TREE-BASED MODELS FOR LONGITUDINAL DATA

Dan Liu

A Thesis

Submitted to the Graduate College of Bowling Green
State University in partial fulfillment of
the requirements for the degree of

MASTER OF SCIENCE

August 2014

Committee:

Peng Wang, Advisor

Hanfeng Chen

Junfeng Shang

ABSTRACT

Peng Wang, Advisor

Classification and regression trees (CART) have been broadly applied due to their simplicity of explanation, automatic variable selection, visualization and interpretation. Previous algorithms for constructing regression and classification tree models for longitudinal data suffer from the computational difficulties in the estimation of covariance matrix at each node. In this paper, we proposed regression and classification trees for longitudinal data, utilizing the quadratic inference functions (QIF). Following the CART approach and taking the correlation of longitudinal data into consideration, we developed a new criterion, named RSSQ, to select the best splits. The proposed approach could incorporate the correlation between the repeated measurements on the same subject without the estimation of correlation parameters. Therefore, the efficiency of the partition results and prediction accuracy could be improved. Simulation studies and real data examples are provided to illustrate the promise of the proposed approach.

KEY WORDS: Longitudinal data; Classification and regression trees; Quadratic inference functions.

ACKNOWLEDGMENTS

I would like to express my very deep gratitude to Dr. Peng Wang, my advisor, for his valuable and constructive suggestions during the planning and development of this research work. His willingness to give his time so generously has been very much appreciated. I would like to thank Dr. Hanfeng Chen and Dr. Junfeng Shang, the committee members, for their patient guidance and useful comments of this research work. I would also like to thank Dr. Tong Sun and Ms. Marcia L. Seubert, for their advices and assistance in keeping my progress on schedule.

Table of Contents

CHAPTER 1: INTRODUCTION **1**

CHAPTER 2: CLASSIFICATION AND REGRESSION TREES **5**

CHAPTER 3: GENERALIZED ESTIMATING EQUATIONS AND QUADRATIC INFERENCE FUNCTIONS **7**

CHAPTER 4: CART FOR LONGITUDINAL STUDIES **10**

4.1 Regression Trees 11

4.2 Classification Trees 15

CHAPTER 5: SIMULATION RESULTS **18**

5.1 Regression Trees 18

5.2 Classification Trees 21

CHAPTER 6: REAL-DATA EXAMPLES **23**

6.1 Regression Trees for Ozone Data 23

6.2 Classification Trees for Indonesia Children's Health Study 28

CHAPTER 7: DISCUSSION **32**

BIBLIOGRAPHY **34**

List of Figures

5.1 The simulation scenario for regression trees 19
5.2 The simulation scenario for classification trees 22

6.1 Plots of pairs of response and predictor variables for ozone data 24
6.2 The regression tree for Ozone Data by regular CART 26
6.3 The regression tree for Ozone Data by the RSSQ criterion 27
6.4 The classification tree for ICHS data by regular CART 30
6.5 The classification tree for ICHS data by the RSSQ(M) criterion 30

List of Tables

5.1 SRE of splits for regression trees 20
5.2 SRE of prediction for regression trees 21
5.3 SRE of predictive misclassification rate 22

6.1 Ozone data description 24
6.2 ICHS data description 28
6.3 Error rates of two approaches 31

CHAPTER 1

INTRODUCTION

The classification and regression trees (CART, Breiman et al., 1984) are tree-based nonparametric methods which offer piecewise-constant estimates of regression functions in single-response settings. The CART methods construct tree-based models by recursively partitioning the predictor space into distinct and non-overlapping regions. Then, the mean or mode of response values of the training observations in each region is used to make estimation or prediction for observations belonging to the same region. CART has been broadly applied due to its simplicity of explanation, automatic variable selection, visualization and interpretation.

Longitudinal studies involve repeated measurements from the same subject over a period of time. As a result, observations from the same subject are usually correlated. Most of longitudinal studies are intended to detect developmental trends of the same subject over the long period of time. In this project, we proposed a new approach to fit tree-based models to longitudinal data. The analytic goals are to identify strata with common covariate values, homogenize outcomes of longitudinal data, and detect developmental trends for each subject. We hope to construct tree-based models that accurately classifies observations to their classes, using only the predictor information. For example, in the Indonesian children's health study (ICHS, Sommer et al., 1984), preschool children in Indonesia were examined

for up to six consecutive quarterly visits for the presence of respiratory infection. A primary question of interest is whether vitamin A deficiency, as indicated by the occurrence of the ocular disease Xerophthalmia, is associated with a higher prevalence of respiratory infection. The chronic change in the prevalence of respiratory infection is also of interest. Since tree-based methods can select influential predictors automatically, we can fit a classification tree to ICHS data to address these questions, taking into account the correlation of repetitive measurements. The analyses of these questions are presented in section 6.2 in this paper.

Segal (1992) modified the node impurity functions based on likelihood-ratio statistics so that tree-constructed methodology can be extended to longitudinal data. He generalized the least square split function to multivariate by introducing covariance parameters of some simple variance models, such as compound symmetry and first-order autoregressive models.

To construct classification trees for binary multiple responses, Zhang (1998) developed generalized entropy criterion as the maximum of log-likelihood of an exponential family distribution. The estimates of parameters are obtained by the maximum likelihood estimation (MLE). Additionally, he applied Segal's (1992) impurity functions to binary multiple responses and found both criteria generated remarkably similar trees without stability problems. To extend this method to ordinal responses, Zhang and Ye (2008) transformed ordinal responses to binary responses by indicator functions.

From a different perspective, Yu and Lambert (1999) regarded longitudinal data responses as curves instead of vectors. They presented two approaches to reduce the curves into lower-dimensional vectors on which a regression tree was constructed. One approach is to employ a linear combination of spline basis functions to fit vector responses and then build a multivariate tree to the lower-order estimated coefficient vectors. The other is to treat first several principal component scores as a response vector. Both methods use squared error loss as impurity function.

Loh (2002) proposed the univariate GUIDE algorithm for regression trees to reduce the selection bias and computational difficulties. Unlike the one-step CART method which

selects the splitting variables and split set simultaneously, GUIDE is a two-step method: first, select a splitting variable with the smallest chi-squared p-value by contingency table chi-squared tests; second, search the maximum reduction in Residual sum of squares (RSS) in all possible partitions of the selected predictor variable. Since GUIDE can be directly applied to longitudinal data only if the observations have a fixed grid with a small number of grid points, an extended GUIDE to longitudinal data was developed. Loh and Zheng (2013) applied regression trees to longitudinal and multiresponse data with GUIDE by dividing the range of time into disjoint intervals of equal length. Then they fitted lowess (Cleveland, 1979) curve to all the subjects to estimate the mean of response values. A contingency table is constructed by indicator functions based on whether the number of observations above lowess curve is larger than that below lowess curve for each time interval. However, correlations of longitudinal response values are oversimplified by the smooth lowess curve, and it will be difficult to decide a threshold for p-values when there are a large number of time intervals.

Hsiao and Shih (2007) extended the curvature test of GUIDE by building a three-way contingency table with an additional dimension for the residuals of multivariate response component. The splitting variable is selected with the smallest associated p-value in the conditional independence test (Simonoff, 2003) for a three-way contingency table.

Unlike methods advocated by Breiman et al. in which partitioning predictor variables was done before fitting piecewise-constant model, Lee (2004) firstly applied the generalized estimating equations (GEE, Liang and Zeger, 1986) to fit a regression model based on the raw predictor variables in each node and then classified subjects into subgroups based on their signs of the averaged Pearson's residuals. Then a two-sample t-test for differences along each covariate between the two groups is performed and the best splitting variable is the one with the largest absolute t-statistic. If there is strong linear relation between the response variable and the predictors, Lee's method may improve the accuracy of tree-based models. However, Lee's approach may not work well for non-linear data sets since GEE is efficient

for only linear regression on longitudinal data.

In the framework of generalized linear model (GLM), Liang and Zeger (1986) proposed the generalized estimating equations (GEE) to estimate the regression parameters for longitudinal data by using a working correlation matrix which involves only a few nuisance parameters. This approach yields consistent estimates of the regression parameters and the variance matrices even when the assumed correlation is not correctly specified. Qu et al.(2000) improved GEE by introducing a method of quadratic inference functions (QIF) which approximates the inverse of the working correlation matrix with a linear combination of basis matrices. The QIF approach improves the efficiency of GEE estimators when the correlation structure is misspecified without the estimation of correlation parameters. Moreover, QIF provides an effective tool for statistical inference.

In this paper, we propose tree-based models for longitudinal data, utilizing QIF. The proposed approach could incorporate the correlation between the repeated measurements on the same subject without the estimation of correlation parameters. Therefore we could improve the efficiency of the partition results and the prediction accuracy. Chapter 2 concisely presented CART by Breiman (1984). In chapter 3, we briefly introduced GEE and QIF for longitudinal data. In chapter 4, a new criterion, Sum Squared Residuals of Quadratic Inference Functions (RSSQ), was proposed to construct tree-based models for longitudinal data. Chapter 5 showed the simulation results to evaluate the performance of RSSQ compared with CART. In chapter 6, we applied RSSQ on two real data sets to build regression trees and classification trees respectively. Finally, we discussed the strengths and weaknesses as well as further developments of the RSSQ method.

