Published in final edited form as:

Comput Stat Data Anal. 2015 February 1; 82: 152–163. doi:10.1016/j.csda.2014.08.013.

Ordinal Logic Regression: A classifier for discovering combinations of binary markers for ordinal outcomes

Bethany J. Wolf^{1,*}, Elizabeth H. Slate², and Elizabeth G. Hill¹

¹ Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC 29464

² Department of Statistics, Florida State University, Tallahassee, FL 32306

Abstract

In medicine, it is often useful to stratify patients according to disease risk, severity, or response to therapy. Since many diseases arise from complex gene-gene and gene-environment interactions, patient strata may be defined by combinations of genetic and environmental factors. Traditional statistical methods require specifying interactions *a priori* making it difficult to identify high order interactions. Alternatively, machine learning methods can model complex interactions, however these models are often difficult to interpret in a clinical setting. Logic regression (LR) enables modeling a binary outcome using logical combinations of binary predictors yielding easily interpretable models. However LR, as currently available, cannot model ordinal responses. This paper extends LR to model an ordinal response and the resulting method is called Ordinal Logic Regression (OLR). Several simulations comparing OLR and Classification and Regression Trees (CART) demonstrate that OLR is superior to CART for identifying variable interactions associated with an ordinal response. OLR is applied to data from a study to determine associations between genetic and health factors with severity of adult periodontitis.

Keywords

Logic regression; ordinal response; decision trees

1. Introduction

There is growing evidence that many common diseases arise from multiple factors including gene-gene and gene-environment interactions [4, 9, 19, 20, 24]. These factors can affect an individual's risk of developing disease, disease severity, or response to treatment [13, 16, 29]. Therefore, understanding a patient's risk profile can be beneficial to determining the

© 2014 Elsevier B.V. All rights reserved

*Corresponding author: MUSC, 135 Cannon Street, Charleston, SC 29464 phone: 843-876-1940 wolfb@msc.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Conflict of Interest: None declared.

best course action for disease prevention or treatment, but may require modeling complex interactions among genetic and environmental factors [3, 9, 20, 21, 22, 28].

Identifying high order interactions such as those that might describe disease status can be difficult using traditional statistical methods. For example, the cumulative logit model is a popular traditional statistical method for modeling ordinal outcomes [1, 2], however, interactions must be selected *a priori* and sufficient data must be available in order to develop a model containing interactions and all associated main effects. Also, as the number of predictors increases, the space of possible predictor interactions becomes prohibitively large limiting the effectiveness of traditional statistical methods.

Nonparametric tree-based methods are easily interpretable and have flexibility to identify relationships among predictor variables [6]. Classification and Regression Trees (CART) [8] is one such method capable of classifying ordinal outcomes. CART offers an additional advantage by allowing interactions to occur over a subset of the support space rather than across the entire support as is necessary in regression models. That is to say, branches in a CART model (tree) represent distinct variable interactions predictive for unique data subsets. When predictors are binary, however, this structure of an interaction is limited because a predictor may be used at most once in a branch.

A CART model is constructed by recursive partitioning of the response into increasingly homogeneous subsets defined by splits (i.e. dichotomizations) of predictor variables [15]. A common approach for identifying optimal splits uses the Gini impurity index (available in the *rpart* package in R), a measure that associates the same cost with all misclassifications [12]. Alternatively, a generalized Gini index (available in the *rpartScore* package in R [12]) allows the misclassification cost to increase as the distance between the true and predicted response category increases [5, 8, 12, 25]. Therefore, while the Gini impurity index is appropriate as a splitting criterion for multinomial response data, the generalized Gini index is more suitable for ordinal data.

Logic regression (LR) [26] is an alternative tree-based method that can be used to classify a binary response using Boolean combinations (“and”= $\wedge$, “or”= $\vee$, and “not”= $\neg$) of binary predictors. The use of $\vee$ in these relationships allows greater flexibility in modeling a response than is available in CART models. One major limitation of LR for classification is that logic trees are only able to predict binary outcomes.

