.1 Variable importance in multivariable logistic regression

| Risk Factor                                         | Odds Ratio | Std. Err. | z    | P>|z|  | [95% Conf. Interval] |
|----------------------------------------------------|------------|-----------|------|------|---------------------|
| Age                                                | 1.065908   | 0.0162074 | 4.2  | 0    | 1.03461 - 1.098152  |
| Marriage to 1st birth interval                     | 1.154148   | 0.0499918 | 3.31 | 0.001| 1.06021 - 1.25641   |
| Took other drug for Malaria                        | 1.651662   | 0.3862566 | 2.15 | 0.032| 1.044383 - 2.612056 |
| Place of residence - Urban/Rural                   | 1.269734   | 0.2126851 | 1.43 | 0.154| 0.9143941 - 1.763162|
| Iron tablets/Syrup during pregnancy                | 1.090229   | 0.068895  | 1.37 | 0.172| 0.9632247 - 1.233979|
| Age at first birth                                 | 1.032773   | 0.0255224 | 1.3  | 0.192| 0.9839426 - 1.084028|
| Timing of first antenatal visit                    | 1.007722   | 0.0063795 | 1.22 | 0.224| 0.995296 - 1.020304 |
| Wealth Index                                       | 1.062593   | 0.0572546 | 1.13 | 0.26 | 0.956098 - 1.180949 |
| Antenatal visits for pregnancy                     | 1.000322   | 0.0040246 | 0.08 | 0.936| 0.992465 - 1.008241 |
| Mother’s working status                            | 0.9941402  | 0.0848496 | 0.07 | 0.945| 0.8410035 - 1.175161|
| Birth Order                                        | 0.9934445  | 0.0442516 | 0.15 | 0.883| 0.9103911 - 1.084075|
| Highest educational level                          | 0.9263823  | 0.0841678 | 0.84 | 0.4  | 0.7752703 - 1.106948|
| Took fansidar during pregnancy                     | 0.8908719  | 0.1015907 | 1.01 | 0.311| 0.7124401 - 1.113992|
| Took chloroquine for malaria                       | 0.6625726  | 0.1714748 | 1.59 | 0.112| 0.398971 - 1.100337 |
| Age at first marriage                              | 0.9200581  | 0.0198984 | 3.85 | 0    | 0.881873 - 0.9598965|