CHAPTER 2

CLASSIFICATION AND REGRESSION TREES

In this section, we provide a brief introduction on the classification and regression tree (CART) by Breiman et al. (1984) in single-response settings. CART is constructed by recursively partitioning the predictor space into J distinct and non-overlapping regions, $R_1, R_2, \ldots, R_J$. The partitions are conducted by binary splits with respect to the values of predictors in the form of $X \in A$ if X is categorical, and $X \leq c$ if X is ordinal. Then, the mean or mode of response values of the training observations in each region is used to make estimation or prediction for observations which belong to the same region. Therefore, CART estimates y_i, the response of subject i, based on the following piece-wise constant model,

$$f(X_i) = \sum_{j=1}^{J} c_j \cdot I(X_i \in R_j). \tag{2.1}$$

where X_i is the covariate vector, c_j is a constant and $I(X_i \in R_j)$ indicates whether X_i belongs to predictor region R_j or not. In the piecewise-constant model, the entries of the predictor matrix contain 1 and 0's only, and there is exactly one 1 in a row and the remaining are 0's. Since each observation must belong to exactly one terminal node, the predictor matrix of a piecewise-constant model is orthogonal and full rank. According to the CART methods,

parameter c_j can be estimated by the mean or mode of the outcome values belonging to the jth terminal node.

If the response variable is quantitative, a regression tree is built and the node impurity function is the residual sum of squares (RSS) $\sum_{j=1}^{J}\sum_{i\in R_j}\left(y_i-\hat{y}_{R_j}\right)^2$, where $\hat{y}_{R_j}$ is the sample mean of responses within the jth region. On the other hand, a classification tree is constructed for a data set with qualitative response variable. In a classification tree, it is desired to minimize the misclassification rate, simply the proportion of training observations in a region which do not belong to the most common class. However, misclassification rates are not sensitive enough to grow a tree, two alternative criteria (purity functions) for making binary splits are advocated in practice, namely Gini index, $G=\sum_{k=1}^{K}\hat{p}_{jk}(1-\hat{p}_{jk})$, and cross-entropy, $C=-\sum_{k=1}^{K}\hat{p}_{jk}log\left(\hat{p}_{jk}\right)$, where $\hat{p}_{jk}$ is the proportion of training observations in the jth region that are from kth category. Both Gini index and cross-entropy will take on a value close to zero if the $\hat{p}_{jk}$'s are all near zero or near one. By exhaustive search, a split which makes the maximum reduction in the splitting function (RSS, Gini index, or cross-entropy) is selected over all possible splitting variables and all possible cutoffs. CART continues recursively partitioning until the pre-specified stopping rule is reached, or either the y or the X values become constant in a node. Since the stopping rule of CART intends to grow a large tree, the tree may need to be pruned by cross-validation, and a subtree which leads to the lowest test error rate is preferred.

Although tree-based methods, to some extent, compromise their prediction accuracy when competing with some supervised learning approaches, such as linear regression, their simple interpretation, neat graphical presentation, and high efficiency for categorical data are attractive. Moreover, tree-based models may outperform classical approaches if the predictors and the response variable have a highly non-linear and complex relationship. Also, there are no assumptions of statical models in CART. Another advantage of CART is that the significant variables are selected as splitting variables automatically.

CHAPTER 3

GENERALIZED ESTIMATING EQUATIONS AND QUADRATIC INFERENCE FUNCTIONS

To establish notation for longitudinal data, let y_{it} be a response variable and x_{it} be a $p \times 1$ vector of covariates, observed at equally-spaced time points ($t = 1, \ldots, n_i$) for the ith subject ($i = 1, \ldots, N$). Thus, besides the $n_i \times p$ matrix of covariates $X_i = (x_{i1}, \ldots, x_{in_i})^T$, each subject has a $n_i \times 1$ vector of responses $y_i = (y_{i1}, \ldots, y_{in_i})^T$. In this paper, we assume that in a longitudinal data set the observations within the same subject are correlated and all subjects have the same correlation structure, but observations from distinct subjects are independent. In this section, we discussed the commonly applied approaches for longitudinal studies under the framework of generalized linear models. Liang & Zeger (1986) proposed the generalized estimating equations (GEE) to obtain the estimates of regression parameter β for nonnormal correlated longitudinal data. The covariance matrix of the ith response vector, V_i, is approximated by $A_i^{\frac{1}{2}} R(\alpha) A_i^{\frac{1}{2}}, i = 1, \ldots, N$, where $R(\alpha)$ is called the working correlation matrix which is fully characterized by a nuisance parameter vector α, A_i is a diagonal matrix in which the entries are the variances of observations within the ith subject.

As an extension of the score equations from a likelihood analysis under the independence assumption, the general estimating equations are defined as $\sum_{i=1}^{N} \dot{\mu}_i^T V_i^{-1}(y_i - \mu_i) = 0$, where $\dot{\mu}_i$ is the first derivative of μ_i with respect to regression parameters β. The estimator $\hat{\beta}$ is defined to be the solution of the generalized estimating equation.

Qu et al.(2000) improved GEE by developing a method of quadratic inference functions (QIF) in which a linear combination of basis matrices, $\sum_{k=0}^{m} a_k M_k$, was used to model $R(\alpha)^{-1}$. the matrices $M_0, \ldots, M_m$ are known and $a_0, \ldots, a_m$ are unknown constants depending on the correlation parameter α. The linear combination is able to accommodate most commomly used working structures for longitudinal studies. In the following, we illustrate the idea with examples of three commonly used working correlations: independent, AR-1 and exchangeable structures.

Example 1. (Independent) Suppose $R(\alpha)$ is an identity matrix which has 1's on the diagonal and 0's everywhere off the diagonal. Then R^{-1} is also an identity matrix and can be written as γM_0 where $\gamma = 1$, and M_0 is an identity matrix.

Example 2. (Exchangeable) Suppose $R(\alpha)$ has exchangeable correlation structure in which the diagonal entries are 1's and the remaining are $\alpha's$. Then R^{-1} is equal to $\gamma(M_0 + a_1 M_1)$, where M_0 is an identity matrix, M_1 has 0's on the diagonal and 1's elsewhere, $a_1 = -\alpha/\{(n-2)\alpha + 1\}$, $\gamma = \{(n-2)\alpha + 1\}/\{(n-1)\alpha^2 - (n-2)\alpha - 1\}$, and n is the dimension of R.

Example 3. (AR-1) Suppose $R(\alpha)$ has first-order autoregressive (AR-1) correlation structure in which the entries $R_{ij} = \alpha^{|i-j|}$. Ignoring the edge effect of AR-1 process, R^{-1} can be approximated by a linear combination of two matrices $\gamma(M_0 + a_1 M_1)$, where M_0 is an identity matrix, M_1 has 1 on the two main off-diagonals and 0 everywhere else, $a_1 = -\alpha/(1+\alpha^2)$ and $\gamma = (1+\alpha^2)/(1-\alpha^2)$.

Example 4. (Unconstructed) If $R(\alpha)$ is totally unspecified, then we use the inverse covariance matrix of the data set directly instead of the approximation $V_i^{-1} \approx A_i^{-\frac{1}{2}} R(\alpha)^{-1} A_i^{-\frac{1}{2}} (i = 1, \ldots, N)$.

Applying the linear approximation to GEE, we have

$$\sum_{i=1}^{N} \dot{\mu}_i^T A_i^{-\frac{1}{2}} (\sum_{k=0}^{m} a_k M_k) A_i^{-\frac{1}{2}} (y_i - \mu_i) = 0. \tag{3.1}$$

Instead of obtaining the exact values of $a_1, \ldots, a_m$ and solving (3.1), Qu et al. (2000) defined the extended score in the form of the quasi-score to be

$$g_N(\beta) = \frac{1}{N} \sum_{i=1}^{N} g_i(\beta) = \frac{1}{N} \begin{pmatrix} \sum_{i=1}^{N} \dot{\mu}_i^T A_i^{-\frac{1}{2}} M_0 A_i^{-\frac{1}{2}} (y_i - \mu_i) \\ \vdots \\ \sum_{i=1}^{N} \dot{\mu}_i^T A_i^{-\frac{1}{2}} M_m A_i^{-\frac{1}{2}} (y_i - \mu_i). \end{pmatrix} \tag{3.2}$$

The estimating equations in vector $g_N(\beta)$ can be combined optimally with the generalized method of moments (Hansen, 1982) since (3.1) is a linear combination of the extended score $g_N(\beta)$. The quadratic inference function is then defined as

$$Q_N(\beta) = g_N^T \, C_N^{-1} \, g_N, \tag{3.3}$$

where $C_N(\beta) = \frac{1}{N^2} \sum_{i=1}^{N} g_i(\beta) g_i^T(\beta)$ is a weighting matrix, and $\hat{\beta}$ is obtained by minimizing the quadratic inference function: $\hat{\beta} = \arg\min_\beta Q_N(\beta)$, without estimation of the nuisance parameters α. In the cases where the working correlation is misspecified, the estimator by the QIF approach is more efficient than that obtained by GEE. More importantly, QIF provides an inference vehicle, since $Q_N(\hat{\beta})$ has an asymptotic χ^2 distribution.

CHAPTER 4

CART FOR LONGITUDINAL STUDIES

The typical CART approaches ignore the correlation between observations, leading to inefficient model estimation (Wang et al., 2014). Here we extend CART to longitudinal settings with the QIF approach. The advantage of our approach is that unlike the current existing methods, we do not need to estimate the correlation parameters at each split, which greatly reduces the computational cost and improves the prediction accuracy.

