2017-01-01

Evaluating Binary Splits On Nominal Inputs

Isaac Xoese Ocloo
University of Texas at El Paso, ocloox1@gmail.com

EVALUATING BINARY SPLITS ON NOMINAL INPUTS

ISAAC XOESE OCLOO

Master's Program in Mathematical Sciences

APPROVED:

Xiaogang Su, Ph.D., Chair

Amy E. Wagler, Ph.D.

Wen-Yee Lee, Ph.D.

Charles H. Ambler, Ph.D.
Dean of the Graduate School

to my

FAMILY

with love

EVALUATING BINARY SPLITS ON NOMINAL INPUTS

by

ISAAC XOESE OCLOO

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Master's Program in Mathematical Sciences

THE UNIVERSITY OF TEXAS AT EL PASO

August 2017

Acknowledgements

Bless the Lord, O my soul and all that is within me, bless his holy name. I am thankful to you Lord for bringing me to Ebenezer. I am also very grateful to my supervisor and mentor Dr. Xiaogang Su, whose guidance and supervision made this work a success. You didn't give up on me when I even gave up on myself. You are like a father to me, always encouraging and guiding me through my studies.

I wish to thank the members of my committee, Dr. Amy E. Wagler, Associate Chair of Department of Mathematical Sciences, and Dr. Wen-Yee Lee, Department of Chemistry, all at The University of Texas at El Paso. Your guidance and assistance were greatly valuable to the completion of this work.

I also wish to thank Dr. Joan Staniswalis, Dr. Panagis Moschopoulos, Dr. Naijun Sha, and Dr. Ori Rosen, all of the Mathematical Sciences Department at The University of Texas at El Paso who taught and mentored me through my studies. Additionally, I wish to thank all professors and staff of the Mathematical Sciences Department here in The University of Texas at El Paso for all their support, enabling me to complete my degree.

I appreciate my wonderful family in Ghana. Your love and prayers has made this journey a success. To all well-wishers whose names are not mentioned, I say "God richly bless you".

Abstract

The maximally selected statistic approach in building tree models is shown to be a cause of variable selection bias. In this study we propose three methods to solve this problem in building regression trees with nominal predictor variables. Out of the three methods proposed we explored only two in detail and defer one for further research. We developed an exact method to compute the p-value corresponding to the maximized splitting statistic in regression trees for nominal predictor variables with at most 10 distinct levels and a method to estimate the best cutoff point as a parameter in a parametric nonlinear mixed-effect model in regression trees for nominal predictor variables with any number of distinct levels. The methods are shown to overcome the variable selection bias in an extensive simulation study and in a real data example.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Tree based methods or prediction trees follow a recursive partitioning algorithm known as exhaustive or greedy search to split data into disjoint subsets. This method is classified into two depending on the type of dependent variable, classification trees for categorical dependent variables and regression trees for continuous response variables.

The application of linear regression and logistic regression in fitting a predictive model for data consisting of a continuous response variable or a categorical response variable respectively with one or more independent predictor variables has been very popular and heavily used in various fields. This idea of fitting a single predictive formula holding over the entire data-space faces a challenge when the data have lots of features which interact in complicated, nonlinear ways, and thus developing a single global model can be very difficult.

Tree based methods become handy in these situations, but not only exclusive to such. With a wide area of application, tree based methods inspire intensive research that is aimed to improve many aspects of its current methodology. Consider an application of classification trees in a diabetics study with two predictor variables, age and blood sugar level in a medical research. It looks more appealing to tell a patient, based on your age beyond a certain cutoff point, and your blood pressure below a certain cutoff point, you are classified as having the disease or not rather than just merely telling the patient you have being diagnosed with the disease or not.

One application of a regression tree is its use to explain differences in the salaries of major

league baseball players and to answer the question "Are players paid according to their performance?". Su et al. (2016) developed a regression tree using this dataset with response variable being salary of 1987 major league baseball players and 23 predictor variables mostly being performance measures recorded on the players.

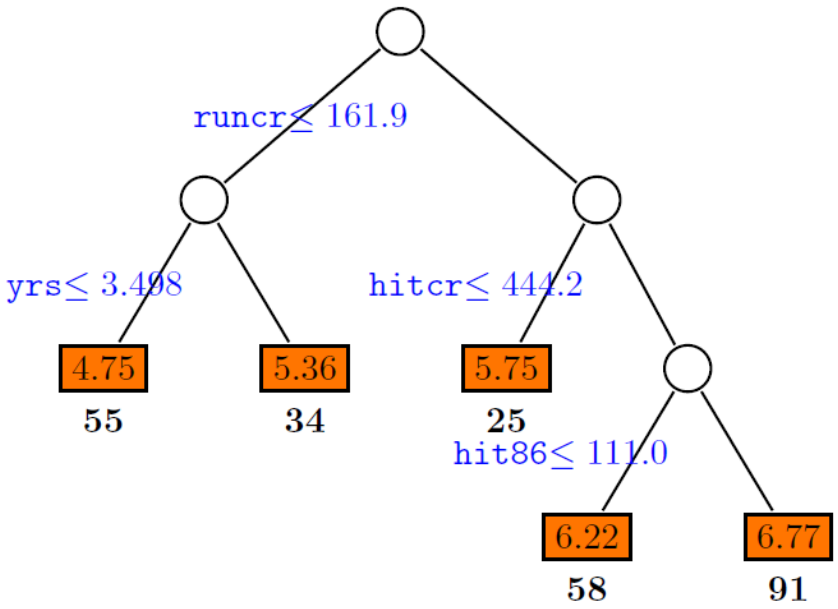


Figure 1.1: Analysis of 1987 Baseball Salary Data.Within each terminal node is the mean response(log-transformed salary); underneath is the sample size. Figure copied from Su et al. (2016)

The tree developed is shown in Figure 1.1 with its interpretation as follows. Starting at the root node and considering the left child node, it is estimated that if the number of runs of the player in 1986 was less than or equal to 161.9 and he played for at most 3.498 years his estimated salary for 1987 would be 4.75 (log-transformed salary) which is approximately 116 thousand dollars. Also a player whose number of runs for 1986 was more than 161.9 and played for more than three and half years had an estimated salary of 5.36 (log-transformed salary) which is approximately 213 thousand dollars in 1987.

These interesting applications among others have made this area of study very interesting. Tree based methods have been shown to have the following advantages as stated in the CART book by Breiman et al. (1984).

- It can be applied to many data structures, especially those involving nonlinear patterns or many predictors that possibly interact in a complicated manner. Tree methods handle both ordered and categorical variables in a simple and natural way through appropriate formulation.
- It does stepwise variable selection, interactions and complexity reduction in an automatic manner. And it makes powerful use of conditional information in handling nonhomogeneous relationships.
- It is invariant under all monotone transformations of individual ordered variables.
- It is extremely robust with respect to outliers and wrong recordings.
- The tree procedure output gives easily understandable and interpreted information regarding the predictive structure of the data.

In spite of the simplicity enjoyed in fitting tree based methods together with its numerous advantages it is confronted with some challenges such as variable selection bias (VSB) and end-cut preference (ECP). End-cut preference usually results from the presence of a few outliers in the data which may distort both the average assigned to a terminal node as well as have a strong influence in the choice of the best split, resulting in splits that have a very small sub-set of cases in one of the branches Torgo (2001). Variable selection bias is induced by the exhaustive or greedy search approach used to determine the best split. This has a selection bias toward variables which provide more split points, Doyle (1992). VSB is a problem that can undermine the reliability of inferences from a tree structure,Loh (2002). The VSB can be corrected by making a decision to split the data based on the probability value associated with the maximized split for each predictor variable Shih and

Tsai (2004). Loh (2002) also proposed a remedy to VSB in regression trees by employing chi-square analysis of residuals and bootstrap calibration of significance probabilities implemented through the GUIDE package. VSB can also be corrected by approximating the indicator threshold function in greedy search with a smooth sigmoid surrogate (SSS) function,Su et al. (2016).

In building classification and regression trees with predictor variables of various types except for a nominal predictor with a specified number of distinct levels, various approaches have been developed to correct VSB either through the p-value or some other alternative approach. However finding the p-value associated with the maximized split induced by a nominal predictor variable or an alternative approach in correcting VSB in classification and regression trees have received less attention, thus setting the tone for this thesis.

1.0.1 Illustrative Example 1 of Variable Selection Bias

The variable selection bias attributable to the maximally selected statistics approach use in building regression trees is illustrated by generating a data of size 100 from the model $Y = 0.5I.(X_1 \in A) + \epsilon$ where $\epsilon \overset{IID}{\sim} \mathbf{N(0,1)}$. X_1 is a nominal variable with two levels and is related to Y. X_2 is also a nominal variable with 10 levels but not related to Y. The maximally selected statistics approach selects a split on X_2 since it has the highest maximized splitting statistics, table 1.1, even though X_2 is not related to Y. This wrong decision to split on X_2 is because it has more distinct levels than X_1 thus inducing more binary splits leading a selection bias.

Table 1.1: Splitting Data by Comparing the Maximized Splitting Statistics

Levels	Maximally Selected Statistic
K=2	0.3242
K=10	0.3761

1.0.2 Illustrative Example 2 of Variable Selection Bias

In the regression tree analysis of the 1987 baseball salary data one final tree is shown in Figure 1.1 using SSS and another is presented in figure 1.2 using the CART procedure. Commenting on the two splits based on team86 and team87 (highlighted with dashed lines), Loh (2002) stated that these splits are hard to interpret and may be attributable to the selection bias of greedy search.