In Section 2, we present an adaption of LR for prediction of ordinal responses that we refer to as Ordinal Logic Regression (OLR). In Section 3 we present the results of several simulation studies comparing the ability OLR and CART (using nominal and ordinal splitting criteria) to identify predictor interactions associated with an ordinal response. We then illustrate OLR by exploring associations among genetic and health factors with severity of adult periodontitis among African Americans with diabetes in Section 4. We conclude with additional discussion in Section 5.

2. Definitions and Notation

Let $\mathbf{W}_i = (y_i, \mathbf{x}_i)$ be the observed data on subject i , $i = 1, 2, \dots, n$, with y_i the ordinal response taking values $1, 2, \dots, K$ in increasing order, and $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ a row vector of p binary predictors. The p predictors $\mathbf{x}_i$ are also called the features associated with observation i and the set of the 2^p possible values of $\mathbf{x}_i$ is called the feature space. We use $\mathbf{W} = (\mathbf{W}_i) = ((y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n)) = (\mathbf{y}, \mathbf{x})$ for the collection of data on all n subjects.

2.1. Classification and Regression Trees

A CART model recursively partitions the observed data $\mathbf{W} = (\mathbf{y}, \mathbf{x})$ into subsets that are increasingly homogeneous in values of $\mathbf{y}$ based on values of the predictors $\mathbf{x}$. The splitting process stops when a pre-specified stopping criterion, usually based on a measure of fit quality, is met [8, 15]. Once the stopping criterion is met, the tree may be pruned (i.e. some nodes deleted) to prevent over fitting.

The most common method for determining the best split at a node in a CART tree is to maximize the reduction in impurity as measured by the Gini impurity index [8]. The Gini index at node t is given by

$$G(t) = \sum_{k=1}^K \sum_{j \neq k}^K p(k|t) p(j|t), \quad (2.1)$$

where $p(k|t)$ is the proportion of observations with response k in node t . A generalized Gini index [5, 8, 12] is defined as

$$GG(t) = \sum_{k=1}^K \sum_{j \neq k}^K C(j|k) p(j|t) p(k|t), \quad (2.2)$$

where $C(j|k)$ is the cost associated with misclassifying an observed response k as j . The Gini Index is the generalized Gini index with $C(j|k) \equiv 1$, which makes no distinction among possible misClassifications. An ordinal response, however, suggests that the cost of misClassification should increase with $|j - k|$. Although there are alternative ways to incorporate category specific misClassification costs [8, 12, 25], we adopt the generalized Gini index with either the absolute cost function, $C(j|k) = |j - k|$, or the quadratic cost function, $C(j|k) = (j - k)^2$, as implemented in *rpartScore* [12].

A fitted CART tree contains M terminal nodes, each corresponding to a rectangular region R_m in the feature space ($m = 1, 2, \dots, M$). The branches in a CART tree are Boolean combinations of predictors for Classification of $\mathbf{y}$ that can be thought of as interaction terms predictive of the response. The interaction defined by branch m is associated with the response category predicted at the m^{th} terminal node. We adopt the notation $\text{CART}(\mathbf{W}, K) = T$ to represent CART tree T fit to data $\mathbf{W}$ with K -category response variable y .

2.2. Logic Regression

Logic regression can be used when the response is binary, in which case the $K = 2$ categories are usually denoted as 0-1. LR constructs a tree T describing Boolean combinations of predictors for Classification of $\mathbf{y}$. All logic expressions, T , generated by LR can be expressed in disjunctive normal form (DNF), that is a series of *prime implicants* (PIs) joined by $\vee$ operators [11]. PIs capture predictor interactions as $\wedge$ combinations that cannot be further reduced [27]. For example $T = (X_3 \wedge X_4 \wedge X_5) \vee (X_3 \wedge X_5 \wedge X_{11})$, with PIs $X_3 \wedge X_4 \wedge X_5$ and $X_3 \wedge X_5 \wedge X_{11}$, is the DNF of the expression $T = (X_4 \vee X_{11}) \wedge (X_3 \wedge X_5)$. If the predictors for an observation generate a 1 for at least one PI in an LR model, the predicted binary response is 1.