We adapt the recursive partitioning procedure, dividing the predictor space into J distinct and non-overlapping regions, $R_1, R_2, \ldots, R_J$, and fitting piecewise-constant models at each step. Tree-based models for longitudinal data can be expressed as

$$Ey_{it} = h(U_{it}\beta^*) = h\left(\sum_{j=1}^{J} I(X_{it} \in R_j) \times \beta_j^*\right), i = 1, \ldots, N; t = 1, \ldots, n_i \qquad (4.1)$$

where h is a link function, $U_{it} = (I(X_{it} \in R_1), \ldots, I(X_{it} \in R_J))$ is a $1 \times J$ vector, and β^*, different from the linear regression parameters β, is the parameter vector for the tree-based model. Here R_j is the jth $(j = 1, \ldots, J)$ predictor region, and $I(X_{it} \in R_j)$ $(j = 1, \ldots, J)$ are indicator functions denoting the terminal node that an observation belongs to. To establish matrix notation, let $U_i = (U_{i1}^T, \ldots, U_{in_i}^T)^T, i = 1, \ldots, N$ be the predictor matrices for the

ith subject and then $E(y_i) = h(U_i\beta^*)$; also, let $U = (U_1^T, \ldots, U_N^T)^T$ be the predictor matrix for the whole data set. In the tree-based piecewise-constant models, the predictor matrix consists of 1 and 0 only, and there is exactly one 1 in a row and the remaining entries are 0's, so the predictor matrix of a piecewise-constant model is orthogonal and full rank. We know that for independent data, the parameter β_j^* $(j = 1, \ldots, J)$ is estimated by the mean of the outcome values which belong to the jth terminal node.

It is well-known that ignoring the correlation information of longitudinal studies leads to inefficient model estimation, reduced power for statistical testing, and incorrect variance estimators. Therefore, at each split $R_1, R_2, \ldots, R_J$, we estimate models with QIF described in chapter 3 rather than GLM. The advantage of QIF is that we do not need to estimate the correlation parameter at each split for each node, which is a major obstacle in computation (Segal, 1992; Zhang, 1998). Song (2009) developed both R package and SAS macro to perform parameter estimation for longitudinal data with the QIF approach. Using simulation results, Liang and Zeger (1986) presented substantial improvement in the efficiency of point estimates of β^* with the GEE method after the correlation structure was considered. Moreover, Qu et al. (2000) demonstrated the equivalence of QIF and GEE methods when the working correlation structure is correctly specified as well as the higher efficiency of the QIF approach over the GEE method when the working correlation structure is misspecified. Thus, when it comes to construct a tree-based model for longitudinal data, the efficiency of β^* estimation with the QIF approach is expected to exceed the efficiency of generalized linear estimation of β^* in the regular CART method.

4.1 Regression Trees

Another aspect of the CART procedure is to develop a proper criterion to select the best split. Just as CART for independent data, criteria for regression trees and classification trees must differ to produce the best results. Here we discuss the criterion and procedure

for estimating regression trees for longitudinal studies.

Inspired by Segal's (1992) generalization of the least squares split function to multivariate cases and QIF for longitudinal data, we define the objective function for the selection of the best split as $\sum_{i\in R_j}(y_i-\mu_i)^T V_i^{-1}(y_i-\mu_i)$, where $V_i^{-1}\approx A_i^{-\frac{1}{2}}R(\alpha)^{-1}A_i^{-\frac{1}{2}}\approx A_i^{-\frac{1}{2}}\left(\sum_{k=0}^m a_k M_k\right)A_i^{-\frac{1}{2}}$ by the idea of GEE and QIF methods. We call this new criterion

$$RSSQ=\sum_{i=1}^N(y_i-\mu_i^*)^T A_i^{-\frac{1}{2}}\left(\sum_{k=0}^m a_k M_k\right)A_i^{-\frac{1}{2}}(y_i-\mu_i^*). \tag{4.2}$$

where $\mu_i^*=h(U_i\beta^*)$ is the estimator of y_i obtained from a regression tree by the QIF approach, and $\sum_{i=0}^m a_i M_i$ is a class of matrices to model $R(\alpha)^{-1}$. The linear combination of matrices $M_0,\ldots,M_m$ and constants $a_0,\ldots,a_m$ can form a considerably rich class to approximate or even accommodate commonly used correlation structures, such as independence, exchangeable, and AR-1 correlation structures.

RSSQ is different from traditional node impurity function RSS for regression trees because the correlation structure of longitudinal data is taken into account. In addition, RSSQ is a measure of impurity for all current terminal nodes with all observations rather than a node impurity function for observations in only one node. By minimizing RSSQ, we try to find the best node to perform further split as well as the splitting variable and the cutoff. We expect that RSSQ as a criterion for classification procedure not only yields accurate classifiers but also provides insight and understanding into the predictive structure of the data. Another interpretation of RSSQ is that its first derivative with respect to the regression parameter β is equivalent to the score function (3.1)

$$\frac{d(RSSQ)}{d\beta}=-2\sum_{i=1}^N\dot{\mu}_i^{*T}A_i^{-\frac{1}{2}}(\sum_{k=0}^m a_k M_k)A_i^{-\frac{1}{2}}(y_i-\mu_i^*)=0 \tag{4.3}$$

Thus, the estimation of β by the QIF approach is consistent with the minimization of RSSQ, making it analog to RSS or likelihood functions for independent data. Given a data set, the values of $a_k(k=0,\ldots,m)$ in $R(\alpha)^{-1}\approx\sum_{k=0}^m a_k M_k$ are fixed but unknown. Thus,

we can use C_N^{-1} obtained from QIF to derive the initial value of $a_k, k = 0, \ldots, m$, denoted as $\hat{a}_k$ for RSSQ. Note that our simulation studies show that the final results actually are not sensitive to $\hat{a}_k$, as long as $\hat{a}_k$ is not too far way from a_k .

In the case of independent correlation structure, $R^{-1} = I_{n_i \times n_i}$, hence a_0 is 1. In the cases of exchangeable and AR-1 structures in QIF, a linear combination of two basis matrices is used to approximate $R(\alpha)^{-1} \approx a_0 M_0 + a_1 M_1$. If a_0 and a_1 are known, it follows by the generalized method of moments that the solution $\hat{\beta}$ of estimating function (3.1) is equivalent to compute

$$\hat{\beta} = \arg\min_{\beta} g_N^T \, B \, g_N \tag{4.4}$$

where B is a blocked matrix,

$$B = \left(\begin{array}{ccc:ccc} a_0^2 & & & a_0 \cdot a_1 & & \\ & \ddots & & & \ddots & \\ & & a_0^2 & & & a_0 \cdot a_1 \\ \hdashline a_0 \cdot a_1 & & & a_1^2 & & \\ & \ddots & & & \ddots & \\ & & a_0 \cdot a_1 & & & a_1^2 \end{array}\right)$$

It is a blocked diagonal matrix with 4 blocks and the diagonal elements within each block are the same.

Since $\hat{\beta} = \arg\min_\beta g_N^T \, C_N^{-1} \, g_N$ in QIF, then $C_N(\beta)^{-1} = (\frac{1}{N^2} \sum_{i=1}^N g_i(\beta) g_i^T(\beta))^{-1}$ is approximately proportional to the blocked matrix B. Therefore, we can reduce the dimension of C_N^{-1} by computing the traces of its four blocks accordingly or equivalently the mean of diagonal entries within each block. Then we get a 2×2 matrix A, which is approximately proportional to the following rank-1 matrix

$$\left(\begin{array}{cc} a_0^2 & a_0 \cdot a_1 \\ a_0 \cdot a_1 & a_1^2 \end{array}\right) = (a_0, a_1)^T (a_0, a_1). \tag{4.5}$$

As a_0 and a_1 are fixed constants conditional on the correlation parameter α and $a_0 \neq 0$, we can factor out a_0 to get

$$\frac{RSSQ}{a_0} = \sum_{i=1}^{N} (y_i - \mu_i^*)^T A_i^{-\frac{1}{2}} \left(M_0 + \frac{a_1}{a_0} M_1 \right) A_i^{-\frac{1}{2}} (y_i - \mu_i^*). \tag{4.6}$$

Therefore, minimizing RSSQ is equivalent to minimizing (4.6). As a result, we only need to estimate the ratio a_1/a_0. Then we perform a singular value decomposition (SVD) on the 2×2 symmetric matrix A, and let the first eigenvector from the SVD analysis be $v^1 = (v_0^1, v_1^1)$. It follows from (4.5) that a_1/a_0 could be approximated by v_1^1/v_0^1.

We learn from the numerical studies that the selection of the best split is not sensitive to the approximation of a_1/a_0, as long as v_1^1/v_0^1 is not too far way from a_1/a_0. Therefore, when deciding a new split, we use C_N^{-1} and v_1^1/v_0^1 computed from last split rather than calculating v_1^1/v_0^1 for all the possible splits. We obtain the initial value of C_N^{-1} and v_1^1/v_0^1 with the model and the estimates from a regular CART analysis. This not only substantially reduces our computational cost, but also stabilizes the partition procedure, since v_1^1/v_0^1 computed from certain split might be extremely large or small.

We summarized our algorithm for building regression trees using the QIF approach and the RSSQ criterion as follows

Algorithm building a regression tree:

1. Obtain the initial approximation of a_1/a_0 with the v_1^1/v_0^1 obtained from a regular CART analysis.

2. For each possible split, fit the model using the QIF approach with a prespecified correlation structure, and calculate the RSSQ with the current values of v_1^1/v_0^1.