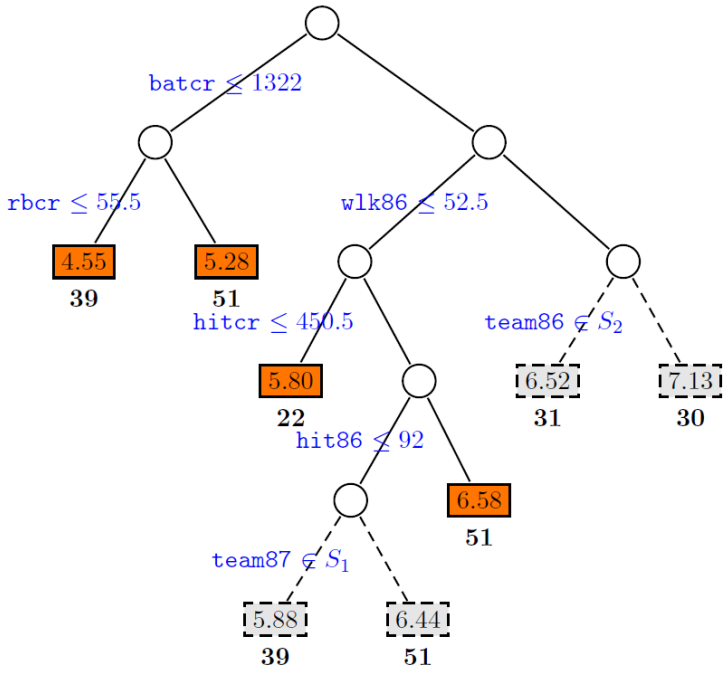


Figure 1.2: Analysis of 1987 Baseball Salary Data.With splits on team86 and team87 attributable to variable selection bias. Figure copied from Su et al. (2016)

1.1 Problem Statement

Variable selection bias(VSB) has led to variables which hitherto should not be considered as the most important variables to split the data and thus weaken confidence in explaining the results Shih and Tsai (2004) and difficulty in interpreting the results Loh (2002). Even though much effort has been devoted into solving the problem of variable selection bias by various pairs of the response variable and each possible type of the predictor variable, e.g., categorical response variables and various predictor types, Miller and Siegmund (1982), Boulesteix (2006b), Boulesteix (2006a) etc in classification trees as well as continuous response variables with various predictor variable types Loh (2002),Shih and Tsai (2004),Su et al. (2016) in regression trees; not much attention has been given to variable selection bias in regression trees where the predictor variable is nominal.

1.2 Objective of the Research

The objective of this thesis is to develop a methodology to evaluate binary splits on nominal predictors, which are capable of addressing the problem of variable selection bias in a regression tree where the predictor variable is nominal. More specifically we seek appropriate ways for computing the p-values associated with the maximized split induced on a continuous response variable in a regression tree by a nominal predictor variable. Depending on the number of distinct levels of the nominal predictor variable and the splitting criteria, three methods are proposed:

- An exact distribution for nominal predictor variables with at most 10 distinct levels.
- Two approaches to be used for predictor variables with any number of distinct levels.

1.3 Significance of the study

The problem caused by variable selection basis cannot be over emphasized and thus a research targeted at solving this problem has much to contribute. The success of this study would be to improve the building of regression trees for any data where at least one of the predictor variable is nominal. Nominal predictors are ubiquitous, e.g., sales data containing a state variable in which the product is sold, in educational research a common nominal predictor is the students college or field of study and in customer satisfaction studies where a nominal predictor is which bank or credit card company the customer uses.

1.4 Outline of the thesis

The remaining parts of the thesis are organized in this manner. Chapter 2 provides a literature review on constructing a regression tree. We summarize various approaches used in computing the p-value associated with the optimal split in both regression and classification trees. In Chapter 3, we present three proposed methods in computing the p-value associated with the best split induced by a nominal predictor variable in a regression tree. We will then use simulations studies and a real data example to show the variable selection bias and how it is rectified by our proposed methods in Chapter 4. Finally, we will discuss the results obtained from the simulation study and real data example, followed by our conclusion, and provide areas for future work in Chapter 5.

Chapter 2

Literature Review

2.1 Introduction

The aim of this chapter is to present a literature review on constructing a regression tree. This is followed by a literature review on how the selection differential could be used as an alternative binary splitting procedure. A novel approach of replacing the step function in greedy search with a smooth sigmoid surrogate function is reviewed and its numerous advantages are explored. We then review literature on how variable selection bias has been addressed in both regression and classification trees.

2.2 Definitions

Node: Is a subset of the set of variables in the data set, and it can be a terminal or non-terminal node.

Root Node: It consists of the entire data set used to build the tree and is located at the top of the tree.

A non-terminal (or parent) node is a node that splits into two daughter nodes (a binary split).

A terminal node is a node that does not split into two daughter nodes and is assigned a class label in a classification tree or the average of the response value in a regression tree.

Recursive partitioning is the step-by-step process by which a decision tree is constructed by either splitting or not splitting each node on the tree into two daughter nodes.

A regression tree is a tree based model for predicting a continuous response from a predictor or set of predictor variables which could be of mixed types assuming that the response variable and predictor variables are related.

A classification tree is a tree based model for predicting a categorical response from a predictor or set of predictor variables which could be of mixed types assuming that the response variable and predictor variables are related.

2.3 Constructing Regression Trees

This section reviews literature on how a regression tree is constructed. We explore how the splitting decision is made on each predictor variable type and how this leads to splitting the entire dataset into a left child node and a right child node. Some challenges inherent with the procedure are also explained.

One single split of data is a two step process.

1. Finding the best cutoff point for each predictor variable.
2. Compare the best cutoff points in step 1 and select the variable with the best of bests cutoff point to split the entire data. This step is were VSB usually occurs.
3. Repeat step 1 and step 2 at each daughter node until a terminal node is reached and assign the mean of the observations to that node.

2.3.1 Finding the Best Cutoff Point for each Predictor Variable.

Suppose that Y is a continuous response variable and $X_1, X_2, \ldots, X_p$ is a set of predictor variables which are of missed type. The number of possible binary splits that can be induced on the response variable depends on the type of predictor variable. If the predictor variable is nominal with K distinct levels it induces $2^{\mathbf{K}-1} - 1$ possible binary splits on the response variable. On the other hand if the predictor variable is ordinal or continuous with N cases then it induces N-1 possible binary splits on Y. The procedure for finding the best cutoff point for each predictor variable is as follows:

- Define the impurity associated with node t by $R(t) = \sum_{i \in t}(Y_i - \bar{Y}(t))^2$ where $\bar{Y}(t)$ is the complete sample Y-mean at node t.

- If the predictor variable is nominal with K distinct levels then a binary split is induced on the response variable by asking the question is "$x_i \in A?''$, where A is a subset of the K categories. This partitions the sample into two parts; the left child node t_L and the right child node t_R where t_L contains all the cases with $x_i \in A$ and t_R contains the other cases.

- If the predictor variable is continuous or at least ordinal then a binary split is induced on the response variable by asking the question is "$\mathbf{x}_i \leq c?''$, where c is called a cutoff point. This partitions the sample into a left child node t_L and a right child node t_R where t_L contains all the cases with $x_i \leq c$ and t_R contains the other cases.

- The impurity associated with the left and right child nodes, $R(t_L) = \sum_{i \in t_L}(Y_i - \bar{Y}(t_L))^2$ and $R(t_R) = \sum_{i \in t_R}(Y_i - \bar{Y}(t_R))^2$ are computed respectively.

- If the predictor variable is nominal, the decrease in the impurity at node t is define as $\Delta(A, t) = R(t) - R(t_L) - R(t_R)$. The best split is chosen to be the one which induces the maximum decrease in the impurity over all $a_k = 2^{\mathbf{K}-1} - 1$ possible binary splits. That is the best split $x_i \in A^*$ satisfies $\Delta(A^*, t) = \max_{a_k} \Delta(A, t)$.

- If the predictor variable is ordinal or continuous, the decrease in the impurity at node t is defined as $\Delta(c,t) = R(t) - R(t_L) - R(t_R)$. The best split is chosen to be the one which induces the maximum decrease in the impurity over all N-1 possible binary splits. That is the best split $x_i \leq c^*$ satisfies $\Delta(c^*,t) = \max_c \Delta(c,t)$.

In the second step a decision is made to split the entire data into two. The maximized reduction in node impurity corresponding to the best cutoff points for each predictor variable are compared and the variable with the maximum reduction in the impurity among all the variables is selected to split the entire data into a left child node and a right child node. The splitting of the entire data is done at the cutoff point of the selected variable.

2.3.2 Challenges With the Tree Building Procedure

Some challenges which are inherent with the tree building procedure includes:

1. The number of binary splits induced on the response variable by a nominal predictor increases exponentially as the number of distinct levels increase. The procedure in CART Breiman et al. (1984) to address this problem is to find the average of the observations in each category and order them $\bar{Y}_1 \leq \bar{Y}_2 \leq, \ldots, \leq \bar{Y}_K$. Based on this ordering, the category in the nominal predictor variable with the minimum average is assigned a value of 1, followed by 2 for the next category in this increasing order till the last category where K is assigned to the category with the maximum average. Ordering the data transforms the nominal predictor variable into an ordinal predictor variable with K cases, and thus inducing $K-1$ possible binary splits on the response variable, which is easier to handle. However this ordering is shown to introduce correlations where none may exist.

2. Variables with more levels or cases are more likely to be selected to split the data by the maximally selected statistic approach. Loh (2002) explains it as, Suppose X_1

and X_2 are two ordered predictors with n_1 and n_2 distinct values, respectively, with $n_1 \gg n_2$. All other things being equal, X_1 will have a higher chance to be selected than X_2. This so because as more splits are considered on X_1 it is more likely to yield a split with a high maximized reduction in the impurity and thus be selected as the variable to split the data on leading to a variable selection bias.

2.4 Selection Differential

Definition Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a continuous distribution with mean μ and variance σ^2 and distribution function (df) F. Let $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$ denote the order statistics of this sample. Suppose we select the top k X-values. Then k^{-1} $\sum_{i=n-k+1}^{n} (X_{i:n} - \mu)$ represents the average difference between the selected group and the population mean. This quantity expressed in standard deviation units is called the selection differential and may be written as $D_{k,n} = \frac{1}{k} \sum_{i=n-k+1}^{n} (X_{i:n} - \mu)/\sigma$

Selection differential referred to as intensity selection by geneticists and breeders represents a measure of improvement in the given trait due to selection. It is used in constructing suitable breeding plans and also for comparing different breeding plans in both plants and animals. Its application can be extended to other kinds of selection problems. Nagaraja (1981)

The selection of the best cutoff point essentially requires maximizing a two sample statistic; the between node impurity of the two child nodes. The selection differential is also a statistic which evaluates the data after its been split into two. It can equivalently be used as a statistic for the binary split. This method for splitting the data is proposed even though the distribution of the selection differential is not obtained in closed form except for the case where the random variable is exponential Nagaraja (1981). It is our purpose to use

this method of splitting the data since this falls perfectly in line with the CART approach of evaluating binary splits on nominal predictors. This procedure and the method of selection differential share the same idea of order statistics and would be a very fine approach to adapt in determining the best split.