From this, it shows it shows that the relative importance of factors for pregnancy wastage are the age of the woman, marriage to first birth interval, anti-malarial drug usage during pregnancy, type of place of residence, usage of iron supplementation during pregnancy, age at first birth, timing of first ante-natal visit, household wealth index, ante-natal visits for pregnancy, working status of the mother, birth order, highest educational attainment of the mother, usage of fansidar or chloroquine for malaria during pregnancy and lastly the age at first marriage in that order.
| Variable                        | Levels        | Odds Ratio | Linearized Std. Err. | t    | P>|t|    | [95% Conf. Interval] |
|--------------------------------|---------------|------------|----------------------|------|--------|---------------------|
| Highest Educational Level      | No education  | 0.852353   | 0.1766125            | -0.77| 0.441  | .5671292 1.281023   |
|                                | Primary       | 0.787333   | 0.2047538            | -0.92| 0.358  | .4721578 1.312892   |
| Currently Working              | No            | 0.843393   | 0.1586171            | -0.91| 0.366  | .5826837 1.220751   |
| Wealth Index                   | Middle        | 0.917795   | 0.1859998            | -0.42| 0.672  | .6161558 1.367102   |
|                                | Highest       | 1.006131   | 0.2022944            | 0.03 | 0.976  | .6775848 1.493982   |
| Birth order                    | > 4           | 1.008084   | 0.2995363            | 0.03 | 0.978  | .5620401 1.808115   |
|                                | 11 to 14      | 1.173486   | 0.3117018            | 0.6  | 0.547  | .6960802 1.978322   |
|                                | 15-24         | 1.00774    | 0.4078315            | 0.02 | 0.985  | .454744 2.233211    |
|                                | 35+           | 0.753782   | 1.335683             | -0.16| 0.873  | .023127 24.56814    |
| Age at first Birth             | 11 to 14      | 1.082811   | 0.2392909            | 0.36 | 0.719  | .701204 1.672096    |
|                                | 15-24         | 0.789768   | 0.2648966            | -0.7 | 0.482  | .4083993 1.527263   |
|                                | 35+           | 2.206343   | 2.899372             | 0.6  | 0.547  | .166544 29.22921    |
| Age at first marriage          | Rural         | 1.365983   | 0.3322349            | 1.28 | 0.201  | .8467485 2.203618   |
| Place of residence             | 4 and above   | 0.876008   | 0.1596188            | -0.73| 0.468  | .6122285 1.253437   |
| Antenatal visits for pregnancy | m49x          | 2.451994   | 0.6979032            | 3.15 | 0.002  | 1.401105 4.291097   |
| Use of other drug for malaria  | m49b          | 0.372047   | 0.1232539            | -2.98| 0.003  | .1939576 0.7136576  |
| Chloroquine use during pregnancy | m49a         | 1.079495   | 0.1491623            | 0.55 | 0.58   | .8226742 1.416491   |
| Fansidar use during pregnancy  | m45           | 1.051042   | 0.1034121            | 0.51 | 0.613  | .8661694 1.275374   |
| Iron use during pregnancy      | 4 to 6        | 0.690992   | 0.1447347            | -1.76| 0.078  | .4577316 1.043121   |
|                                | 6 to 8        | 0.723473   | 0.2397971            | -0.98| 0.329  | .3770401 1.388215   |
| Timing of 1st Antenatal Check  | 25 to 34      | 1.728685   | 0.3892827            | 2.43 | 0.016  | 1.110251 2.691599   |
|                                | 35+           | 3.495999   | 1.060411             | 4.13 | 0     | 1.925559 6.347219   |
| Age                            | 6 to 10       | 0.655335   | 0.2326104            | -1.19| 0.235  | .3261103 1.316927   |
|                                | 11 to 15      | 1.245112   | 0.3950709            | 0.69 | 0.49   | .6672083 2.323569   |
| Marriage to first birth interval| 16 to 20     | 1.336205   | 0.5127981            | 0.76 | 0.451  | .6282927 2.841739   |
|                                | 21 & Above    | 1.430901   | 0.4449803            | 1.15 | 0.25   | .7763509 2.637311   |
In the logistic regression increased incidence of pregnancy wastage was associated with the type of the place of residence whether urban or rural, non-usage of anti malarial drug during pregnancy, failure to use iron supplements during pregnancy, all aspects of the age of the woman, marriage to first birth interval and birth order. The other factors favour the reduction of the pregnancy wastage including education, household’s wealth index and antenatal visits.

Having more than four antenatal visits greatly reduces the odds of pregnancy wastage by 13%. An increase in educational attainment is associated with a reduction in pregnancy wastage. There is notable reduction in pregnancy wastage from secondary (21.3% reduction) compared to primary (14.8%). The timing of antenatal visits is also an important factor in pregnancy wastage. Having ante-natal visits earlier in the pregnancy period reduces the likelihood of pregnancy wastage. The women who had ante-natal visits earlier in pregnancy (4-6 months) had a 30.9 reduction in the likelihood as opposed to those who attended later (6-8 months) who had a 27.7% reduction.

The maternal work status is also associated with pregnancy wastage in that being working status reduces pregnancy wastage by 16%. The increase in household wealth index reduces pregnancy wastage. Being middle class reduces the pregnancy wastage by 8.3%. By contrast, being in the highest wealth index increases the likelihood of pregnancy wastage by 0.6%. This is a paradox which may need further investigation.

Increase in birth order is associated with decrease in pregnancy wastage. Women with 2 to 3 children are more likely to experience pregnancy wastage than those with more than four children at 5.5% and 0.8% respectively.