3. Choose the split with the smallest RSSQ and construct two new sub-nodes accordingly. Update the values of v_1^1/v_0^1 under the current tree.

4. Repeat step 2-3 until either the pre-specified minimum node sample size or the pre-specified minimum of the maximum reduction in RSSQ is reached. A large tree T_0 will be obtained.

5. Following Breiman's(1984) idea, obtain a series of subtrees of T_0, $T(\lambda) \subset T_0$, by minimizing the cost complexity $R_\lambda(T) = R(T) + \lambda|T|$, where $|T|$ represents the number of terminal nodes of subtree T, $R(T) = \sum (y_i - \mu_i^*)^T A_i^{-\frac{1}{2}} \left(\sum_{k=0}^m a_k M_k\right) A_i^{-\frac{1}{2}} (y_i - \mu_i^*)$ is the mean square prediction error of validation data, and λ tunes the penalty on the tree size. Then conduct a five-fold cross-validation on the training data to select the best tuning parameter λ and the subtree $T(\lambda)$. Consequently, the original tree T_0 is pruned to a subtree $T(\lambda)$.

4.2 Classification Trees

In this section, we proposed a novel criterion to grow classification trees for longitudinal data. We restricted the discussion to binary responses here, one can, however, extend our approach directly to ordinal data using the idea of Zhang and Ye (2008).

To build a classification tree for multiple binary responses, Zhang (1998) introduced a measure of within-node purity

$$w(t) = -\frac{1}{n_t} \sum_{i \in \text{node t}} (y_i - \bar{y}(t))^T V^{-1} (y_i - \bar{y}(t)) \tag{4.7}$$

where n_t is the node size, V^{-1} is the inverse covariance matrix of y_i in the root node, and $\bar{y}(t)$ is the average of y_i within node t. A classifier is selected by maximizing the within-node purity $w(t)$ or equivalently minimizing $-w(t)$. In fact, $-w(t)$ is the least square function for observations within node t.

A common criterion to evaluate the efficiency and accuracy of a classification tree is the misclassification rate, namely the proportion of observations which are incorrectly classified

in the opposite class. Since prediction accuracy is our goal, we intend to minimize the misclassification rate when selecting the best splits for a classification tree. However, the misclassification rate alone is not sufficiently sensitive to grow a tree as some tentative splits may result in equal misclassification rates.

Inspired by Zhang's (1998) within-node purity function and also considering the misclassification rate, we modify RSSQ for classification trees which simultaneously minimizes the misclassification error and node impurity (heterogeneity). We defined RSSQ(M) to be

$$RSSQ(M) = \delta \cdot |M| + \sum_{i=1}^{N} D_i^T A_i^{-\frac{1}{2}} \left(\sum_{k=0}^{m} a_k M_k \right) A_i^{-\frac{1}{2}} D_i \tag{4.8}$$

where $|M|$ is the number of observations which are incorrectly classified in the opposite class, δ is a weight for $|M|$, $D_{it} = y_{it} - \mu_{it}^*$ if all observations are classified in the same class, otherwise

$$D_{it} = \begin{cases} \mathrm{y}_{it} - \mu_{it}^* & \text{if the observation } y_{it} \text{ is correctly classified,} \\ 0 & \text{if the observation } y_{it} \text{ is misclassified,} \end{cases} \quad (t = 1, ..., n_i)$$

and $D_i = (D_{i1}, \ldots, D_{in_i})^T$.

RSSQ(M), similar to RSSQ, is a measure of impurity for the whole data set rather than the node impurity function for observations within a node. $|M|$ accounts for the misclassification error. Since the scale of $|M|$ and the second term may be close, we can avoid the scale problem by multiplying $|M|$ by a big weight, δ, if the reduction of misclassification errors is desired. If all observations are classified into the same category, the second term, $\sum_{i=1}^{N} D_i^T A_i^{-\frac{1}{2}} \left(\sum_{k=0}^{m} a_k M_k \right) A_i^{-\frac{1}{2}} D_i$, is the least square error (LSE) for all observations; otherwise, the second term is the LSE of observations excluding the misclassified ones.

In the case when all observations are classified into the same class, $|M|$ is fixed, equal to the number of observations belonging to the less frequent class because those observations would be classified into the more frequent class. This happens frequently in practice. Taking

the Indonesian children's healthy study (ICHS) as an example, Sommer et al. (1984) reported only 444 accounts of respiratory infection out of 1200 total records. If a classification tree is fitted to ICHS data, the first few classifiers may not split records into different classes, but later classifiers may produce different classes and result in smaller misclassification rates. In this case, minimizing the LSE of all observations tends to make the estimators of correct-classified observations closer to their classes, meanwhile producing estimators of misclassified observations as close to 0.5 as possible. Then, the next split may be selected in the node with estimator close to 0.5.

In the case when observations are classified in different classes, we compute LSE of observations excluding the misclassified ones. One reason of excluding misclassified observations in the second term is that $|M|$ already represents the misclassification error. More importantly, minimizing the second term including misclassfied observations can mess up the procedure of choosing the best classifier because we hope to make the estimator of class-0 as close to 0 as possible and the closer the estimator of class-1 to 1 the better, but the misclassified observations tend to drag both the estimators of class-0 and class-1 to the midway, 0.5. Thus, excluding the misclassfied observations and computing LSE for only correct-classified observations meet our goal of selecting the best splits. The best classifiers are selected by minimizing RSSQ(M). Specifically, if two splits result in the same misclassification error $|M|$, then minimizing RSSQ(M) tends to select the one with the smaller least square error of the correct-classified observations. On the other hand, if two splits have similar LSE's of the correct-classified observations, then minimizing the RSSQ(M) can select a split producing a smaller misclassification error.

The algorithm of building a classification tree is almost the same as building a regression tree based on RSSQ, the only differences are to replace the splitting function RSSQ with RSSQ(M) as well as to utilize $R_\lambda(T) = \delta \cdot |M| + \sum_{i=1}^{N} D_i^T A_i^{-\frac{1}{2}} \left(\sum_{k=0}^{m} a_k M_k\right) A_i^{-\frac{1}{2}} D_i + \lambda|T|$ as the cost complexity for cross-validation to perform pruning.

CHAPTER 5

SIMULATION RESULTS

To compare the performance of our method with the regular CART for longitudinal data, we conducted simulations with simple non-linear scenarios. For each simulation, simulated data are randomly assigned into training set and validation set. Training data are used to fit the models which are expected to perform well for training data, therefore the performance of those models is tested on validation set.

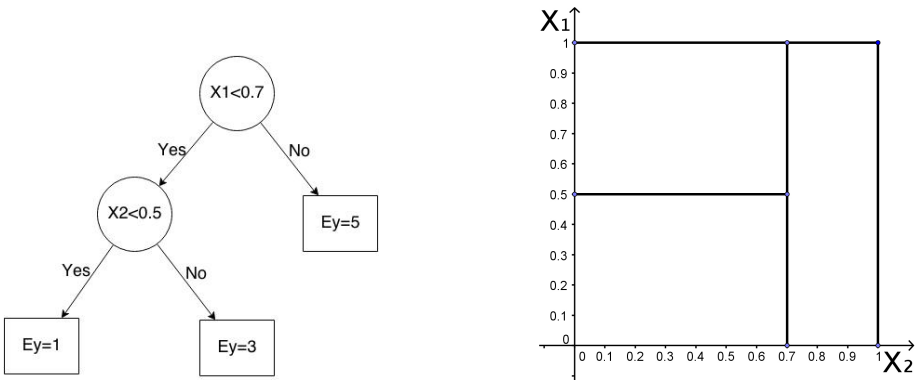
5.1 Regression Trees

For simplicity but without loss of generality, we generated data for each subject with equal length, $n_i = 10$ $(i = 1, \ldots, 20)$. The set-up of non-linear longitudinal data for regression trees is as follows:

$$y_i = I(X_1 < 0.7 \& X_2 < 0.5) + 3 \cdot I(X_1 < 0.7 \& X_2 \geq 0.5) + 5 \cdot I(X_1 \geq 0.7) + \varepsilon_i \qquad (5.1)$$

where X_1 is generated from a multinormal distribution with mean $(0.1, 0.2, \ldots, 1.0)$ and scaled identity matrix (scalar is 0.0001) as covariance matrix, X_2 is generated from Uniform(0, 1) distribution independently, and $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})^T$ has correlation structure with nuisance parameter α. MASS package in R (Ripley et al. 2002) can simulate multiple

response variable with specific covariance matrix. We used both AR-1 and exchangeable correlation structures in our scenario.



(a) True tree (b) True partition

Figure 5.1: The simulation scenario for regression trees

The mean square error (MSE) of prediction is used to assess the goodness of the predictive structure for a regression tree. The mean square error of prediction is estimated by averaging $\sum_{j=1}^{J}\sum_{X_{it}\in R_j}(y_{it}-\hat{y}_{R_j})^2$ over all samples, where y_{it} is the response in the validation set and $\hat{y}_{R_j}$ is the terminal node estimator obtained by the training set. Training sets contain 20 subjects with 10 longitudinal measures each, and validation sets consist of 10 subjects with 10 observations each. A regression tree is constructed on the training set and then the mean square error of prediction is computed on the validation set.

To evaluate the accuracy of classifiers, the mean square error (MSE) of splits is estimated by averaging the sum of the squared differences of estimated splits and the true splits over all samples. In our simulation scenario, average $(split(X_1)-0.7)^2+(split(X_2)-0.5)^2$ over all samples.