2.3. Ordinal Logic Regression

An Ordinal Logic Regression (OLR) model is constructed using $K - 1$ cumulative response category indicators, defined as $I_\ell(y) = 1$ indicating $y \geq \ell$ and $I_\ell(y) = 0$ for $y < \ell$, $\ell = K, K - 1, \dots, 3, 2$. Category 1 is the reference category. OLR sequentially fits a logic regression tree for response indicator I_ℓ , $\ell = K, K - 1, \dots, 2$, reweighting the data for each tree to remove the influence of response observations already correctly classified in higher categories. The OLR algorithm proceeds as follows:

1. Initialize the weights as $w_i = 1, i = 1, 2, \dots, n$.
2. For $\ell = K, K - 1, \dots, 3, 2$
 - a. Create $I_\ell(y_i) = I_\ell(y_i) - \emptyset, i = 1, 2, \dots, n$.
 - b. Fit a logic regression model to $\{I_\ell(y_i)\}$ with weights $\{w_i\}$, obtaining T_ℓ

$$T_\ell = \underset{T}{\operatorname{argmin}} \sum_{i=1}^n I[T(\mathbf{x}_i) \neq I_\ell(y_i)] w_i$$

where $T(\mathbf{x})$ is a logic tree.

- c. Update the weights as $w_i \leftarrow w_i [1 - T_\ell(\mathbf{x}_i) I_\ell(y_i)]$.
3. This yields a collection of logic regression trees $T = \{T_\ell, \ell = K, K - 1, \dots, 2\}$.

The final OLR model therefore contains $K - 1$ trees and is denoted as $\text{OLR}(\mathbf{W}, K) = T$.

The prediction algorithm successively evaluates the predictions from individual logic trees in T starting with tree T_K . Given new predictors $\mathbf{x}^*$, let $S = \{\ell : T_\ell(\mathbf{x}^*) = 1\}$ be the collection of indices for the subset of trees in T that evaluate to 1. The OLR prediction is then given by

$$T(\mathbf{x}^*) = \max(1 \cup S). \quad (2.3)$$

2.4. Prime Implicant Matching

It is valuable to identify PIs that explain the response and, more specifically, that are associated with each response category. For a CART model, $\text{CART}(\mathbf{W}, K) = T$, let F_k denote

the set of PIs associated with the branches that predict the response category k , $k = 1, 2, \dots, K$. If a PI Q is an element of F_k for the CART tree T , we say that Q is an “exact match” for T associated with response category k . With a small abuse of notation, we will refer to PIs associated with each response category as F_k for both CART ($k = 1, 2, \dots, K$) and OLR ($k = 2, 3, \dots, K$). Thus for the ordinal logic regression $OLR(\mathbf{W}, K) = T = \{T_2, T_3, \dots, T_K\}$, let F_k be the PIs in T_k . The PI Q is said to be an “exact match” for T_k if $Q \in F_k$. Since $\hat{y}(T_k, \mathbf{x}) = 1$ implies $\hat{y}(T_k, \mathbf{x}) \geq k$, a PI $Q \in F_k$ is associated with response categories k and higher.

A PI Q is said to be a “subset match” for F_k if Q is an exact match for F_k or $Q \wedge P$ is an element of F_k for some PI P . This concept applies to both the sets $F_1, F_2, \dots, F_K$ associated with the CART($\mathbf{W}, K$) model and the sets $F_2, F_3, \dots, F_K$ associated with the OLR($\mathbf{W}, K$) model. Subset matching enables us to fully detect contributions of PIs to the fitted trees in OLR and CART models.