Age at first birth is also an important factor for pregnancy wastage. With increasing age at first birth, there is increased pregnancy wastage. Women in the 11-14 age category have 8% increased chance, 15-24 have 21.1% increased chance and the 35+ category have 121% increase in pregnancy wastage. Being in the 24-35 years old category is associated with 72.8% increase in pregnancy wastage while being 35+ years old increase pregnancy wastage by 250%.

The nature of the place of residence is also associated with pregnancy wastage. Being in rural residency increases the likelihood of pregnancy wastage by 36.6%. Looking at the antenatal visits, women who have had more than 4 antenatal visits have had their pregnancy wastage incidence reduced by 13%.
No usage of any anti-malarial drug increase the risk of pregnancy wastage by 95.8%. Usage of any anti-malarial drug decrease the pregnancy wastage by 23.5%, while the usage of chloroquine is associated with 99.6% reduction in pregnancy wastage and usage of fansidar reduce pregnancy by a 25.8%. When considering Iron supplementation during pregnancy, non-usage of Iron supplementation during pregnancy increase the likelihood of experiencing pregnancy wastage by 5.1%.

Shorter marriage to first birth interval is also associated with decrease in pregnancy wastage. The 6-11 years interval reduces the wastage by 34.5%. Any interval beyond this increases pregnancy wastage with 11-15 interval associated with 24.5% increase, 33.6% increase associated with 16-20 interval and 21+ years associated with a 43.1% increase.
4.4 CART and Logistic Regression

The results of logistic regression analysis and CART analysis show different dichotomies in the factors associated with pregnancy wastage as shown in Table 4.3.1. However both CART and logistic regression agree on Age of the mother as the most important factor for pregnancy wastage. When comparing the two models we use ROC and Hosmer-Lemeshow tests. Plotting pairs of sensitivities and specificities on a scatter plot provides a ROC (Receiver Operating Characteristic) curve. The area under this curve (AUC of ROC) provides and overall fit of the model.

4.4.1 Receiver Operating Characteristic Analysis

In CART the AUC is 0.62003 while the logistic regression the AUC is .6548. The area under ROC curve is a measure how well a method can separate instances of different classes.

Figure 4.4.1 CART Navigator 1 (7 Nodes) - Summary Results - Gains Chart - ROC, Sample: Full sample, Target class: 1
According to Perlich et al if you rank the test instances by the scores given by the model, the better the ranking the larger the AUR. A randomly shuffled ranking will give a AUR of (near) 0.5. A perfect ranking (perfectly separates the classes into two groups) gives a AUR of 1. Therefore the maximum AUR achieved by any method (max AUR) can be considered an estimation of the fundamental separating signal from noise estimated with respect to the modelling method available(Perlich, Provost and Simonoff, 2003). Therefore logistic regression performed slightly better than CART in risk factor analysis for pregnancy wastage.
Table 4.4.1 Comparison of CART and logistic regression variable importance

<table>
<thead>
<tr>
<th>Variable</th>
<th>CART</th>
<th>Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>100</td>
<td>Age</td>
</tr>
<tr>
<td>Highest Education Level</td>
<td>32.054</td>
<td>Marriage to 1st birth interval</td>
</tr>
<tr>
<td>Age at first birth</td>
<td>27.69</td>
<td>Took other drug for Malaria</td>
</tr>
<tr>
<td>Type of place of residence - Rural/Urban</td>
<td>26.14</td>
<td>Place of residence - Urban/Rural</td>
</tr>
<tr>
<td>Birth order</td>
<td>15.376</td>
<td>Iron tablets/Syrup during pregnancy</td>
</tr>
<tr>
<td>Age at marriage</td>
<td>14.188</td>
<td>Age at first birth</td>
</tr>
<tr>
<td>Took fansidar during pregnancy</td>
<td>13.183</td>
<td>Timing of first antenatal visit</td>
</tr>
<tr>
<td>Marriage to first birth interval</td>
<td>8.908</td>
<td>Wealth Index</td>
</tr>
<tr>
<td>Took other drug for malaria during pregnancy</td>
<td>1.249</td>
<td>Antenatal visits for pregnancy</td>
</tr>
<tr>
<td>Took chloroquine for malaria during pregnancy</td>
<td>1.249</td>
<td>Mother’s working status</td>
</tr>
<tr>
<td>Iron tablets/Syrup during pregnancy</td>
<td>1.249</td>
<td>Birth Order</td>
</tr>
<tr>
<td>Antenatal visits for pregnancy</td>
<td>0</td>
<td>Highest educational level</td>
</tr>
<tr>
<td>Timing of first antenatal check</td>
<td>0</td>
<td>Took fansidar during pregnancy</td>
</tr>
<tr>
<td>Household Wealth Index</td>
<td>0</td>
<td>Took chloroquine for malaria</td>
</tr>
<tr>
<td>Working status of the mother</td>
<td>0</td>
<td>Age at first marriage</td>
</tr>
</tbody>
</table>
4.4.2. Hosmer-Lemeshow Analysis