If the correlation structure is specified as independence, RSSQ is equal to RSS, indicating our method is equivalent to Breiman's CART with independence structure. The simulated relative efficiency (SRE) is defined as

$$SRE=\frac{\text{MSE with independence structure}}{\text{MSE with specified correlation structure}} \tag{5.2}$$

SRE's of splits and SRE's of prediction are calculated from 500 simulations respectively over a variety of working assumptions. Table 5.1 records SRE of splits and shows that, in general, our RSSQ method outperforms regular CART regarding to the accuracy of classifiers. When the true correlation α is large, $\alpha = 0.7$, the efficiency of the RSSQ method is much higher than CART, especially when the correlation structure is correctly specified. In particular, when the true correlation is AR-1 with autocorrelation $\alpha = 0.7$, $SRE = 1.828$ with correct working assumption; and when the true correlation has exchangeable structure with common correlation $\alpha = 0.7$, $SRE = 1.832$ with correct working assumption. However, there is only small differences between CART and the RSSQ method when the true correlation α is moderate, say 0.3. Particularly, in the case of simulations with exchangeable correlation structure with common correlation $\alpha = 0.3$, the RSSQ method is less efficient than CART when AR-1 working structure is assumed. It is because the estimation of the coefficient of basis matrix M_1 from C_N^{-1} tends to make $\hat{a}_1$ smaller than its true value, then the selection of the best classifier based on minimum RSSQ may slightly suffer less accuracy.

		Working R	
True R	α	AR-1	Exchangeable
AR-1	0.3	1.012	1.014
	0.7	1.828	1.071
Exchangeable	0.3	0.849	1.162
	0.7	1.641	1.832

Table 5.1: SRE of splits for regression trees

Table 5.2 records SRE's of prediction on validation sets over various working assumptions and demonstrates that the RSSQ method and CART are almost equivalent regarding to the goodness of predictive structure of regression trees. In general, exchangeable working assumption outperforms AR-1 working assumption. As mentioned before, it is due to the underestimation of a_1, the coefficient of basis matrix M_1, from C_N^{-1}. Therefore, an exchangeable correlation structure is more stable than an AR-1 correlation structure in RSSQ.

		Working R	
True R	α	AR-1	Exchangeable
AR-1	0.3	0.972	0.997
	0.7	0.984	1.000
Exchangeable	0.3	0.970	1.003
	0.7	1.000	1.004

Table 5.2: SRE of prediction for regression trees

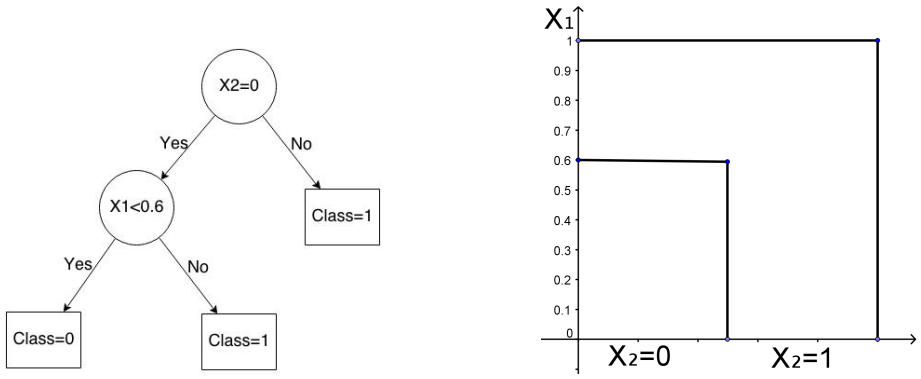
5.2 Classification Trees

For simplicity but without loss of generality, we generated data for each subject with equal length, $n_i = 15$ ($i = 1, \ldots, 20$). We used MultiOrd package in R (Amatya & Demirtas, 2006) to simulate binary responses with specific correlation structures, including exchangeable and AR-1. Due to the limitations of simulation package for binary responses, an appropriate value of nuisance parameter for correlation structure should be not larger than 0.5 and the difference of the expectation of responses cannot be too large. The set-up of non-linear longitudinal data for classification trees is as follows:

$$Ey_i = 0.35 \cdot I(X_1 < 0.6 \& X_2 = 0) + 0.65 \cdot I(X_1 \geq 0.6 \& X_2 = 0) + 0.65 \cdot I(X_2 = 1) \quad (5.3)$$

where X_1 is generated from a multinormal distribution with mean $(0.1, \ldots, 1.0)_{1\times 15}$ and scaled identity matrix (scalar is 0.0001) as covariance matrix, X_2 is a categorical variable in which each subject is randomly assigned either 0 or 1, y_i are simulated binary responses having a specific correlation structure with nuisance parameter α.

The misclassification rate of prediction is used to assess the goodness of the predictive structure for a classification tree. In each sample, the training set consists of 20 subjects with 15 longitudinal measures each, and the validation set contains 10 subjects with 15 observations each. A classification tree is constructed on the training set and then the model's performance is assessed by the validation set. The overall misclassification rate of prediction is estimated by averaging the misclassification rates of validation sets over



(a) True tree (b) True partition

Figure 5.2: The simulation scenario for classification trees

all samples. Since the true tree in simulation scenario has three terminal nodes, we made comparison of our RSSQ(M) approach and regular CART with respect to the mean square error of misclassification in trees with three terminal nodes for simplicity.

The simulated relative efficiency (SRE) for classification is defined as

$$SRE = \frac{\text{mean of misclassification rates by regular CART}}{\text{mean of misclassification rates by RSSQ(M)}} \tag{5.4}$$

SRE's of misclassification rate are calculated from 500 simulations over a variety of working assumptions. Table 5.3 shows that our approach slightly outperforms regular CART regarding to SRE's of predictive misclassification when the true correlation structure is AR-1. Specifically, the prediction efficiency of trees under independence and AR-1 working assumptions is a little better than exchangeable assumption. However, if the true correlation structure is exchangeable, regular CART has slightly higher efficiency than our approach by RSSQ(M). If making comparison among three working assumptions, exchangeable which is the true correlation structure is more efficient than the other two working assumptions.

		Working R		
True R	α	independence	AR-1	Exchangeable
AR-1	0.5	1.015	1.010	1.002
Exchangeable	0.5	0.974	0.973	0.985

Table 5.3: SRE of predictive misclassification rate

CHAPTER 6

REAL-DATA EXAMPLES

6.1 Regression Trees for Ozone Data

We applied a regression tree with RSSQ criterion on Ozone Data (Weiss, 2005). In late July 1987 at 20 sites in and around Los Angeles, California, USA, the ozone was measured over a three-day period. Twelve recordings were taken hourly from 07:00 am to 18:00 pm each day, providing us $20 \times 20 \times 3$ ozone records. Measurement units are in parts per hundred million. The original data set gives the four-letter abbreviation for the sites, the full names of the sites, and the longitude, latitude, and altitude of each site. The longitude, latitude and altitude of a site was simplified into a categorical variable, valley. Valley is categorical variable to indicate whether the site is in the Simi or San Fernando Valleys (SF) or San Gabriel Valleys (SG). The remaining sites are adjacent to the ocean or otherwise do not have mountain ranges between them and the ocean. We treat the data as having $60 = 20 \times 3$ subjects with 12 longitudinal measures each. There is no missing value in this data set. 30 subjects were randomly chosen to training data set, and the remaining 30 subjects were in the validation set.

According to *figure 6.1*, ozone and variable hour are highly correlated, but the relation between ozone and hour is nonlinear because ozone rises in the daytime and returns to a

variable	variable name	Description
y	ozone	Numerical; Ozone measurements ranging from 0.0 to 28.6
x_1	hour	Numerical; Time from 07:00 am to 18:00 pm
x_2	day	Categorical; The day on which ozone measured
x_3	valley	Categorical; "SF" means site in the Simi or San Fernando Valleys, "SG" means site in San Gabriel Valleys, "NO" means site in neither the Simi/San Fernando Valleys nor San Gabriel Valleys.
ID	ID	Each subjects identification number.

Table 6.1: Ozone data description

baseline value each night. Furthermore, ozone slightly increases by day. In addition, San Gabriel Valleys (SG) tend to have higher ozone than Simi or San Fernando Valleys (SF) and remaining sites. To analyze the pattern of ozone, a classification tree is a reasonable method.