3. Comparison of OLR and CART

We compare the performance of OLR and CART in three different scenarios encompassing six simulation cases described in Table 1. The first scenario is designed to assess the effect of response error, the second assesses the effect of latent (i.e., omitted) predictor variables, and the third assesses the impact of a misspecified response. In the latent predictor scenario, one or more of the prime implicants associated with the response includes a latent predictor that is not observed when building OLR and CART models. In the misspecified response scenario, the prime implicants associated with the response are the same for all categories greater than 1 and thus the response is not truly ordinal. Each simulation is defined by a true underlying logical relationship L relating the binary predictors $\mathbf{x}_i$ for an individual to the ordinal response y_i and a response error probability π , taken constant across the K categories. These logical relationships L are selected to represent PIs of varying sizes, PIs containing a latent predictor, and PIs including predictor complements. Simulated data are produced by first generating 50 Bernoulli predictors $\mathbf{x}_i$ red for each of a large number of subjects. The logical relationship L given in Table 1 is used to construct the underlying true response category $L(\mathbf{x}_i)$, and then the observed category y_i is determined such that $\Pr(y_i = L(\mathbf{x}_i)) = 1 - \pi$, and for $j \neq L(\mathbf{x}_i)$, $\Pr(y_i = j) = \pi / (K - 1)$. Training data $\mathbf{W}$ of size n is then sampled as a case-control study such that $n/2$ observations are drawn with $y_i = 1$ (no disease) and $n / [2(K - 1)]$ observations are uniformly drawn from each of the remaining response categories. These training data $\mathbf{W}$ are used to construct OLR ($\mathbf{W}, K$) and CART ($\mathbf{W}, K$) models. As an example, for the first simulation case in Table 1, the 50 binary predictors are generated independently as $X_j \sim \text{Bernoulli}(p_j)$ $j = 1, 2, \dots, 50$ with $p_j = 0.35$ for $j \in \{10, 25, 45\}$, $p_j = 0.45$ for $j \in \{5, 30\}$, $p_j = 0.55$ for $j \in \{20, 40\}$, and $p_j = 0.50$ for all other j . For subject i , the true response is $L(\mathbf{x}_i) = 3$ if

$L_{\geq 3}^A(\mathbf{x}_i) = 1, L(\mathbf{x}_i) = 2$ if $L_{\geq 3}^A(\mathbf{x}_i) = 0$ and $L_{\geq 2}^A(\mathbf{x}_i) = 1$, else $L(\mathbf{x}_i) = 1$. The observed response y_i is then taken equal to $L(\mathbf{x}_i)$ with probability $1 - \pi = 0.9$, and otherwise the value is drawn uniformly from the response categories excluding $L(\mathbf{x}_i)$. The simulated case-control data are then sampled from these observations as described.

We consider sample sizes of 36-1000. For each combination of simulation case and sample size we generate 2500 training data sets matching the criteria described in Table 1 and fit an OLR model and two CART models using: (1) the nominal Gini index with misClassification costs set to 1, and (2) the generalized Gini index with the absolute cost function. For CART, we also considered the generalized Gini index with the quadratic cost function however the results were similar to the absolute cost so only the results using the absolute cost function are reported. We evaluate the performance of each method based on how well the fitted trees recover the true model's prime implicants. Model comparisons included: (a) exact identification of PIs in L , (b) identification of PIs in L as a subset match according to the definition in Section 2.4, and (c) model error rates of the fitted models, estimated by the proportion of times each method incorrectly classifies the observed values in independent test data of 100 observations generated in the same way as the training data.

The *rpart* package in R v 2.15.3 is used to fit a CART model using the Gini index as the splitting criterion and using the default pruning parameters provided in the package. The *rpartScore* package [12] is used to fit an ordinal CART model using the generalized Gini index with the absolute cost function. CART models fit with *rpartScore* are pruned using the default pruning parameters provided in the package. The OLR method described in Section 2.3 is used to generate all OLR models. The simulated annealing algorithm for model optimization that is available in LR [17] is used to construct all OLR models. At each step in the OLR algorithm, the initial annealing temperature is selected such that approximately 90% of the "new" models are accepted. The final temperature is set to achieve acceptance of fewer than 5% of new models. The cooling schedule is set so that 50,000 iterations are required to get from start to end temperature.