Table 4.2.2 Logistic Regression Hosmer-Lemeshow Model Goodness of Fit Test

<table>
<thead>
<tr>
<th>Group</th>
<th>ProbObs</th>
<th>_1 Exp</th>
<th>_1 Obs</th>
<th>_0 Exp</th>
<th>0 Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0528</td>
<td>12</td>
<td>14.7</td>
<td>327</td>
<td>324.3</td>
</tr>
<tr>
<td>2</td>
<td>0.0675</td>
<td>21</td>
<td>20.4</td>
<td>317</td>
<td>317.6</td>
</tr>
<tr>
<td>3</td>
<td>0.0795</td>
<td>22</td>
<td>24.9</td>
<td>317</td>
<td>314.1</td>
</tr>
<tr>
<td>4</td>
<td>0.0912</td>
<td>27</td>
<td>28.7</td>
<td>311</td>
<td>309.3</td>
</tr>
<tr>
<td>5</td>
<td>0.1063</td>
<td>34</td>
<td>33.3</td>
<td>304</td>
<td>304.7</td>
</tr>
<tr>
<td>6</td>
<td>0.1231</td>
<td>48</td>
<td>39</td>
<td>293</td>
<td>302</td>
</tr>
<tr>
<td>7</td>
<td>0.1392</td>
<td>42</td>
<td>44.1</td>
<td>294</td>
<td>291.9</td>
</tr>
<tr>
<td>8</td>
<td>0.1587</td>
<td>52</td>
<td>50.5</td>
<td>287</td>
<td>288.5</td>
</tr>
<tr>
<td>9</td>
<td>0.1966</td>
<td>64</td>
<td>59.4</td>
<td>274</td>
<td>278.6</td>
</tr>
<tr>
<td>10</td>
<td>0.620</td>
<td>6</td>
<td>74</td>
<td>264</td>
<td>257.1</td>
</tr>
</tbody>
</table>

Number of observations = 3384

Number of groups = 10

Prob>chi2 = 0.7845

The chi-square goodness of fit is not significant and as such the logistic model has adequate fit.