2.5 Replacing the Indicator Function in GS with SSS

The splitting statistic employed in splitting the data can be cast into the following linear model: $y_i = \beta_0 + \beta_1 \delta(x_i; c) + \epsilon_i$,

where $\delta(x_i; c)$ is an indicator function

$$\delta(x_i; c) = \begin{cases} 1 & \text{if} \quad i \in t_L \\ 0 & \text{otherwise} \end{cases}$$

and $\epsilon \sim (0, \sigma^2)$.

The greedy search makes use of the indicator function in the linear model for splitting the data. This greedy search algorithm is shown to be biased towards variables with many levels. The greedy search is also shown to be attributable to the computational cost in terms of time spent in evaluating each possible binary split and the tendency of selecting a cutoff point close to the end of the data referred to as the end cut preference problem. Su et al. (2016) suggested replacing the indicator function with a Smooth Sigmoid Surrogate (SSS) function. A sigmoid function is a mathematical function with an, 'S' shape. This results in a non-linear model for estimating the best cutoff. The non-linear model can be appropriately reformulated into a one-dimensional optimization problem that can be quickly solved. Replacing the step function in the greedy search with the Smooth Sigmoid Surrogate function leads to the following interesting results. The SSS is observed to facilitate a parametric smoothing or regularization to the erratic splitting statistics in GS, yielding improved stability in the objective function to be optimized. As a result, SSS

are more capable of identifying weak signals. The smoothing effect also leads to substantial amelioration to the end-cut preference problem in practice. Furthermore, since the search of best cutoff is cast in a nonlinear regression framework, conventional statistical inference can be exploited to facilitate convenient comparisons across different predictors for finding the best split of data. SSS is shown to have potentially reduced computational cost without sacrifice in performance. SSS can be flexibly extended to many other recursive partitioning methods designated for different analytic purposes. This alternative splitting method is shown to alleviate the variable selection bias and is incorporated in this work.

2.6 Variable Selection Bias Correction Approaches

This section presents a review of literature on various approaches employed to solve the problem of variable selection bias in both regression and classification trees.

Classification trees are constructed based on the same idea underlying regression trees. This type of tree is used when the response variable is categorical or nominal. The Chi-square test statistics is used as the goodness of split measure or more commonly an impurity measure based on entropy or Gini index are used.

Since the best cutoff point in either regression or classification entails finding the maximum of the splitting statistics this method is broadly referred to as maximally selected statistics. Maximally selected statistics for the estimation of simple cutpoint models are embedded into a generalized conceptual framework based on conditional inference procedures,(see, e.g., Zeileis et al. (2008)).

For a binary response variable and a given continuous predictor variable, each point within the interval of the range of values of the predictor variable is selected to split the response variable into two. A total of N-1, 2×2 contingency table is formed with the numbers

of observation above and below the cut point in each sample, where N is the number of distinct levels of the predictor variable. Miller and Siegmund (1982) showed that when the cut point is selected to maximize the standard χ^2 statistic the χ^2 percentile points are inappropriate, the p-value is underestimated. The maximized chi-square was shown to converge to a normalized Brownian bridge under the null-hypothesis of no association between X and Y, which is different from the known chi-square distribution. Some further work was done by Halpern (1982) in a simulation study to examine the distribution of the maximally selected chi-square statistic in the small sample case. Also Koziol (1991) derived the exact distribution of the maximally selected chi-square by a combinatorial approach.

In determining the association between a binary response variable and at least ordinal scaled predictor variable,Boulesteix (2006b) derived the exact finite-sample distribution of the maximally selected χ^2 statistic. In a further study Boulesteix (2006a) derived the exact distribution of the maximally selected chi-square statistic using a combinatorial approach, when the predictor variable is nominal with several categories. However, it is best applicable to scenarios when X has only a small or moderate number of distinct levels.

In the case when the outcome is ordered, quantitative or censored variable, Lausen and Schumacher (1992) developed the asymptotic null distribution of maximally selected rank statistic. This asymptotic null distribution of maximally selected rank statistic is then compared with Monte Carlo simulation results by using continuous predictive variable X (Lausen 1992; Lausen 2004).

Loh (2002) propose an algorithm for regression tree construction called GUIDE which controls bias by employing chi-square analysis of residuals and bootstrap calibration of significance probabilities. The algorithm fits a piecewise constant model at each node and the residuals are computed. Then the cases in the node are divided into two groups, with one group defined by the positive residuals and the other by the non-positive residuals. The

Pearson chi-square test is then used to detect associations between the signed residuals and groups of predictor values. If X is a c-category predictor, the test is applied to the $2 \times c$ table formed by the two groups of residuals as rows and the categories of X as columns. These studies are all response and predictor type specific. It is obvious not much studies have been devoted into addressing the problem of variable selection bias in the case where the response variable is continuous and the predictor variable is nominal. It is always a good thing to explore alternatives to methods so as to have a plethora of tools in the data analysis process. This helps to compare the performance of various methods on the same data to confirm results or to trace possible violations in an experimental design setup.

Chapter 3

Methodology

This chapter contains three proposed methods for evaluating binary splits on nominal predictor variables in a regression tree. Two of the proposed methods are explored into detail whiles the third one is considered for further research due to some difficulties which came up. The proposed methods are shown to address the problem of variable selection bias.

3.1 Proposed Methods

The first proposed method is used when the nominal predictor variable has a small number of distinct levels. The restriction on this method is due to the inability of the multivariate normal package **mvtnorm** and **mnormt** (see, R Core Team, 2016), which are employed in computing the p-value with dimension up to 1000. Since $2^{10-1} - 1 = 511 < 1,000 < 2^{11-1} - 1 = 1,023$, we are allowed to evaluate the best binary split on categorical predictors with up to K=10 distinct levels. The other two methods can be used even when K is large. In these latter two approaches we make use of the selection differential in Nagaraja (1981) and approximation of the indicator threshold function with a smooth sigmoid surrogate (SSS) function in Su et al. (2016), respectively.

To briefly explain the proposed methods are summarized as follows:

1. Exact distribution of the maximally selected splitting statistic for small K.

2. Maximizing the selection differentials over K-1 possible splits, assuming a balanced

design.

3. Estimating the Best cutoff point as a Parameter in a Parametric Nonlinear mixed-effect Model.

3.2 Exact distribution of the maximally selected splitting statistic for small K

Consider data that consist of $\{(y_i, x_i) : i = 1, \ldots, n\}$, where y_i is the continuous response and the predictor x_i is nominal with K distinct levels. Without loss of generality (WLOG), we assume that the response has been centered so that $\bar{y} = 0$. Let $\{n_k, \bar{y}_k, s_k^2\}$ denote the sample size, mean, and variance of observed response value in the k-th category or group, respectively, for $k = 1, \ldots, K$. Let μ_k denote the theoretical k-th population mean. To proceed, we assume $\bar{y}_k \sim \mathcal{N}(\mu, \sigma^2/n_k)$ with error variance σ^2, which holds by central limit theorem (CLT) as long as n_k is moderately large.

Let s denote such a split that bisects the data node t into the left child node t_L and the right child node t_R. Let n_L and n_R be the sample size in t_L and t_R, respectively.

Let $u_k = n_k \bar{y_k} = \sum_{i=1}^{n_k} y_i$ be the sum of response values in the k-th category. Let $\mathbf{u} = [u_k] \in \mathbb{R}^K$. By independence of the K groups, it follows that $\mathbf{u} \sim \mathcal{N}\{\mathbf{0}, \sigma^2 \cdot \mathrm{diag}(n_k)\}$.

With within-node variation being the impurity measure and irrelevant constants omitted, it can be shown (see, e.g.,Hawkins (1977)) that the best split $s^\star$ maximizes

$$Q(s) = \frac{\left|\sum_{i \in t_L} y_i\right|}{\sqrt{n_L \, n_R}}. \tag{3.1}$$

Let $q^\star = Q(s^\star)$ be the maximized quantity that is computed with an observed data set. To

assess the statistical significance of $s^\star$, consider the corresponding p-value given below

$$\begin{aligned}
\text{p-value} &= \Pr\left\{\max_s Q(s) \geq q^\star\right\} \;=\; 1 - \Pr\left\{\max_s Q(s) < q^\star\right\} \\
&= 1 - \Pr\left\{Q(s) < q^\star, \forall\, s\right\} \;=\; 1 - \Pr\left\{\frac{\left|\sum_{i\in t_L} y_i\right|}{\sqrt{n_L\, n_R}} \;<\; q^\star, \forall\, s\right\} \\
&= 1 - \Pr\left\{-q^\star\sqrt{n_L\, n_R} \;<\; \sum_{i\in t_L} y_i \;<\; q^\star\sqrt{n_L\, n_R}, \forall\, s\right\}. \qquad (3.2)
\end{aligned}$$

The statistical distribution of

$$u_L = \sum_{i\in t_L} y_i = \sum_{k\in t_L} u_k$$

for any split s is studied below. In other words, we shall consider the sum of response values for all possible subsets of the K categories that go to one child node. These include a total of $a_K = (2^{K-1} - 1)$ subsets for the partitioning purpose. Let $\mathcal{A}$ denote the set of all these a_K permissible subsets. We let $\mathbf{e}_A \in \mathbb{R}^K$ denote a vector with elements being 1 for positions in A and 0 otherwise for any element $A \in \mathcal{A}$ and form matrix $\mathbf{E} \in \mathbb{R}^{K\times a_K}$ that has $\{\mathbf{e}_A : A \in \mathcal{A}\}$ as columns. Also, introduce vector $\mathbf{n} = (n_k) \in \mathbf{R}^K$. It follows that the sample size $n_L(A)$ associated with A is $n_L(A) = \mathbf{e}_A^T\mathbf{n}$. Let $\mathbf{j} = (1) \in \mathbb{R}^K$ and hence $n = \mathbf{j}^T\mathbf{n}$ and $n_R(A) = (\mathbf{j} - \mathbf{e}_A)^T\mathbf{n}$. The p-value can be expressed in matrix form as:

$$\text{p-value} = 1 - \Pr\left\{-\mathbf{b} < \mathbf{E}^T\mathbf{u} < \mathbf{b}\right\}, \qquad (3.3)$$

where vector $\mathbf{b} \in \mathbb{R}^{a_K}$ has element $q^\star\sqrt{\mathbf{n}^T\mathbf{e}_A(\mathbf{j} - \mathbf{e}_A)^T\mathbf{n}}$ for every $A \in \mathcal{A}$.
In (3.3), the random vector $\mathbf{E}^T\mathbf{u}$ follows a degenerate multivariate normal distribution

$$\mathbf{E}^T\mathbf{u} \;\sim\; \mathcal{N}_{a_K}\left\{\mathbf{0}, \sigma^2 \cdot \mathbf{E}^T diag(n_k)\mathbf{E}\right\}.$$

A closer look at the variance-covariance matrix $\sigma^2 \cdot \mathbf{E}^T(n_k)\mathbf{E}$ shows that it has diagonal elements $\sigma^2\mathbf{e}_A^T diag(n_k)\mathbf{e}_A = n_A\sigma^2$ for every A and off-diagonal elements $\sigma^2\mathbf{e}_A^T diag(n_k)\mathbf{e}_{A'} = n_{A\cap A'}\sigma^2$, where n_A denotes the sample size in categories falling in A and similar meaning holds for $n_{A\cap A'}$.