3.1. Response Error

Simulation cases 1 and 2 (Table 1) examine the effect of error in the response on the performance of nominal and ordinal CART and OLR. Although our focus is on recovery of PIs truly associated with the ordinal response, we first report model error rates in Table 2; these support that OLR has comparable or better prediction of the test data than CART procedures, and that error rates decrease with increasing sample size. The standard error of the mean model error rates ranged from 0.50 ($n = 36$) to 0.02 ($n = 1000$) in all cases and varied little by method.

Figures 1 and 2 show the proportion of times each method recovers PIs exactly or as subset matches known to be in L^A for data with 10 and 25% error in the observed response. A PI is counted as recovered only when it matches (i.e. predicts) the correct response category. In this scenario, OLR is more likely to exactly recover those PIs for response category 2, the lower response category, relative to both CART methods at both error levels. When response error is 10%, OLR more frequently exactly identifies PIs for the highest response category (category 3) relative to the CART methods. However, as error in the response increases, the ordinal CART models exactly identified PIs associated with the highest response category more often than OLR or nominal CART at smaller sample sizes. Under subset matching, all four methods show improved ability to recover the associated PIs for each response category. OLR recovers via subset matching all PIs in L^A more frequently

than the two CART methods for $n = 100$ at both error levels. The performance of ordinal CART improves for larger sample sizes, becoming comparable to, or, in the case of the PI $X_{20} \wedge X_{30} \wedge X_{40}$, slightly better than OLR.

We also examined the response error scenario in data where the observed response included 4 categories and 25% error. The results when $K = 4$ are similar to case 2. OLR exactly identifies PIs in the lower response categories, 2 and 3, more frequently than the two CART methods. However, the ordinal CART model identifies those PIs in category 4 more frequently than OLR at smaller sample sizes. Under subset matching, OLR identifies PIs for all categories more frequently than the CART models for $n = 150$. Details of this simulation and results are provided in the supplemental material (Supplemental Figure 1).

3.2. Latent Predictors

Simulation cases 3-5 (Table 1) examine the effect of latent predictors on the performance of nominal CART, ordinal CART, and OLR. Corresponding to the latency of predictor(s) X_{51} and X_{52} , these predictors are not available for constructing OLR and CART based models for these simulation cases. Average model error rates for select sample sizes are shown in Table 2.

Prime implicants that include the latent predictor cannot be identified exactly. Thus we determine the proportion of times the observed components of those PIs are recovered with no additional components. Figures 3 and 4 show the proportion of times each method recovers the observed components of the PIs known to be in L^B and L^C (Table 1). Results for L^D are shown in Supplemental Figure 2. Across all three latency cases (case 3-5 in Table 1), OLR more frequently exactly recovers those PIs that include a latent predictor. OLR also exactly recovers all PIs associated with response category 2 more often than both CART methods for $n = 200$, regardless of latency. In case 3 where a category 2 PI includes the latent predictor, the three methods recover PIs in category 3 with similar frequency to case 1 (10% response error scenario). In cases 4 and 5 where a PI associated with category 3 includes a latent predictor, all of the methods recover the PI $X_{25} \wedge X_{45}$ more often than in the response error scenario. However, ordinal CART recovers $X_{25} \wedge X_{45}$ significantly more often than either CART or OLR for $75 \leq n \leq 750$. All three methods improve under subset matching with the most notable improvements occurring for PIs that include the latent predictors. However, OLR recovers all PIs in L^B , L^C , and L^D via subset matching more frequently than the two CART methods for $n = 100$.

3.3. Misspecified Response

Simulation case 6 (Table 1) examines the impact of a misspecified ordinal response on the performance of nominal CART, ordinal CART, and OLR. In this scenario, prime implicants associated with response are the same for response levels 2 and 3 and thus the response is not truly ordinal. We considered the model as correctly predicting the response if $\hat{y} = 1$ when $y = 1$ or if $\hat{y} > 1$ when $y = 2$ or 3. In this scenario, both CART methods have smaller prediction error than OLR for $n = 150$ (Table 2).