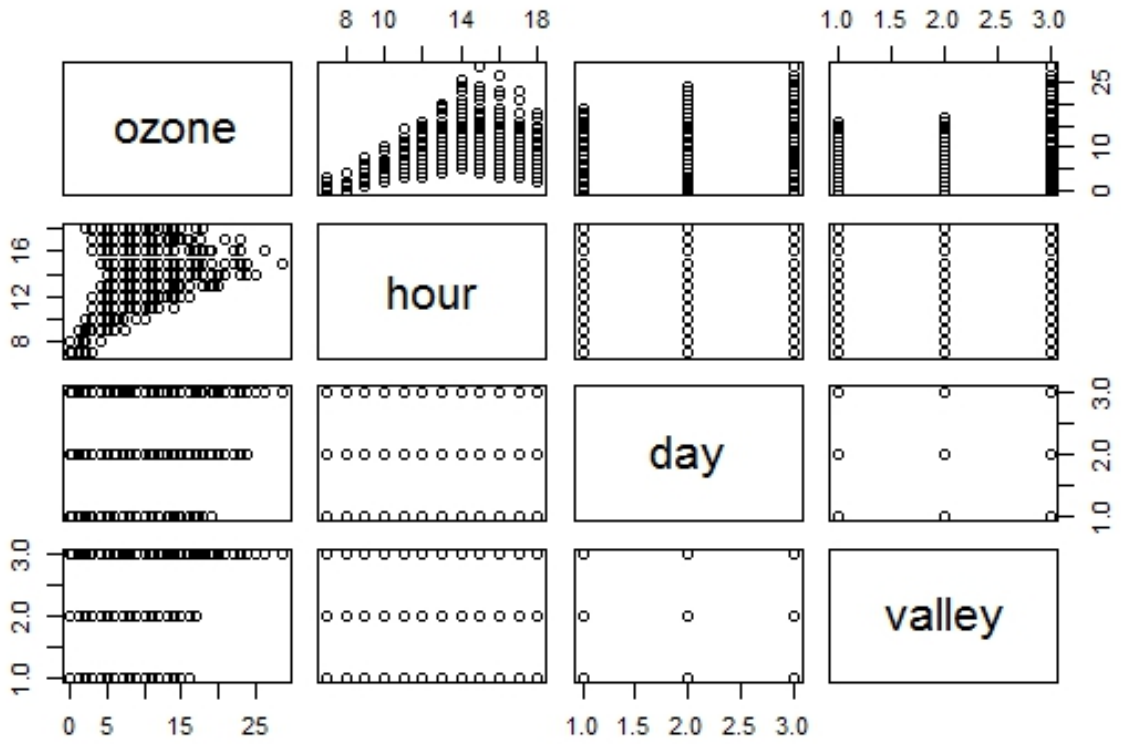


Figure 6.1: Plots of pairs of response and predictor variables for ozone data

Moreover, the measures of ozone are changing over time and are highly correlated within

each subject. Since the repeated measurements are taken at equally-spaced time points, the correlation matrix of responses within subjects can be computed.

$$\begin{pmatrix}
1.00 & 0.28 & -0.02 & -0.16 & -0.23 & -0.26 & -0.26 & -0.26 & -0.30 & -0.29 & -0.25 & -0.23 \\
0.28 & 1.00 & 0.53 & 0.33 & 0.03 & -0.10 & -0.10 & -0.19 & -0.20 & -0.13 & -0.07 & 0.03 \\
-0.02 & 0.53 & 1.00 & 0.75 & 0.49 & 0.32 & 0.24 & 0.21 & 0.29 & 0.37 & 0.38 & 0.47 \\
-0.16 & 0.33 & 0.75 & 1.00 & 0.79 & 0.59 & 0.39 & 0.24 & 0.33 & 0.36 & 0.37 & 0.47 \\
-0.23 & 0.03 & 0.49 & 0.79 & 1.00 & 0.86 & 0.60 & 0.44 & 0.50 & 0.49 & 0.51 & 0.54 \\
-0.26 & -0.10 & 0.32 & 0.59 & 0.86 & 1.00 & 0.83 & 0.69 & 0.68 & 0.64 & 0.59 & 0.51 \\
-0.26 & -0.10 & 0.24 & 0.39 & 0.60 & 0.83 & 1.00 & 0.89 & 0.81 & 0.73 & 0.62 & 0.50 \\
-0.26 & -0.19 & 0.21 & 0.24 & 0.44 & 0.69 & 0.89 & 1.00 & 0.93 & 0.85 & 0.77 & 0.63 \\
-0.30 & -0.20 & 0.29 & 0.33 & 0.50 & 0.68 & 0.81 & 0.93 & 1.00 & 0.93 & 0.84 & 0.75 \\
-0.29 & -0.13 & 0.37 & 0.36 & 0.49 & 0.64 & 0.73 & 0.85 & 0.93 & 1.00 & 0.91 & 0.81 \\
-0.25 & -0.07 & 0.38 & 0.37 & 0.51 & 0.59 & 0.62 & 0.77 & 0.84 & 0.91 & 1.00 & 0.93 \\
-0.23 & 0.03 & 0.47 & 0.47 & 0.54 & 0.51 & 0.50 & 0.63 & 0.75 & 0.81 & 0.93 & 1.00
\end{pmatrix}$$

This correlation matrix can be viewed as a descriptive indication of the relation between responses, and it has an approximate first-order autoregressive correlation structure. Additionally, we performed a Shapiro-Wilk test to verify that the response variable, ozone, satisfies the assumption of normality because p-value $< 2.2 \times 10^{-16}$. Thus, we can assume AR-1 correlation structure and Gaussian family to conduct estimation on Ozone Data using QIF. Therefore, a regression tree with RSSQ can be applied to this Ozone Data set in which the correlation structure is taken into consideration.

Figure 6.2 is the regular regression tree constructed by tree package in R (Ripley, 2014). A regular regression tree treats highly correlated outcomes as independent observations to fit a piecewise-constant tree model. According to *figure 6.1*, it is obvious that ozone changes by time everyday, rising in the daytime and falling to a baseline value at night. The change of ozone by time may be due to the temperature change in a day. In the morning and at night (before 10:30 am and after 17:30 pm), the temperature is relatively low resulting in a low ozone level, while in the daytime (between 10:30am to 17:30 pm) ozone level rises along with temperature increase. Thus, it is reasonable that most cutoffs are split at time (nodes 1, 2, 6, 7, 11, and 14). In addition, node 3 splits on valley, indicating sites in San Gabriel

Valleys have ozone levels quite different from the other sites after 10:30 am. Specifically, the longitudes of sites in San Gabrial Valleys (SG) are much higher than the other two kinds of sites; therefore, high longitude or mountain ranges between the sites and the ocean may lead to high ozone. Node 8 splits at valley on "SF" and "No", suggesting that sites in those two kinds of valleys have different ozone levels from 10:30 am to 17:30 pm, the time period of peak ozone level. Among all terminal nodes, node 12 has a high incidence of effects, containing 108 observations out of a total of 360 records.

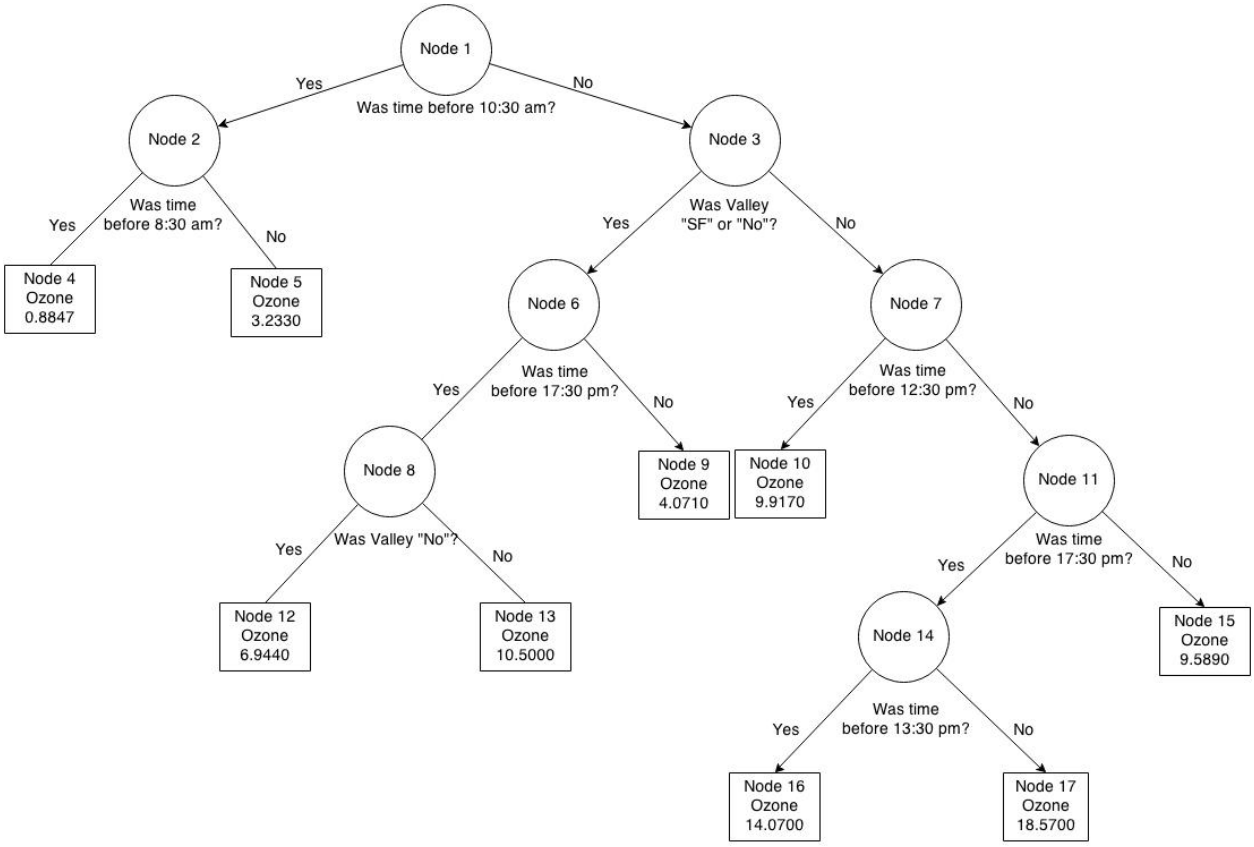


Figure 6.2: The regression tree for Ozone Data by regular CART

Figure 6.3 presents the regression tree obtained by RSSQ criterion which takes the strong correlation among responses into consideration. It shows that the RSSQ criterion generates a regression tree which is remarkably similar to the one by the regular regression tree. They tell similar stories about ozone level. Both trees have exactly the same classifiers at nodes 1, 2, 3, and 7, implying time and valley are influential variables absolutely. Neither regular

CART nor our RSSQ method has a classifier on day, indicating variable-day is not significant. Different from *Figure 6.2*, node 4 is split at valley and node 5 is further split at time. Node 12 has the highest incidence of effects, involving 126 out of a total of 360 observations.