Computation of the p-value given in (3.3) essentially involves multivariate normal probabilities, which are studied by (Genz, 1992) and others. In particular, σ^2 will be replaced with its unbiased estimate of σ^2 given by $\hat{\sigma}^2 = \sum_{k=1}^{K}(n_k - 1)s_k^2/(n - K)$. For relatively smaller sample sizes, the multivariate t distribution can be used instead.

3.3 Maximizing the selection differential over K − 1 possible splits

This approach uses the idea of sorting the averages of the continuous response corresponding to each category Breiman et al. (1984). Consider data with a continuous response Y and a nominal predictor variable X with K distinct levels. The average of the response values in each category are computed and arranged in an ascending order, $\bar{Y}_1 \leq \bar{Y}_2 \leq , \ldots, \leq \bar{Y}_K$. The nominal values in the category with the minimum average are assigned 1; the next category whose average is greater than the minimum average is assigned 2, this follows till the category with the maximum average is assigned K. The data now consists of a continuous response and an ordinal predictor variable. The best split maximizes the selection differential.

$$\begin{aligned} D_{k,K} &= \frac{1}{k}\sum_{i=K-k+1}^{\mathbf{K}} \bar{Y}_{(i)} n_i \\ &= \frac{1}{k}\sum_{i=K-k+1}^{\mathbf{K}} (\bar{Y}_{(i)}\sqrt{n_i})\sqrt{n_i} \\ &= \frac{1}{k}\sum_{i=K-k+1}^{\mathbf{K}} Z_i\sqrt{n_i} \end{aligned} \tag{3.4}$$

Where $Z_i = \bar{Y}_i\sqrt{n_i}$

Let $T_{obs} = D_{k^\star,n}$ be the maximized quantity computed with an observed data set. To assess the statistical significance of $k^\star$, consider the corresponding p-value given below.

$$
\begin{aligned}
\text{P - value} &= P\{\max_k D_{k,n} > T_{obs}\} \qquad (3.5)\\
&= 1 - P\{\max_k D_{k,n} \leq T_{obs}\}\\
&= 1 - P\{D_{k,n} \leq T_{obs},\, \forall\, K-1\}\\
&= 1 - p\{\frac{1}{k}\sum_{i=n-k+1}^{\mathbf{K}} Z_i\sqrt{n_i} \leq T_{obs},\, \forall\, K-1\}
\end{aligned}
$$

The distribution of $\frac{1}{k}\sum_{i=n-k+1}^{\mathbf{K}} Z_i\sqrt{n_i}$ is essential to determine the p-value, however a closed form distribution of $D_{k,n}$ is a challenge. A closed form distribution exists only for the case where the parent population is exponential,Nagaraja (1981). A transformation of the random variable Y into an exponential random variable is proposed in order to make use of the closed form distribution. The selection differential, is also limited to balanced designs which is unrealistic in tree modeling. We wish therefore to defer this proposed method for a further research.

3.4 Estimating the Best Cutoff Point as a Parameter in a Parametric Nonlinear Mixed-Effects Model.

This method is proposed for general cases including scenarios where the nominal predictor variable has a large number of levels. We attempt to estimate the best cutoff point for the nominal predictor as a parameter in a parametric nonlinear model. This is related to the SSS methodology implemented in Su et al. (2016). The nonlinear model derived is compared with the null model, which is the model with only the intercept parameter. The degrees of freedom associated with the chi-square statistic is estimated numerically. With the chi-square statistic and the corresponding degrees of freedom we compute the p-value associated with the binary split.

3.4.1 Model Specification

Let's ponder the mechanism that underlies the data. The original model involves the discrete threshold indicator function, which is then approximated by an SSS function. This leads to a smooth parametric mixed-effects model, on which basis LRT may be applicable. First of all we assume $\{\mu_k\}_{k=1}^{K}$ come from either of random effects-models $\mu_k \sim N(\mu_L, \tau^2)$ or $N(\mu_L, \tau^2)$, yet which one underlies μ_k is unknown.

The nonlinear model is specified as :

$$\begin{aligned} y = \quad & z_1\{\mu_L I(\mu_1 \leq c) + \mu_R I(\mu_1 > c)\} + \quad z_2\{\mu_L I(\mu_2 \leq c) + \mu_R I(\mu_2 > c)\} \quad + \quad \ldots \\ & + \quad z_k\{\mu_L I(\mu_k \leq c) + \mu_R I(\mu_k > c)\} + \quad \epsilon \end{aligned}$$

where z_k is an indicator function

$$z_k = \begin{cases} 1 & \text{if} \quad k^{th} \quad \text{group} \\ 0 & \text{otherwise} \end{cases}$$

and $\epsilon \sim (0, \sigma^2)$.

The model involves the following parameters: $(\mu_1, \mu_2, \ldots, \mu_k, c, \mu_L, \mu_R, \sigma^2)$ of which $(\mu_1, \mu_2, \ldots, \mu_k, \sigma^2)$ are replaced with $(\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_k, S^2)$ obtained from the data and thus leaving (c, μ_L, μ_R) to be estimated.

The Smooth Sigmoid Surrogate function $s(c, \mu_i) = \pi(\mu_i - c;\text{a})$ is introduced to replace the indicator function,$I(\mu_i \leq c) = I(\mu_i, c)$ where $i = (1, \ldots, K)$. The parameter a in the notation is suppressed since it will be fixed a priori. By choosing a large a that corresponds to a sharper 'S' shape and estimating the c corresponding to the best fit, we are essentially finding a sharp step-function-type change in the data.

$$\begin{aligned} y = \quad & z_1\{\mu_L.s(\bar{y}_1, c) + \mu_R.s^{'}(\bar{y}_1, c)\} + \quad z_2\{\mu_L.s(\bar{y}_2, c) + \mu_R.s^{'}(\bar{y}_2, c)\} \quad + \quad \ldots \\ & + \quad z_k\{\mu_L.s(\bar{y}_k, c) + \mu_R.s^{'}(\bar{y}_k, c)\} + \quad \epsilon \end{aligned}$$

In addition if we know the value of c the parameters μ_L and μ_L can be estimated as:

$$\hat{\mu_L} = \frac{\sum_{k=1}^{K} n_k \bar{y}_k . s(\bar{y}_k, c)}{\sum_{k=1}^{K} n_k . s(\bar{y}_k, c)}$$

$$\hat{\mu_R} = \frac{\sum_{k=1}^{K} n_k \bar{y}_k . s^{'}(\bar{y}_k, c)}{\sum_{k=1}^{K} n_k . s^{'}(\bar{y}_k, c)}$$

The estimated response is then finally written as:

$$\begin{aligned} \hat{y} = \quad & \hat{\mu_L}\{z_1 . s(\bar{y}_1, c) + z_2 . s(\bar{y}_2, c) + \ldots + z_k . s(\bar{y}_k, c)\} \\ + \quad & \hat{\mu_R}\{z_1 . s^{'}(\bar{y}_1, c) + z_2 . s^{'}(\bar{y}_2, c) + \ldots + z_k . s^{'}(\bar{y}_k, c)\} \end{aligned}$$

3.4.2 Estimating the Best Cutoff Point

The splitting statistic Δl used in CART can be treated as an objective function for c and rewritten as follows:

$$\Delta l(c) = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \left\{ \sum_{i \in t_L} (y_i - \bar{y_L})^2 + \sum_{i \in t_R} (y_i - \bar{y_R})^2 \right\}$$

$$= \frac{1}{n_L} (\sum_{i \in t_L} y_i)^2 + \frac{1}{n_R} (\sum_{i \in t_R} y_i)^2$$

$$= \frac{1}{n_L} (\sum_{i \in t_L} y_i)^2 + \frac{1}{n - n_L} (\sum_{i=1}^{n} y_i - \sum_{i \in t_L} y_i)^2$$

where $\{n_L, \bar{y}_L\}$ denote the sample size and the average response in the left child node, respectively, and similarly $\{n_R, \bar{y}_R\}$ for the right child node.

$\Delta l(c)$ is approximated by replacing the $\delta(x_i, c)$ in n_L and $\sum_{i \in t_L} y_i$ with $s(c; x_i)$ so that

$$n_L = \sum_{i=1}^{n} \delta(x_i; c) \approx \sum_{i=1}^{n} s(c; x_i)$$

$$\text{and } \sum_{i \in t_L} y_i = \sum_{i=1}^{n} y_i \delta(x_i; c) \approx \sum_{i=1}^{n} y_i s(c; x_i)$$

The approximated objective function, denoted as $\tilde{\Delta} l(c)$, becomes

$$\tilde{\Delta} l(c) = \frac{\mathbf{s}^T(\mathbf{y}\mathbf{y}^T)\mathbf{s}}{(\mathbf{j}^T\mathbf{s})} + \frac{(\mathbf{j\text{-}s})^T(\mathbf{y}\mathbf{y}^T)(\mathbf{j\text{-}s})}{\mathbf{j}^T(\mathbf{j\text{-}s})}$$

where $y = (y_i)$, $s = (s(c; x_i))$, and $j = (1)$ are n-dimensional vectors.

Suppose further that, WLOG, the response has been centered $\mathbf{y} := (\mathbf{I}_n - \mathbf{j}\mathbf{j}^T/n)\mathbf{y}$ so that $\sum_{i=1}^{n} y_i = 0$. It follows that $\sum_{i \in t_L} y_i = -\sum_{i \in t_R} y_i$ and hence $\Delta l(c)$ can be further simplified as $(\sum_{i \in t_L} y_i)^2 / n_L(n - n_L)$ up to some irrelevant constant. Its approximation $\tilde{\Delta} l(c)$ reduces to

$$\tilde{\Delta} l(c) = \frac{\mathbf{s}^T(\mathbf{y}\mathbf{y}^T)\mathbf{s}}{\mathbf{s}^T(\mathbf{j}\mathbf{j}^T)(\mathbf{j\text{-}s})}$$

3.4.3 Likelihood Ratio Test of the Reduced and Current model

The likelihood ratio test is used to compare the reduced model (model with only an intercept term) and the current model.