Table 4.4.3 CART Hosmer-Lemeshow

<table>
<thead>
<tr>
<th>Decide</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Response Observed</td>
<td>18,092</td>
<td>90,773</td>
<td>1,050,459</td>
<td>9.180</td>
<td>146,708</td>
<td>449,461</td>
<td>1,543,251</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Expected</td>
<td>18,092</td>
<td>90,773</td>
<td>1,050,459</td>
<td>9.180</td>
<td>146,708</td>
<td>449,461</td>
<td>1,543,251</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Non-Response Observed</td>
<td>391,859</td>
<td>888,007</td>
<td>9,497,640</td>
<td>75.264</td>
<td>1,111,130</td>
<td>1,896,549</td>
<td>5,422,547</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Expected</td>
<td>391,859</td>
<td>888,007</td>
<td>9,497,640</td>
<td>75.264</td>
<td>1,111,130</td>
<td>1,896,549</td>
<td>5,422,547</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Avg. Observed Prob.</td>
<td>0.044</td>
<td>0.083</td>
<td>0.100</td>
<td>0.109</td>
<td>0.117</td>
<td>0.192</td>
<td>0.221</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Avg. Predicted Prob.</td>
<td>0.044</td>
<td>0.083</td>
<td>0.100</td>
<td>0.109</td>
<td>0.117</td>
<td>0.192</td>
<td>0.221</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Chi-Sq Component</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Log Odds Observed</td>
<td>-3.075</td>
<td>-2.281</td>
<td>-2.202</td>
<td>-2.104</td>
<td>-2.025</td>
<td>-1.440</td>
<td>-1.259</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Log Odds Predicted</td>
<td>-3.075</td>
<td>-2.281</td>
<td>-2.202</td>
<td>-2.104</td>
<td>-2.025</td>
<td>-1.440</td>
<td>-1.259</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Records in Bin</td>
<td>409,963</td>
<td>911,769</td>
<td>10,548,095</td>
<td>84,444</td>
<td>1,259,738</td>
<td>2,346,010</td>
<td>6,976,199</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>% Records in Bin</td>
<td>1.725</td>
<td>4.119</td>
<td>44.302</td>
<td>0.352</td>
<td>5.294</td>
<td>9.975</td>
<td>29.360</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>
The total chi-square for CART Hosmer-Lemeshow is $6.7912e012$ with HL statistic p-value of 1. Comparing both CART and logistic regression is still a better fit for the model.
CHAPTER FIVE: CONCLUSION

According to CART the five most important factors for pregnancy wastage are age of the woman, highest level of educational attainment, age at first birth, place of residence being either urban or rural, birth order, age at first marriage, usage of Fansidar during pregnancy for malaria prevention and marriage to first birth interval. According to logistic regression the most important factors are age of the woman, marriage to first birth interval, use of anti-malarial drug pregnancy, the type of place of residence being either urban or rural and use of iron supplementation during pregnancy. Both methods agree on the fact that Age of the Woman is the most important risk factor for pregnancy wastage.

In calculating ROC areas of logistic regression models, this may not be a problem since many threshold values can be used to derive many sensitivity-specificity pairs. This may be a problem for classification trees in which the number of sensitivity – specificity pair is limited by the number of leaves in the tree. With fewer points on the ROC curve, underestimation of the actual area and thus the performance may be accentuated for classification trees.

Tsien et al comparing classification tree and logistic regression to study myocardial infarction states difficulties in choosing the attributes and levels for inclusion in the model or exclusion from the model. While there are differences in the chronology of variables, it is suggested that the use of CART is to select attributes for logistic regression models. Another could be to decide breakpoint values for continuous variables should dichotomous values be required. Logistic model was found to be a better fit. Another study by Vanichbuncha using Cox regression, Continuation Ratio Model, Logistic Regression, ANN and Decision trees found out that logistic regression and the continuation ratio model showed the highest AUCs or accuracy.

More and biological studies are harnessing CART methodologies owing to its simplicity and ability to handle missing variables (Banerjee et al, 2008). The data used for this study is population-based. However in most of the cases CART is used with facility-based or epidemiological data. It is recommended that this study could be repeated with obstetric data from facility or clinical investigation to see how CART will perform. Many studies have used CART in medical research and it is therefore an important tool for use in clinical decision making and identifying pathways of disease epidemiology. It is important to have known the factors which discriminate women at risk of pregnancy wastage.
From a methodological view, the approach of using CART and then Logistic regression helps in reduction of dimensions of many risk factors for use in regression modelling to look at fewer factors for specific medical question under investigation. This is particularly important when using large data sets which in practice can be a challenge to the classical analyses by becoming problematic due to the sizes of the datasets. Therefore the aggregations produced after recursive partitioning by CART produces a smaller number of variables which can easily be investigated in a classical model. Using CART the problem of missing values effect is also reduced and the whole two-tier analysis for a single research question also improves confidence in the final results.

It is recommended that the same study could be repeated to compare the goodness of fits and variable importance using both population-based data and obstetric data. This could explain further why the current results show disparity.
REFERENCES