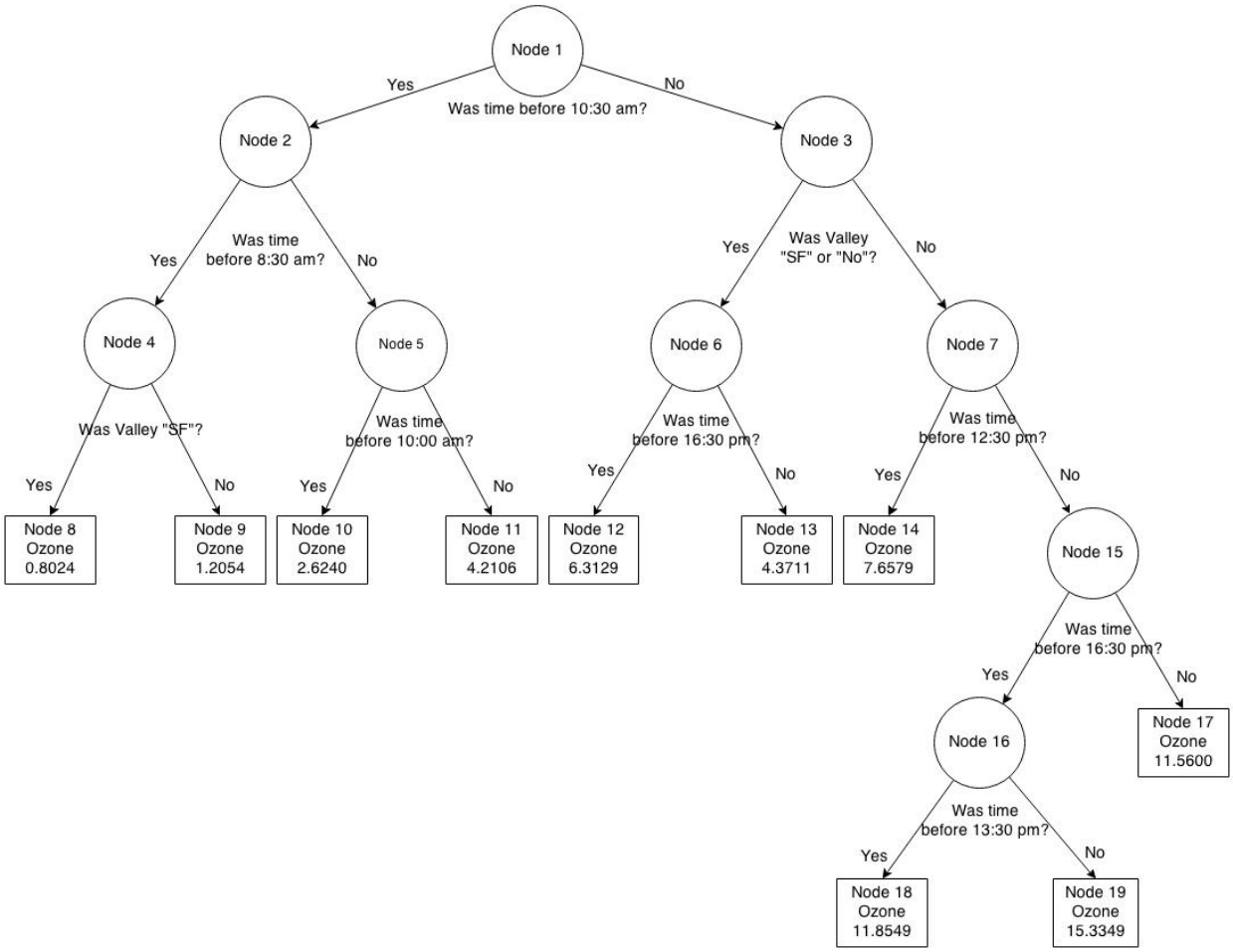


Figure 6.3: The regression tree for Ozone Data by the RSSQ criterion

By regular CART, the mean square errors (MSE) of the training data set and the validation set are 6.354 and 9.697 respectively. By CART with RSSQ, which is our method, MSE's of the training data set and the validation set are 9.136 and 12.814. In terms of MSE, regular CART outperforms our approach with RSSQ criterion. However, RSSQ values of the training data and the validation data by regular CART are 1279.80 and 2080.56; RSSQ values of the training data and the validation data by our method are 778.74 and 1792.08 correspondingly, which are significantly smaller than the ones by regular CART. Considering

the correlation of responses, our method is superior to regular CART regarding to RSSQ values.

6.2 Classification Trees for Indonesia Children's Health Study

We fitted classification trees to Indonesian children's health study data (ICHS, Diggle et al., 2002). This data set is a sub-sample of 250 preschoolers from Indonesian children's health study (Sommer et al., 1984). Preschool children in Indonesia were examined quarterly for up to 6 consecutive visits. Those preschoolers' presence of respiratory infection (RI), gender, presence of the ocular disease Xerophthalmia and age were recorded for each visit. There is no missing value in this data set. A primary question of concern is whether there is an increase in risk of respiratory infection for children who are vitamin A deficient which were indicated by the presence of the ocular disease Xerophthalmia. The chronic change in the prevalence of respiratory infection is also of interest.

variable	variable name	Description
y	presence of respiratory infection	Categorical; not present=0, present=1.
x_1	time	Numerical; time in months.
x_2	gender	Categorical; male=0, female=1.
x_3	Vitamin A	Categorical; not deficient=0, deficient=1.
x_4	age	Numerical; age in years.
ID	ID	Each subjects identification number.

Table 6.2: ICHS data description

An advantage of classification trees is that the variable selection is done automatically, since a classification tree will select the influential variables as splitting variables. Thus, the first problem whether the presence of respiratory infection is associated with vitamin A deficiency can be addressed by a classification tree. In addition, if the variable-time is selected as a splitting variable, then the chronic change in the prevalence of respiratory infection can also be presented. On the contrary, if the variable-time is not a splitting variable, then there

is no obvious chronic change. We randomly assigned 150 children into the training data set and the remaining 100 children into the validation set.

All subjects have 6 consecutive observations, so we can compute the correlation matrix of the response variable. It turns out to be an exchangeable correlation structure. We can try exchangeable correlation structure in RSSQ(M) to construct a classification tree.

$$\begin{pmatrix} 1.00 & 0.58 & 0.45 & 0.46 & 0.52 & 0.45 \\ 0.58 & 1.00 & 0.58 & 0.61 & 0.55 & 0.58 \\ 0.45 & 0.58 & 1.00 & 0.57 & 0.44 & 0.49 \\ 0.46 & 0.61 & 0.57 & 1.00 & 0.47 & 0.54 \\ 0.52 & 0.55 & 0.44 & 0.47 & 1.00 & 0.53 \\ 0.45 & 0.58 & 0.49 & 0.54 & 0.53 & 1.00 \end{pmatrix}$$

In the whole ICHS data set, there are 444 out of 1500 records showing presence of RI. Thus, the probability of RI is relatively low. If we conduct a regular classification tree on the training data set, R package tree (Ripley, 2014) gives us an unpruned tree with splits on gender and age (*Figure 6.4*). Boys' probability of RI is 38.72%, much higher than girls', 24.00%. Within girls, younger ones (less than 6.5 years old) tend to have higher risk of RI, 26.13%, while older ones, with probability 4.17%, seldom suffer from RI. Although the probabilities of presence of RI in the terminal nodes differ, all observations are classified as no presence of RI since the probabilities are smaller than the classification threshold 0.5. If we prune the tree by cross-validation, only a single node remains and all observations are classified into one category, no presence of RI. Thus, the misclassification rate is 444 out of 1500. Although, the type I error (the proportion of preschoolers who are incorrectly classified as presence of RI) of this tree is zero, the type II error (the proportion of observations who suffer from RI but are incorrectly classified as no presence of RI) is high, 29.6%. If detecting the presence of RI is the goal, a regular classification tree cannot meet the goal. An alternative solution to decrease type II error is to lessen the threshold for splitting.