The Likelihood-ratio statistic is stated as:

$$\Delta G^2 = -2 \text{ Log L from reduced model} - (\text{-2 Log L from current model}) \tag{3.6}$$

Owning to the smooth data generating mechanism, we have the conjecture that this LRT follows chi-square distribution under the null hypothesis. However finding its df is difficult due to the complicated model mechanism. In comparing two linear models using the LRT, the degrees of freedom is determined by subtracting 1 from the extra number of parameters in the current model. This method of determining the degrees of freedom is not appropriate for comparing the nonlinear models. A naive approach is to consider that the maximized split has degree of freedom 1 and thus use the chi-square statistic value from the LRT with the 1 degree of freedom to compute the p-value. We estimate the degrees of freedom via

monte carlo numerically. A histogram overlayed with a histogram density curve is plotted with the values obtained from 1000 simulation runs and seen to be best approximated by the chi-square density curve. The degrees of freedom is then estimated to be the value at which the chi-square density best approximates the histogram density curve. This is followed by examining how the degrees of freedom is related to either the sample size or the number of distinct levels the nominal predictor has through a simulation study in other to develop a model to estimate the degrees of freedom.

Chapter 4

Results And Analysis

In this chapter, we present some interesting results from our study. In the first part of this chapter we present a detailed report on how the degrees of freedom is estimated numerically. In the second part we present two simulation examples and a real data example on building regression trees with nominal predictor variables as an illustration to compare our methods with the maximally selected statistics approach of determining a binary split on nominal predictors.

4.1 Numerical Estimation of Degrees of Freedom

For the numerical estimation of the degrees of freedom we start with a plot of histograms overlayed with a histogram density curve (red) which is approximated with a chi-square density curve (green), Figure 4.1 and 4.2. We generated two data sets each consisting of a continuous response variable $Y \sim N(0, 1)$ and a nominal predictor variable. The nominal predictor variables had distinct levels 10 and 50 respectively. For each data set two sample sizes $\{200, 1000\}$ are generated. The LRT values were obtained through a 1000 simulation run using equation 3.6. The estimate of the degrees of freedom for K=10 in the first data set, Figure 4.1 did not change much even with an increase in sample size from 200 to 1000. A similar observation is made for K=50 in Figure 4.2.

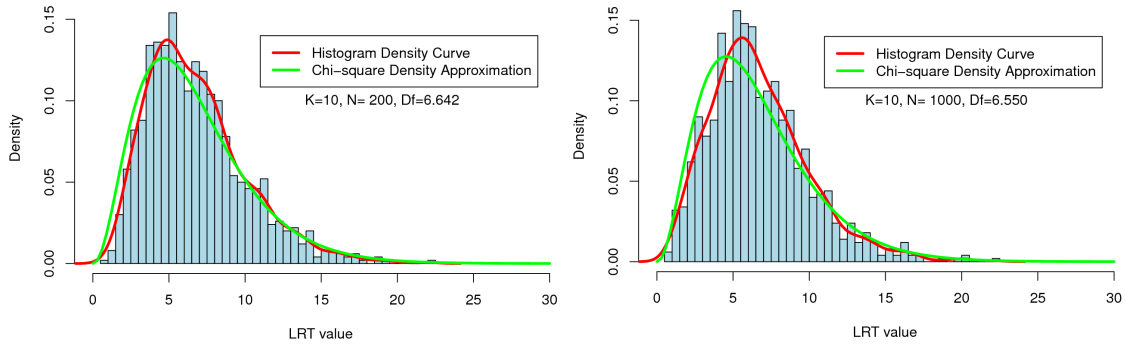


Figure 4.1: Degrees of Freedom for K = 10

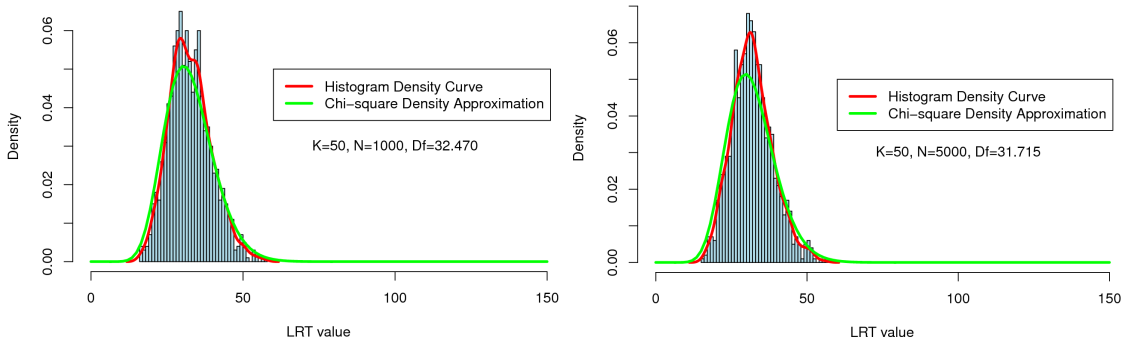
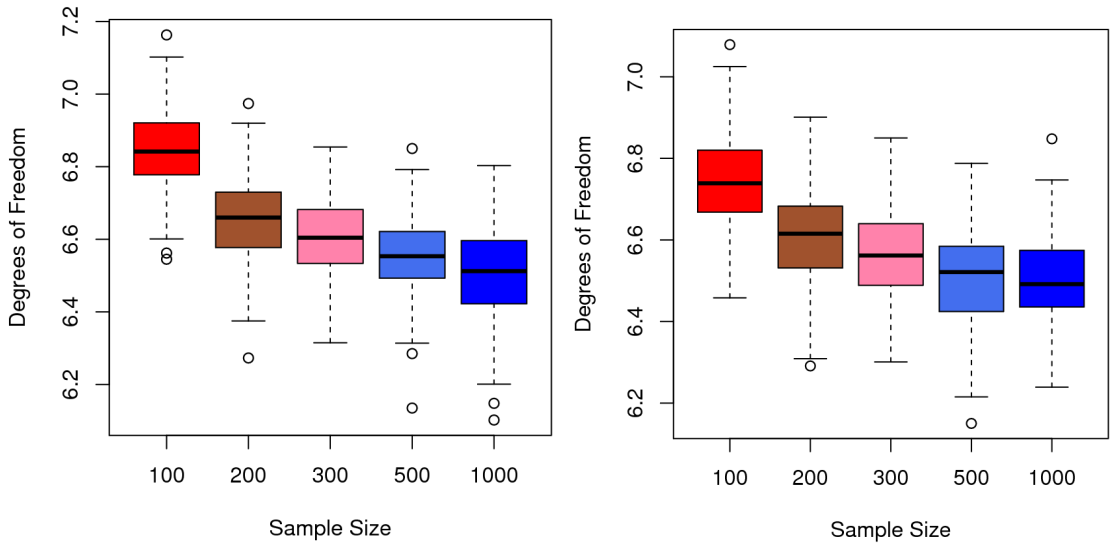


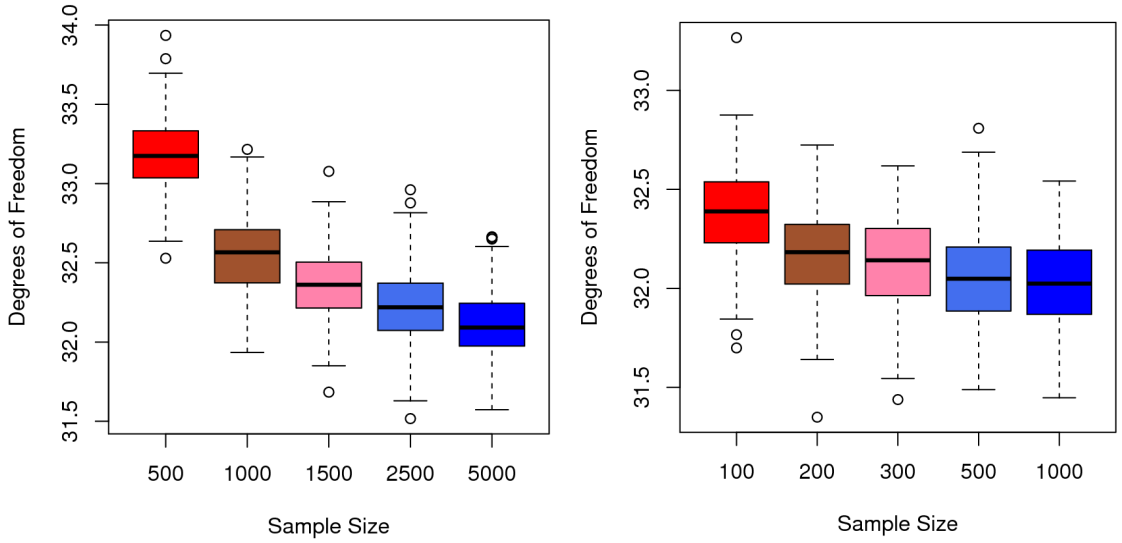
Figure 4.2: Degrees of Freedom for K = 50

In a similar fashion we examine the effect that the number of observations in each category of the nominal predictor may have on the degrees of freedom as the overall sample size of the data increases by comparing their boxplots. Two scenarios, a balanced and an unbalanced case are considered for each K used. The balanced case has an even number of observations in each category whiles the unbalanced case ensures that there is at least $3/4^{th}$ the number of observations in each category in the balanced case. For a sample size of 100 with K=10 each category had 10 observations in the balanced cased. In igures 4.3 and 4.2, each boxplot was constructed using 200 realizations of degrees of freedom computed through a 1000 simulation run each. Apart from the case where the number of observations in each category was 10, the other scenarios did not show much difference in the median degrees of freedom as the sample size increases in both the balanced and unbalanced case as shown in the boxplots.



(a) Balanced **(b)** Unbalanced

Figure 4.3: Boxplot for Degrees of Freedom for K = 10



(a) Balanced **(b)** Unbalanced

Figure 4.4: Boxplot for Degrees of Freedom for K = 50

Numerical summaries, mean and standard deviations in tables 4.1, 4.2 and 4.3 of the degrees of freedom were also computed for the degrees of freedom values obtained in the

simulation scenario explained above. The first column in each table represents the number of distinct levels whiles the other columns represent the number of observations in each category. We observe that the mean and standard deviations corresponding to each level under the various sample sizes did not change much even as the sample size increases.