Since the PIs associated with the response are the same for response levels 2 and 3, we determine the proportion of times the PIs are identified for response categories $Y = 2$. Figure 5 shows the proportion of times each method recovers the PIs known to be in L^E (Table 1) as exact and subset matches. OLR more frequently exactly identifies the PIs relative to the CART methods for $n = 75$. However, under subset matching, both versions of CART more frequently identify the PIs in L^E relative to OLR at larger sample sizes ($n = 200$).

3.4. OLR using Cross-Validation to select LR model size

Ruczinski *et al.* [26] suggest performing a cross-validation (CV) procedure to identify the most appropriate model size to avoid over-fitting an LR model. This procedure splits the data into c data sets with approximately equal numbers of observations and constructs c LR models for a particular model size (i.e. number of predictors in the model) using the original data with the c^{th} data set removed. Model error rate is estimated for the c^{th} data set and average model error rate for models of a particular size is obtained. This is repeated for a range of model sizes. The model size yielding the lowest mean error rate is selected as the “best” size to avoid over-fitting.

The LR trees composing the OLR models constructed in the simulations in sections 3.1 and 3.2 were allowed to grow to maximum size. The fact that PIs are most frequently identified as subset matches in our simulation results suggests over-fitting of trees within the OLR models. In order to correct for over-fitting in OLR models, the LR CV procedure can be employed at each stage of model construction to select a maximum model size for each LR tree, T_ℓ within the final OLR model T . We repeated simulation cases 1, 2, and 5 (Table 1) using 5-fold CV to select maximum model size for each T_ℓ within T and compared these results to OLR without CV. We refer to OLR with use of CV as OLR.CV.

In all three cases, OLR.CV more frequently identifies all PIs as exact matches than OLR except for PI $X_{20} \wedge X_{30} \wedge X_{40}$ (Figure 6 and Supplemental Figures 3 and 4). OLR and OLR.CV exactly recover $X_{20} \wedge X_{30} \wedge X_{40}$ with similar frequency in all three scenarios, though OLR recovers this PI more often for $n = 75$. The difference between OLR.CV and OLR in ability to exactly recover the remaining PIs is most noticeable as response error increases and for recovery of the observed components of PIs including a latent predictor. The differences are also greatest at smaller sample sizes ($n = 50$ to 300). Improvements in performance under subset matching is most noticeable for OLR. The ability to identify these PIs as subset matches is similar for OLR and OLR.CV across these three cases. These results suggest that use of CV in constructing OLR models is particularly useful at smaller sample sizes ($n < 300$).

3.5. Summary of Simulation Results

In the first two simulation scenarios, OLR is more likely to exactly recover those PIs for the lower response categories. OLR also exactly recovers the observed components of the PIs that included a latent predictor relative to nominal and ordinal CART models. However, ordinal CART exactly recovers the PIs in the highest response category more often than either CART or OLR when samples sizes are small ($n = 200$) with the exception of the PI including the latent predictor in scenarios 4 and 5. The performance of the three methods

improves if subset matching is taken into consideration. However, OLR tends to perform better than the CART based methods across all categories for the first two simulation scenarios for $n = 100$. Under subset matching, the performance of the three methods is similar for larger n .

In the scenario where the response is not truly ordinal (case 6, Table 1), OLR exactly identifies the PIs associated with the response more often than either CART method. However, under subset matching both CART methods identify the PIs more often than OLR when n is large.

We also ran simulations in data with $K = 4$, which reinforce the results observed for $K = 3$. OLR exactly identifies PIs in the lower categories more frequently than all CART methods. However, ordinal CART identifies those PIs in category 4 more frequently than OLR when n is small.

4. Severity of Periodontitis in African American Adults with Diabetes

We illustrate OLR by exploring the association of genetic and health factors with the severity of generalized adult periodontitis using data from a study conducted at the Center for Oral Health Research at the Medical University of South Carolina [10]. A subject has generalized adult periodontitis if at least 30% of periodontal sites have clinical attachment level (CAL) ≥ 3 mm. Severity has three categories: none if the subject does not have generalized adult periodontitis; mild if at least 30% of sites have $3 \leq \text{CAL} < 4$ mm; and moderate/severe if at least 30% of sites have $\text{CAL} \geq 4$ mm.