We also constructed a classification tree on ICHS training data based on the RSSQ(M) criterion. After tree pruning procedure, we obtain a 5-leaf classification tree shown in *figure*

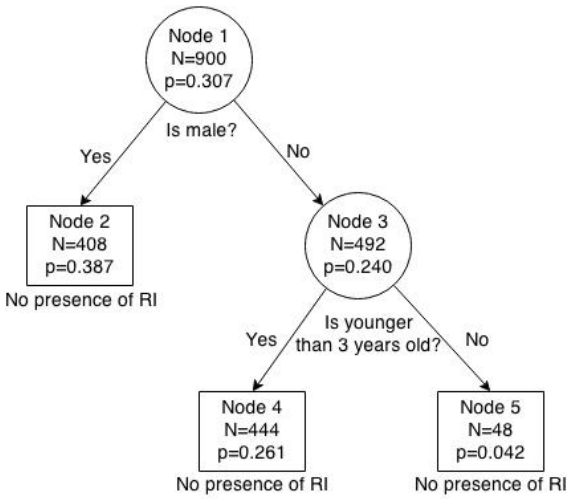


Figure 6.4: The classification tree for ICHS data by regular CART

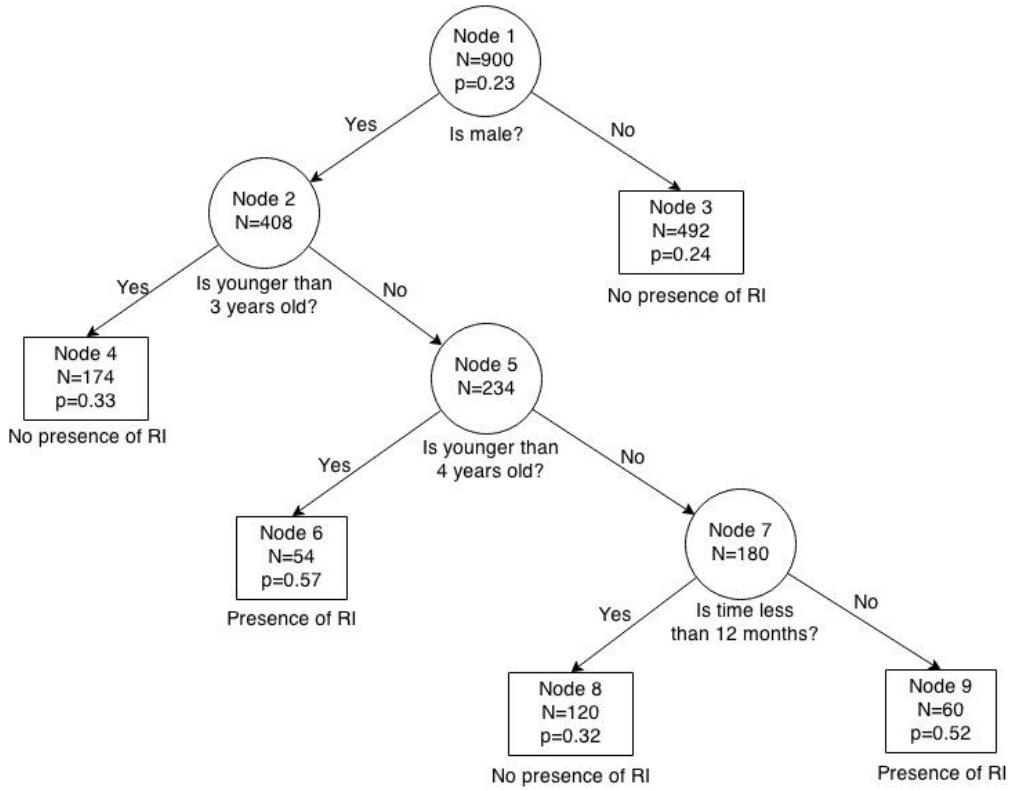


Figure 6.5: The classification tree for ICHS data by the RSSQ(M) criterion

6.5 which tells a different story from the one in *figure 6.4*. The first node is also slpit at gender, indicating boys and girls have significantly different probabilities of presence of RI. Both nodes 2 and 5 are split at age, implying the presence of RI may change by

age. Specifically, 3-year-old boys have higher risk of RI. Moreover, older boys tend to have increasing risk of RI after participating this study 12 months. Furthermore, the classification tree based on the RSSQ(M) criterion has a misclassification rate 0.295, slightly smaller than regular CART. Although the type I error increased a little bit, it is acceptable considering the significant decrease in the type II error. The smaller the type II error is, the higher chance to detect the presence of RI.

The classification matrix gives a sense of the classification accuracy and what type of misclassification is more frequent. Based on the trees built on the training data, we calculated the type I, type II and overall misclassification rate on the validation data set. Even if the RSSQ(M) approach leads to high type I error and misclassification rate comparing with regular CART, the reduction in the type II error is significant. Thus, our approach remains efficient if detecting the presence of RI is desired.

	Training set		Validation set	
CART	Regular	RSSQ(M)	Regular	RSSQ(M)
Type I error	0	0.058	0	0.103
Type II error	0.307	0.238	0.28	0.237
Misclassification rate	0.307	0.295	0.28	0.34

Table 6.3: Error rates of two approaches

All in all, in both approaches of building classification trees on ICHS data, no splits on vitamin A suggests that the presence of RI is not associated with vitamin A deficiency. Regarding to the chronic change in presence of RI, girls younger than 6.5 years old tend to suffer higher risk of RI, and boys aging at 3 have higher probability of presence of RI.

CHAPTER 7

DISCUSSION

In this project, we proposed to extend the very popular CART approaches to longitudinal data with the QIF approach. Unlike other tree-based methods based on estimating equations, the proposed approach does not require the estimation of the correlation parameter within each node, therefore greatly reduce the computational cost and is robust against the estimation of the correlation parameters. For regression trees, we proposed a new criterion that is analog to the residual sum of squares or likelihood functions for independent data. For classification trees, we developed a novel criterion which takes into consideration both the misclassification rate and the within-node purity. Numerical studies show that the proposed procedure consistently outperforms the traditional CART.

Existing literature tends to study longitudinal data and data with multiple responses together. We did not provide a detailed discussion for multiple response data here, however, the proposed methods could be directly applied to such data sets. Moreover, we believe the proposed methods could be extended to longitudinal ordinal responses, although we expect the computational cost would be substantially higher. It might be desirable to develop some algorithms that could screen out the bad splits for faster computation. In addition, there might be faster ways to optimize QIF as we currently just employ a simple Newton-Raphson algorithm. A proper gradient-based technique may help to reduce the computational time.

In this paper, we discuss the classification tree for only the binary response. In general, classification problems with multi-cluster responses are highly challenging and require further research. However, we can extend the research here directly with the idea of Zhang and Ye (2008). We believe the computational advantage of the proposed procedure is even more significant in the case of ordinal responses, due to the increasing number of parameters.

We end the discussion with another potential advantage of the proposed approach. With the existing approaches, it is often difficult to perform hypothesis testing, which is an essential part of statistical inference in many applications. This is due to the fact that observations inside and outside a node are not independent. With our approach, we incorporate all the observations under one linear model at each split instead of fitting a model for each node. This allows us to carry out the hypothesis testing procedure just as testing whether a certain variable is significant in a regular generalized linear model. Qu et al. (2000) provided a detailed discussion on how to perform chi-square test with QIF, and it is can be directly applied to our proposed CART for longitudinal studies.

BIBLIOGRAPHY

[1] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA.

[2] Cleveland,W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829836.

[3] Death, G. (2002). Multivariate regression trees: A new technique for modeling species environment relationships. *Ecology*, 83, 11051117.

[4] Diggle, P. J., Heagerty, P. J., Liang, K. Y. and Zeger, S. L. (2002). *Analysis of Longitudinal Data, 2nd ed*. Oxford Statistical Science Series 25. Oxford Univ. Press, Oxford.

[5] Hansen, L. (1982), Large sample properties of generalized method of moments estimators. *Econometrica*, 50, 1029-1054.

[6] Hsiao, W. C. and Shih, Y. S. (2007). Splitting variable selection for multivariate regression trees. *Statistic & Probability Letter*, 77, 265271.

[7] James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. New York, NY: Springer.

[8] Lee, S. K. (2005). On generalized multivariate decision tree by using GEE. *Computational Statistics & Data Analysis*. 49, 11051119.

[9] Lee, S. K. (2006). On classification and regression trees for multiple responses and its application. *Journal of Classification*, 23, 123-141.

[10] Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 12-22.

[11] Loh, W. Y. and Shih, Y. S. (1997). Split selection methods for classification trees. *Statistica Sinica*, 7, 815840.

[12] Loh, W. Y. (2009). Improving the precision of classification trees. *The Analysis of Applied Statistics*, 3, 17101737.

[13] Loh, W. Y. and Zheng, W. (2013). Regression trees for longitudinal and multiresponse data. *The Analysis of Applied Statistics*, 7, 495-522.

[14] Qu, A., Lindsay, B. and Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika*, 87, 823-836.

[15] Segal, M. R. (1992). Tree structured methods for longitudinal data. *Journal of the American Statistical Association*, 87, 407418.

[16] Simonoff, J. S. (2003). *Analyzing Categorical Data*. Berlin: Springer.

[17] Wang, P., Tsai, G. F. and Qu, A. (2012). Conditional inference functions for mixed-effects models with unspecified random-effects distribution. *Journal of the American Statistical Association*, 107, 725-736.

[18] Weiss, R. E. (2005). *Modeling longitudinal data*. New York, NY: Springer.

[19] Yu, Y. and Lambert, D. (1999). Fitting trees to functional data, with an application to time-of-day patterns. *Journal of Computational and Graphical Statistics*, 8, 749762.

[20] Zeger, S. L., Liang, K. Y., and Albert, P. S. (1988). Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics*, 44, 1049-1960.

[21] Zhang, H. P. (1997). Multivariate Adaptive Splines for Analysis of Longitudinal Data. *Journal of Computational and Graphical Statistics*, 6, 74-91.

[22] Zhang, H. (1998). Classification trees for multiple binary responses. *Journal of the American Statistical Association*, 93, 180193.

[23] Zhang, H. and Ye, Y. (2008). A tree-based method for modeling a multivariate ordinal response. *Stat. Interface*, 1, 169178.