Table 4.1: Mean and Standard deviation for unbalanced observations in each category for K=10

N	100	200	300	500	1000
K=10	6.743(0.115)	6.603(0.114)	6.568(0.112)	6.514(0.111)	6.497(0.106)

Table 4.2: Mean and Standard deviation for unbalanced observations in each category for K=50

N	500	1000	1500	2500	5000
K=50	32.391(0.228)	32.169(0.226)	32.133(0.223)	32.053(0.226)	32.030(0.220)

Table 4.3: Mean and Standard Deviation of Degrees of Freedom for Balanced case.

n	10	20	30	50	100
K=5	3.450(0.091)	3.323(0.094)	3.281(0.085)	3.258(0.088)	3.218(0.089)
K=10	6.846(0.115)	6.659(0.112)	6.604(0.108)	6.553(0.111)	6.589(0.120)
K=20	13.440(0.133)	13.144(0.154)	13.053(0.140)	12.983(0.146)	12.925(0.139)
K=50	33.184(0.236)	32.555(0.250)	32.365(0.218)	32.233(0.227)	32.113(0.223)

The thorough simulation study done above shows clearly that the degrees of freedom is only related to the number of distinct levels the nominal predictor has. The degrees of freedom for each nominal predictor with distinct levels 5 through 100 are therefore computed and presented in Table 4.4.

Table 4.4: Degrees of Freedom for K= 5 through K=100

Levels(K)	Df	Levels(K)	Df	Levels(K)	Df	Levels(K)	Df
5	3.124	30	19.323	55	35.292	80	51.706
10	6.600	35	22.516	60	38.787	85	54.807
15	9.612	40	25.525	65	41.561	90	57.370
20	12.879	45	28.851	70	45.244	95	60.930
25	16.328	50	31.812	75	48.132	100	64.433

A plot of the degrees of freedom against the number of distinct levels shows a positive linear relation. The simple linear regression equation representing this relation is presented in figure 4.5 and a summary of the model in Table 4.5.

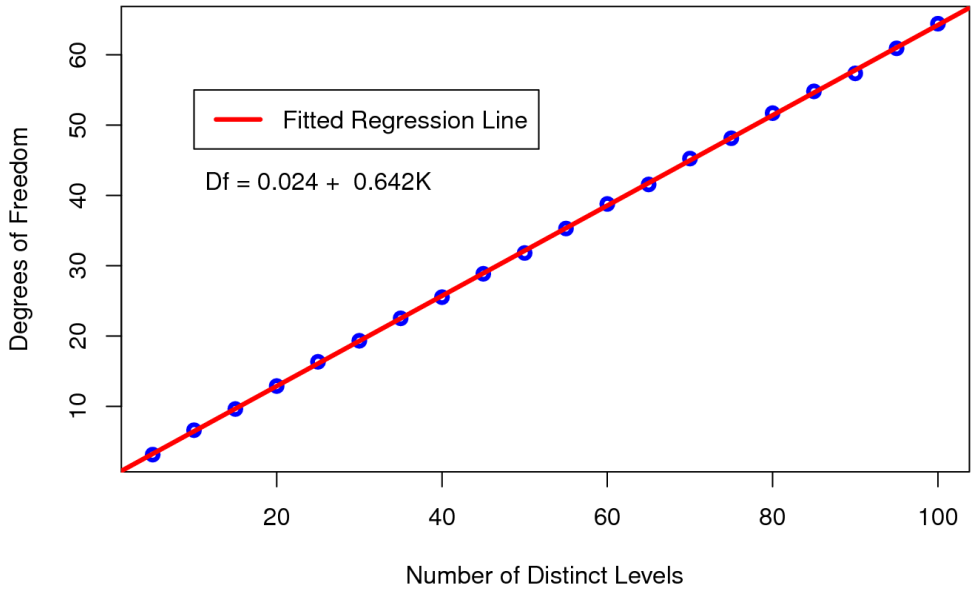


Figure 4.5: Linear Relation between Degrees of Freedom and Distinct levels of a nominal variable

Table 4.5: Summary of Linear Regression Model

Coefficients	Estimates	Std.Error	t value	$Pr(>\|t\|)$
Intercept	0.024	0.100	0.246	0.809
Df	0.642	0.002	392.614	$< 2e^{-16}$

4.2 Illustration Examples for our Proposed Methods

Two simulation examples and one real data example are presented in this section. In each of the examples we compute and compare the maximized test statistics (MTS) for a binary split on each predictor, a p-value referred to as the naive p-value is also calculated using the chi-square value from the LRT corresponding to the maximized test statistic with 1 degree of freedom, the p-value from our exact distribution of the maximally selected splitting statistic for small K and the p-value from our estimating the Best cutoff point as a Parameter in a Parametric Nonlinear mixed-effect Model. For easy of representation we denote each of these quantities as MTS, Naive P-value, Exact P-value and Nonlinear P-value respectively.

4.2.1 Simulation Example 1

Data is simulated from the model $Y = \beta I.(x_1 \in A) + \epsilon$ where $\epsilon \overset{IID}{\sim} \mathbf{N(0,1)}$ and X_1 is a nominal variable with two levels. Three values for $\beta \in \{0.0, 0.5, 1.0\}$ corresponding to null, medial and strong signals respectively for sample sizes 100 and 500 are used. Another nominal predictor variable X_2 with 10 levels is also simulated which has no relation with the response variable Y.

In table 4.6 we observed that for $\beta = 0.00$ corresponding to a null signal strength and for the two sample sizes 100 and 500 respectively, determining a single split by using the maximally selected statistics approach suffers from variable selection bias. The naive p-value computed also agreed with the wrong split decision made by the MTS approach. Our two methods: the exact p-value and the nonlinear p-value yielded accurately probability

values which were non significant and thus did not permit a split on either of the predictor variables since none of them is related to the response Y. For the medial and strong signal strength the four methods where able to determine accurately a split on X_1.

Table 4.6: Results for a single split on the predictor variables.

β	N	Predictor	MTS	Naive P-value	Nonlinear P-value	Exact P-value
0.00	100	X_1	0.151	0.145	0.204	0.331
		X_2	0.237	0.022	0.562	0.910
	500	X_1	0.059	0.186	0.256	0.344
		X_2	0.144	0.001	0.134	0.639
0.05	100	X_1	0.283	$9.938e^{-11}$	$2.057e^{-10}$	$1.390e^{-06}$
		X_2	0.111	0.013	0.455	0.639
	500	X_1	0.228	$9.265e^{-07}$	$1.769e^{-06}$	0.000
		X_2	0.061	0.195	0.962	0.998
1.00	100	X_1	0.486	$2.373e^{-06}$	$4.482e^{-06}$	0.000
		X_2	0.266	0.013	0.456	0.868
	500	X_1	0.500	$3.851e^{-26}$	$9.234e^{-26}$	$3.553e^{-15}$
		X_2	0.104	0.036	0.677	0.941

4.2.2 Simulation Result 2

We need to be careful not to consider any split on a nominal variable with many distinct levels as due to **VSB** since such variables could also be the right variables to induce a split on the response variable. We therefore simulate data of size 200, where X_1 and X_2 are nominal predictors with 7 and 3 levels respectively. X_1 is related with Y and expected to be selected to split the data. The parameter values used to simulate Y based on the levels of X_1 are $\{3.5, 2.2, 1.05, 2.95, 1.12, 1.45, 1.05\}$ with noise $\epsilon \overset{IID}{\sim} \mathbf{N(0,1)}$. Thus we have a data set with a continuous response variable Y and two nominal variables with a known relation

between Y and X_1 only. We apply both the exact p-value method and the parametric nonlinear mixed-effects method since the levels of the two nominal predictors used are less than 10. We then compare the binary split decision made by our two methods to the naive p-value approach and the maximally selected statistic approach. The four methods unanimously selected X_1 as the variable to split on, table 4.7.

Table 4.7: Maximized Test Statistics (MTS) and P-values

Predictor	Levels	MTS	Naive P-value	Nonlinear P-value	Exact P-value
X_1	7	0.963	$8.536e^{-30}$	$2.067e^{-26}$	$3.728e^{-07}$
X_2	3	0.188	0.056	0.155	0.447

We also compared the percentages of the number of times out of a 1000 simulation run the naive p-value is able to select X_1 as the right variable to split on compared to using the p-value obtained from the nonlinear approach. In this simulation study we considered two predictor variables with levels K=2 and K=50 respectively. The response variable Y is simulated to be related with only X_1 using $\beta = \{0.1, 0.5, 1\}$ for N=500 and N=1000 respectively. For a signal strength as low as $\beta = 0.1$ our parametric nonlinear mixed-effects method was able to detect correctly 59% and 74% for N=500 and N=1000 respectively that the split should be based on X_1, table 4.8, whiles the naive p-value was not able to select X_1 as the right variable to split on, even with an increase in sample size. For $\beta = 0.5$ our nonlinear p-value approach was able to perfectly select X_1 as the variable to split on in all the simulation runs unlike the naive p-value which got better under only a strong signal strength and a large sample size.

Table 4.8: Comparing the percentages of the naive and nonlinear method to select X_1 as the variable to split on.

β	N	Naive P-value	Nonlinear P-value
0.1	500	0.00	0.59
	1000	0.00	0.74
0.5	500	0.44	1.00
	1000	0.91	1.00
1.00	500	1.00	1.00
	1000	1.00	1.00

4.2.3 Real Data Examples

Automobile data set (1985) with 205 observations from the Regression category at UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/) was accessed and used for the illustration of the performance of our methods. The original data set consists of three types of entities: specification of an auto in terms of various characteristics, insurance risk rating and normalized losses in use as compared to other cars. Only the nominal predictor variables on specification of an auto in terms of various characteristics are used to illustrate our method of evaluating binary splits on nominal inputs. The response variable is the price at which the car was sold in the US, in thousands of US dollars. The log of price is taken and centered.

A description on the number of distinct levels of each nominal predictor variable used is in table 4.9. Some of the predictor variables such as number of cylinders and number of doors may look ordinal but they are nominal since we cannot for example classify a car with two doors as better than a car with four doors or vice versa. To determine the variable to split on we compute the maximized test statistic for each predictor and the p-values by the naive method and using our two methods. The maximally selected statistic approach

used the maximized test statistics values in column **MTS** to split the data on number of cylinders since it had the highest maximized test statistics value whiles our exact method and nonlinear parametric method chose a split on Drive wheels since it had the smallest p-values in their column, table 4.9.