These data are drawn from 244 African American adults with diabetes. Information on each subject includes seven binary health indicators for total cholesterol (> 200 mg/dL), HDL (> 40 mg/dL), triglycerides (> 150 mg/dL), C-reactive protein levels (> 1 mg/L), HbA1c levels ($> 7\%$), BMI (> 25), and smoking status (ever versus never), together with genotype data for nine single nucleotide polymorphisms (SNPs) believed to play a role in inflammation and/or bone resorption. Among the 244 participants in the study, 155 have no periodontitis, 40 have mild periodontitis, and 49 have moderate to severe periodontitis.

Seven of the SNPs are coded by two dummy variables for analysis. The first dummy variable takes value 1 if the subject has a SNP genotype with at least one copy of the minor allele (dominant effect of the minor allele) and the second takes value 1 if the subject has two copies of the minor allele (recessive effect of the minor allele). This coding allows for consideration of both dominant and recessive genetic effects in the model. The remaining two SNPs, for which no subjects have two copies of the minor allele, are coded for the dominant genetic effect only. The final data set therefore includes 23 binary predictors composed of these 16 SNP dummy variables and seven health indicators. An OLR model using cross-validation, a nominal CART model, and an ordinal CART model using the absolute cost function were constructed from these data, treating individuals with no periodontitis as the reference category.

OLR identified a PI representing the interaction between HbA1c, smoking and the dominant effect of an IL-1 α 2 SNP as associated with moderate/severe periodontitis. It further

identified two SNP interactions as associated with mild periodontitis, namely, the interaction of the recessive effect of an IL-1 α 1 SNP with, respectively, the dominant effects of an IL-1 β and an IL-10 SNP. The separate associations with periodontitis of these SNPs, smoking and HbA1c are consistent with previous reports [7, 14, 18, 23].

The nominal CART model was similar to the ordinal CART model, identifying the interaction between ever having smoked and the recessive effect of IL-1 α 2 SNP as being associated with moderate to severe disease, which is similar to the PI identified by OLR. All three models also identified the interaction between having ever smoked, the absence of the recessive effect of IL-1 α 2 SNP and elevated triglycerides as being associated with moderate to severe disease. Nominal CART did not identify any predictors or interactions as being associated with mild periodontitis. However, the ordinal CART model identified the interaction between not smoking and the dominant effect of TNF α SNP as being associated with mild disease. The OLR model and the two CART models are shown in Supplemental Figures 5 and 6.

The three models identified similar associations for moderate to severe disease. This example emphasizes the flexibility of OLR relative to CART. PIs in OLR are distinctly defined for each cumulative disease severity category. However, in the CART model, when branches have more than one split, the variable that was split on must be included in one interaction term as occurrence, and the other branch must contain the complement. Thus, for example, the ordinal CART models identified the interaction (never smoked $\wedge$ dominant effect of TNF α). It may be that the TNF α SNP is more likely to impact development of periodontitis in non-smokers, however, this interaction may arise due to the less flexible structure of CART trees.

5. Discussion

We developed Ordinal Logic Regression (OLR) to extend logic regression for classification of an ordinal response. Our simulation studies comparing OLR, CART using the Gini index with misClassification cost all equal to 1, and CART using the generalized Gini index with the linear cost function demonstrate that when data include exclusively binary predictors and the response is ordinal, OLR is better able to identify interactions associated with the ordinal response and has smaller model error rate than CART with rare exceptions.