Table 4.9: Maximized Test Statistics(**MTS**) and P-values

Predictor	Levels(K)	**MTS**	Naive P-value	Exact P-value	NonLinear P-value
Fuel Type	2	0.0689	0.050	0.5366	0.0774
Number of doors	3	0.0529	0.134	0.9005	0.3198
Drive Wheels	3	0.3434	$1.425e^{-29}$	0.0081	$1.9209e^{-28}$
Engine Type	7	0.2032	$1.842e^{-9}$	0.1769	$2.3971e^{-7}$
Number of Cylinders	7	0.3507	$2.757e^{-31}$	0.3340	$7.6561e^{-28}$
Fuel System	8	0.3475	$1.559e^{-30}$	0.3290	$1.2873e^{-26}$

Chapter 5

Discussion and Conclusion

In this chapter we present a discussion on how our two methods: Exact distribution of the maximally selected splitting statistic for small K and Estimating the Best cutoff point as a Parameter in a Parametric Nonlinear mixed-effect Model compare to the maximally selected statistics in determining a split on a continuous response when the predictor is nominal. We would also present a suggestion for further research and a conclusion based on our findings.

5.1 Discussion of Results

In the numerical estimation of the degrees of freedom, 1000 simulation runs were considered for each histogram plotted under the varying sample sizes and distinct levels of the nominal predictor. The chi-square density curve approximates well the histogram density curve, figure 4.1 and 4.2. In figure 4.3 and 4.4 we further explored how likely the sample sizes in each category can affect the estimation of the degrees of freedom. We observe the less variability in the 200 observations for the degrees of freedom under each sample size consider both for the balanced and unbalanced cases noting that with 10 samples in each category the variability was quite pronounced. These results are consistent with the mean and standard deviations for the degrees of freedom estimates in tables 4.1, 4.2 and 4.3. In figure 4.5 the relation between the degrees of freedom and the distinct levels of the nominal predictors is a positive linear which is quantified as $df = 0.024 + 0.642K$. With this equation the degrees of freedom for the chi-square value in our parametric nonlinear mixed-effects method is estimated and used with the chi-square value to compute the p-

value corresponding to the best cutoff.

In the simulation example 1 to illustrate our methods for the null signal case where $\beta = 0.00$ meaning that neither X_1 nor X_2 is related with the response Y, the maximally selected statistic approach considered a binary split on X_2, a case of **VSB** due to the high number of levels X_2 has compared to X_1. Our two methods were able to detect that there was no relation between the response and the two predictors since the p-values yielded were non-significant and thus did not allow a split on any of the two predictors. For the medial and strong signal cases the maximally selected statistic approach made decisions which were in the same direction as our p-value methods. In the second simulation example we see that our methods were able to detect rightly a binary split on a variable with many levels.

Following on with our real data example, a single split on the number of cylinders partitions the data based on {two,three and four} and {five,six,eight and twelve} whereas a split on Drive wheels partitions the data based on {Front-wheel, Four-wheel} and {Rear-wheel}. Difficulty interpreting the results of a tree are usually due to **VSB** in most cases. A search on how the number of cylinders or drive wheels can influence the price of a car shows that drive wheel is more likely to influence the price of a car than number of cylinders especially in a US market. Owning to the fact that it snows usually in the US more car buyers are likely to buy cars which can guarantee them safety when driving in the snow. Our check shows that Four-wheel and Front-wheel drives guarantee more safety in the sown than Rear-wheel drives. Thus a split on drive wheels based on this known fact by our method makes the interpretation of the tree more meaningful than a split on the number of cylinders where higher number of cylinders does not necessarily guarantee a better performance of the automobile.

5.2 Conclusion

The simulation studies and the real data example gave much insight into the performance of our proposed methods. Based on our findings we make the following conclusions on building regression trees with nominal predictor variables.

1. Our study has confirmed the need to use p-values instead of directly comparing the values of the maximally selected statistics for the binary split of a data.

2. Comparing the p-values which corresponds to the maximized test statistic for each nominal predictor variable in considering a binary split on a data is better than the maximally selected statistic approach.

3. Either the exact method for predictors with distinct levels of at most 10 or the parametric nonlinear mixed-effects model are suitable for computing the p-value to split the data.

5.3 Recommendation for Future Work

It is our aim that further research be conducted into the use of the selection differential to determine the binary split. Even though the selection differential is ideal for balanced designs we may either consider merging some categories in the data to achieve the balance or sampling an equal number of observations from each category so as to apply this method.

References

Boulesteix, A.-L. (2006a). Maximally selected chi-square statistics and binary splits of nominal variables. *Biometrical Journal*, 48(5):838–848.

Boulesteix, A.-L. (2006b). Maximally selected chi-square statistics for ordinal variables. *Biometrical Journal*, 48(3):451–462.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Doyle, P. (1992). The use of automatic interaction detector and similar search procedures. *Operational Research Quarterly*, pages 465–467.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of computational and graphical statistics*, 1(2):141–149.

Halpern, J. (1982). Maximally selected chi square statistics for small samples. *Biometrics*, pages 1017–1023.

Hawkins, D. M. (1977). Testing a sequence of observations for a shift in location. *Journal of the American Statistical Association*, 72(357):180–186.

Koziol, J. A. (1991). On maximally selected chi-square statistics. *Biometrics*, pages 1557–1561.

Lausen, B. and Schumacher, M. (1992). Maximally selected rank statistics. *Biometrics*, pages 73–85.

Loh, W.-Y. (2002). Regression tress with unbiased variable selection and interaction detection. *Statistica Sinica*, pages 361–386.

Miller, R. and Siegmund, D. (1982). Maximally selected chi square statistics. *Biometrics*, pages 1011–1016.

Nagaraja, H. N. (1981). Some finite sample results for the selection differential. *Annals of the Institute of Statistical Mathematics*, 33(1):437–448.

Shih, Y.-S. and Tsai, H.-W. (2004). Variable selection bias in regression trees with constant fits. *Computational statistics & data analysis*, 45(3):595–607.

Su, X., Kang, J., Liu, L., Yang, Q., Fan, J., and Levine, R. A. (2016). Smooth sigmoid surrogate (sss): An alternative to greedy search in recursive partitioning. *National University of Singapore - Institute of Mathematical Science*.

Torgo, L. (2001). A study on end-cut preference in least squares regression trees. In *Portuguese Conference on Artificial Intelligence*, pages 104–115. Springer.

Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17(2):492–514.

Appendix

R CODES

```
#########################################################################
# FUNCTIONS IN HANDLING A CATEGORICAL (CLASS) VARIABLE
#########################################################################


# ================================================
# FUNCTIONS FOR SSS (SMOOTH SIGMOID SURROGATE)
# ================================================

# THE EXPIT FUNCTION
expit <- function(x) (tanh(x/2)+1)/2   # CORRECT & VERIFIED

# TELL IF A NUMBER IS EVEN
is.even <- function(x) x %% 2 == 0

# THE OBJECTIVE FUNCTION USED IN LEAST SQUARES WITH CONTINUOUS RESPONSE
obj.LS <- function(c, a=50, y, x, scale.y=T){
SS <- NA; n <- length(y)
grp <- expit(a*(x-c)); n.L <- sum(grp); sum.L <- sum(y*grp) # THE ONLY
PLACE THAT NEEDS APPROXIMATION
if (scale.y) SS <- sum.L^2/(n.L*(n-n.L))
else {
```

```
n.R <- n- n.L; sum.R <- sum(y) - sum.L
SS <- sum.L^2/n.L + sum.R^2/n.R
}
return(-SS)
}
# IN FACT, STANDARDIZATION OF y WON'T CHANGE THE CHOICE OF c.star.


# -------------------------------------
# USING obj.ttest() SEEMS ADVANTAGEOUS
# -------------------------------------
obj.ttest <- function(c, a=10, y, x, scale.y=T){
if (scale.y) y <- scale(y, center = T, scale = T)  # STANDARDIZATION OF Y
 MIGHT HELP WITH NUMERICAL STABILITY (PREFERRABLE)
score <- NA; n <- length(y)
grp <- expit(a*(x-c))
n1 <- sum(grp); n0 <- n- n1
y1 <- y*grp; y0 <- y*(1-grp)
ybar1 <- sum(y1)/n1; ybar0 <- (sum(y)-sum(y1))/n0
sp2 <- (sum(y^2) - n1*ybar1^2 - n0*ybar0^2) / (n-2)  # COMPUTE POOLED S2
t <- (ybar1-ybar0)/sqrt(sp2 *(1/n1 + 1/n0))
score <- t^2
return(-score)
}

# FIND THE BEST CUTOFF POINTS FOR A CONTINUOUS VARIABLE X
bestcut.LS <- function(x, y, a=NULL, scale.y=T, alpha.endcut=.02,
method=c("ReducedSS", "ttest"), multi.start=T, n.starts=5)
```

```
{
n <- length(x)
# FIX a
if (is.null(a)) a <- sqrt(n)
# FINDING THE SEARCH RANGE TO AVOID ENDCUT PREFERENCE PROBLEM
sigma <- sd(x); mu <- mean(x)
x <- scale(x)  # IMPORTANT TO STANDARDIZE x IN ORDER TO APPLY A CONSTANT a
LB <- quantile(x, probs = alpha.endcut); UB <- quantile(x, probs =1-alpha.endcut);
if (method=="ReducedSS") obj <- obj.LS
else obj <- obj.ttest
if (multi.start==T) {
B <- seq(LB, UB, length.out=n.starts)
Q.min <- 1e15
for (b in 2:n.starts) {
OPT <- optimize(obj, lower=B[b-1], upper=B[b], maximum=F,
a=a, y=y, x=x, scale.y=scale.y)
if (OPT$objective < Q.min) {
Q.min <- OPT$objective
cstar <- OPT$minimum
}
}
} else {
cstar <- optimize(obj, lower=LB, upper=UB, maximum=F,
a=a, y=y, x=x, scale.y=scale.y)$minimum
}
cstar <- cstar*sigma + mu # TRANSFORM BACK
return(cstar)
}
```

```
# ====================================================================
# FUNCTION THAT SORTS THE LEVELS OF A CATEGORICAL (CLASS) VARIABLE
# ====================================================================


order.categories <- function(dat, col.y, cols.cat, details=T){
results <- list(NULL)
vnames <- colnames(dat)
y <- dat[, col.y]
p <- length(cols.cat)
OUT <- as.list(1:p)
names(OUT) <- colnames(dat)[cols.cat]
for (j in 1:p){
col.cat <- cols.cat[j]
vname <- vnames[col.cat]
x <- as.character(dat[, col.cat])
x.level <- sort(unique(x))
out <- NULL
if (details) print(x.level)
M0 <- aggregate(y, by=list(x), FUN=mean)
M0 <- data.frame(vname, M0[order(M0$x),])
colnames(M0) <- c("var", "level", "mean")
M0$group <- 1:NROW(M0)
if (details) print(M0)
x1 <- ordered(x, levels = M0$level)
```

```
dat[, col.cat] <- as.numeric(x1)
OUT[[j]] <- M0
}
results$OUT <- OUT
results$dat <- dat
return(results)
}



# ============================================================
# EMPIRICAL NULL DISTRIBUTION OF LIKELIHOOD RATIO TEST (LRT)
# (NONLINEAR WITH SMOOTH SIGMOID)
# ============================================================

# CONTROL SETTING FOR FUNCTION rpart()
library(rpart)
control.0 <- rpart.control(minsplit = 6, minbucket=round(6/3),
cp = 0, xval = 1, maxdepth = 1)

# CONTROL SETTING FOR FUNCTION nls()
options(warn=-1)
ctr0 <- nls.control(maxiter = 50, tol = 1e-03, minFactor = 1/1024,
            printEval = FALSE, warnOnly = TRUE)

# FUNCTION mapvalues() IN {plyr} IS USED
# install.packages("plyr");
library(plyr)
```

```
source("Functions-SSS.R")
#set.seed(1011)

K <- 95 # NUMBER OF GROUPS
n0 <- 5000;   # NUMBER OF OBSERVATIONS IN EACH GROUP
nrun <- 1000
sigma <- 1;
a0 <- 50  # PARAMETER a IN SSS

LRT <- matrix(0, nrow=nrun, ncol=3)
for (i in 1:nrun){
print(i)
 #z <- sample(x=paste("a", 1:K, sep=""), size=100, replace=TRUE) # THIS
 CASE n IS THE TOTAL SAMPLE SIZE.
 #y <- rnorm(100, mean=0, sd=sigma); y <- y-mean(y)  # CENTERING


z <- rep(paste("a", 1:K, sep=""), rep(n0, K)) # IN THIS WAY, EVERY
GROUP HAS n OBSERVATIONS.
y <- rnorm(n0*K, mean=0, sd=sigma); y <- y-mean(y)  # CENTERING

# USING rpart
fit.rpart <- rpart(y~z, method="anova", control=control.0)
pred <- predict(fit.rpart, newdata=data.frame(z=z), type ="vector")
z.split <- as.numeric(as.factor(pred))-1
LRT[i, 1] <- 2*(logLik(lm(y~z.split)) - logLik(lm(y~1)))
```

```
# USING SSS APPROACH
# -------------------------
ymeans <- tapply(y, z, mean);
x <- mapvalues(z, from=names(ymeans), to=rank(ymeans))
x <- as.numeric(as.character(x))
dat0 <- data.frame(y=y, x=x)
c.sss <- bestcut.LS(x=x, y=y, a=a0, scale.y=T,
alpha.endcut=.02, method="ttest")

# LRT1 - SELF-COMPUTED
# ----------------------
s0 <- tanh(a0*(x-c.sss))  # EQUIVALENT TO s0 <- expit(a0*(x-c.sss))
LRT[i, 2] <- 2*(logLik(lm(y~s0)) - logLik(lm(y~1)))

# LRT2 - USING nls
# --------------------
fit0 <- nls(y~ beta0, data=dat0, start=list(beta0=0))
c00 <- ifelse(is.even(K), K/2, (K+1)/2)
fit1 <- nls(y ~ beta0 + beta1 * tanh(a0*(x-c0)), data=dat0,
start = list(beta0=0, beta1=0.5, c0 =c00),
algorithm="port", control=ctr0,
lower = c(-5, -5, 1),
upper = c(5, 5, K))
LRT[i, 3] <- as.numeric(2*(logLik(fit1) - logLik(fit0)))
print(LRT[i,])
}
head(LRT)  # NOTE THAT THE FIRST TWO COLUMNS ARE (NEARLY) IDENTICAL.
```

```
dir.create(path="./LRT/")
filename <- paste("./LRT/lrt-K", K, ".Rdata", sep="")
save(LRT, file=filename)


# ====================================
# FIND DF FOR CHI-SQUARE APPROXIMATION
# ====================================

chi2.approx <- function(df, x){
x <- sort(x); n <- length(x)
p <- c((1:(n-1))/n, (n-1)/n*1/100 + 1*99/100)
sum((x - qchisq(p, df=df))^2)
}
LRT
lrt  <- LRT[,2]

df0 <- optimize(chi2.approx, lower = 1, upper = K, x=lrt)$minimum;df0


setEPS()
postscript("k50100.eps",paper="special",horizontal=FALSE,
onefile=FALSE, height=5, width=7)

hist(lrt, prob=TRUE, col="lightblue", nclass=50,
xlim=c(0, 3*K),main="",xlab="LRT value")
lines(density(lrt), col="red", lwd=3)
```

```
curve(dchisq(x, df=df0), col='green', add=TRUE,lwd=3)

legend(x=60,y=0.05,legend=c("Histogram Density Curve",
"Chi-square Density Approximation"),
col=c("red","green"),lty=1,lwd=3)

text(x=100,y=0.03,label="K=50, N=5000, Df=31.715")
dev.off()
#

#============================
#Data Simulation For Example 1
#============================
set.seed(151)
n <- 100
x <- sample(LETTERS[1:2], size=n, replace=TRUE)
mu <- ifelse(x=="A", 0, 1)
y <- mu + rnorm(n)
y <- y-mean(y)
x2 <- sample(LETTERS[1:10], size=n, replace=TRUE)

#============================
#Data Simulation For Example 2
#============================
set.seed(141)
n=100
x0 <- sample(1:7, size=n, replace=TRUE)
x1 <- LETTERS[x0]
```

```
install.packages("plyr")
library(plyr)
mu <- mapvalues(x1, from=LETTERS[1:7], to=c(5.5, 1, 1.25, 8.7, 1.85, 2.7, 3.05))
mu <- as.numeric(mu)
y <- mu + rnorm(n)
x2 <- sample(LETTERS[1:3],size=n,replace=TRUE)



#=====================================================
#P-value for Parametric Nonlinear Mixed-Effects Model
#=====================================================
y <- y
z <- x
# CONTROL SETTING FOR FUNCTION rpart()
install.packages("rpart")
library(rpart)
control.0 <- rpart.control(minsplit = 6, minbucket=round(6/3),
cp = 0, xval = 1, maxdepth = 1)

fit.rpart <- rpart(y~z, method="anova", control=control.0)
pred <- predict(fit.rpart, newdata=data.frame(z=z), type ="vector")
z.split <- as.numeric(as.factor(pred))-1
LRT <- as.numeric(2*(logLik(lm(y~z.split)) - logLik(lm(y~1))));LRT
K <- length(unique(x))
df0 <- 0.024+K*0.642
p.value <- pchisq(LRT, df=df0, lower.tail=FALSE)
p.value
```

```
#=======================================
#Naive P-value using degrees of freedom 1
#=======================================
p.value <- pchisq(LRT, df=1, lower.tail=FALSE)
p.value

#================================================
#Computing the P-value For the Exact Distribution
#of the Maximally Selected Statistics
#================================================
library(rpart)
dat <- data.frame(y=y, x=x)
fit.rpart <- rpart(y~factor(x), data=dat, method="anova",
control=rpart.control(minsplit = 6, minbucket=round(6/3),
cp = 0, xval = 1, maxdepth = 1))
pred <- predict(fit.rpart, newdata=dat, type ="vector")
z.split <- as.numeric(as.factor(pred))-1
n.L <- sum(z.split==0)
n <- length(y)
n.R <- n-n.L
sum.L <- sum(dat$y[z.split==0])
q.star <- abs(sum.L)/sqrt(n.L*n.R) #Maximized Statistics


k <- length(unique(x))
fit <- lm(y~factor(x))
sigma2 <- (summary(lm(y~factor(x)))$sigma)^2; sigma2
```

```
grps <- sort(unique(x))
k <- length(grps)
n.k <- aggregate(y, by=list(x), FUN=length)$x
u <- aggregate(y, by=list(x), FUN=sum)$x

tmp <- expand.grid(rep(list(c(TRUE, FALSE)), k));
k1 <- 2^(k-1)-1
E <- as.matrix(tmp[2:(k1+1), ]+0)
row.names(E) <- NULL

mu.vec <- rep(0, k1)
Sigma <- sigma2* E %*% diag(n.k)%*% t(E) #Variance Covariance


# MULTIVARIATE NORMAL PROBABILITY
install.packages(c("mvtnorm", "mnormt"))
library(mvtnorm)
UB <- as.vector(q.star*sqrt((E%*%n.k)*((1-E)%*%n.k)))
LB <- -UB
pvalue <- 1- pmvnorm(lower=LB, upper=UB, mean=mu.vec, sigma=Sigma)
pvalue
```

Curriculum Vitae

Isaac Xoese Ocloo is a disciplined and determined young man with an inquisitive mind to determine statistically the causes and effects of problems inimical to human existence with the aim of finding ways to solve them.

After completing Chemu Senior High School, Tema community 4, where he studied Technical Drawing, he continued his college education a year later in the Kwame Nkrumah University of Science and Technology (KNUST), Kumasi where he pursued a bachelor's degree in Mathematics, graduating with First Class Honors. Isaac's excellent academic record earned him a position of a Teaching Assistant at KNUST for his one year national service upon graduation.

In Fall 2015, he entered the Graduate School of The University of Texas at El Paso (UTEP) as a graduate student to purse a master's degree in Statistics. While in UTEP, he worked as a Teaching Assistant and was assigned various duties such as grading and tutoring. Even though Isaac had some initial challenges in keeping up with his studies in his new environment his determination to succeed paid off when he was awarded with an Academic and Research Excellence Graduate Student Statistics award by the department of mathematical sciences, UTEP during their pre commencement.

He has drawn much inspiration from his academic advisor Dr. Xiaogang Su and other faculty at the department of Mathematical Sciences, UTEP to continue with his doctoral studies. In Fall 2017 Isaac will start his PhD. Statistics at Bowling Green State University Ohio.

Contact Information: ocloox1@gmail.com

This thesis was typed by Isaac Xoese Ocloo.