OLR has a less rigid model structure than CART, regardless of the choice of splitting rule, providing more flexibility in identifying interactions predictive of an outcome. For example, if individuals in category k are defined by the PIs $X_{10} \wedge X_{45} = 1$ or $X_{25} \wedge X_{45} = 1$, CART may initially split on $X_{45} > 0.5$ dividing individuals into those for whom $X_{45} = 1$ and those for whom $X_{45} = 0$. A second split in the branch $X_{45} = 1$ might be $X_{10} > 0.5$, corresponding to the PI $X_{45} \wedge X_{10} = 1$ which predicts a response of k . Here the PI $X_{10} \wedge X_{45}$ would be exactly matched and this branch correctly classifies individuals in category k having predictor values $X_{10} = 1$ and $X_{45} = 1$. However, there is a set of individuals in category k due to PI $X_{25} \wedge X_{45}$ not correctly classified by this branch. Thus CART must split the branch $X_{45} = 1$ and $X_{10} = 0$ on $X_{25} > 0.5$. These individuals are classified by the PI $X_{45} \wedge \neg X_{10} \wedge X_{25}$ which is not an exact match for $X_{25} \wedge X_{45}$. Due to this rigid structure, CART cannot exactly

match both $X_{10} \wedge X_{45}$ and $X_{25} \wedge X_{45}$. In contrast, by allowing for the use of both $\wedge$ and $\vee$ operators, OLR has greater flexibility in modeling the response.

The flexibility of OLR is apparent in its ability to exactly identify PIs relative to CART. The greatest improvements in performance for OLR relative to CART occur at smaller sample sizes ($n = 300$). OLR exactly identifies the true PIs with greater frequency than CART in data with noisy predictors. In all of the latent predictor cases, OLR more consistently identifies true PIs that do not include the latent predictor relative to CART. The performance of both methods improves with subset matching although OLR identifies subset matches more often than CART in scenarios with an ordinal response for $n = 300$. Our results suggests that, when data include exclusively binary predictors, OLR is better suited to handle smaller data sets and will provide superior performance to CART. If the response has been misspecified as being ordinal when it is not, OLR more consistently exactly matches the PIs, while CART performs better relative to OLR under subset matching for $n = 200$. However, even under these circumstances, OLR still correctly identifies the PIs associated with the response as subset matches approximately 80% of the time for $n = 150$.

Like CART, the performance of OLR degrades as instability is introduced into the data (i.e. noisy predictors and latent predictors). Random Forests, an ensemble adaptation of CART, has been shown to improve predictive performance of CART. Additionally Random Forests provides a measure of predictor importance. In the future we plan to extend our work with OLR to an ensemble application to improve performance and provide quantitative measures of both predictor and PI importance.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was partially supported by National Institute of General Medicine Grant T32GM074934, National Cancer Institute Grant R03CA137805, National Institute of Dental and Craniofacial Research Grant K25DE016863, National Institute of Dental and Craniofacial Research Grant P20RR017696, National Science Foundation Division of Mathematical Sciences Grant 0604666, and South Carolina Clinical and Translational Research Institute NIH/NCRR Grant UL1RR029882. We thank the Center for Oral Health Research and the Medical University of South Carolina for providing the illustrative data of Section 4. We also thank the reviewers for their insightful comments that have greatly improved this manuscript.

References

- [1]. Agresti A. Tutorial on modeling ordered categorical response data. *Psychological Bulletin*. Mar; 1989 105(2):290–301. [PubMed: 2648444]
- [2]. Agresti A. Modeling ordered categorical data: recent advances and future challenges. *Statistics in Medicine*. 1999; 18:2191–2207. [PubMed: 10474133]
- [3]. Alaiya AA, Franzn B, Auer G, Linder S. Cancer proteomics: from identification of novel markers to creation of artificial learning models for tumor Classification. *Electrophoresis*. 2000; 21(6): 1210–1217. URL <http://dx.doi.org/3.0.CO;2-S>. [PubMed: 10786893]
- [4]. Alvarez-Castro JM, Carlborg O. A unified model for functional and statistical epistasis and its application in quantitative trait loci analysis. *Genetics*. 2007; 176(2):1151–1167. URL <http://dx.doi.org/10.1534/genetics.106.067348>. [PubMed: 17409082]

- [5]. Archer K. rpartOrdinal: an R package for deriving Classification trees for predicting an ordinal response. *Journal of Statistical Software*. 2012; 34(7):1–17.
- [6]. Austin PC. A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. *Statistics in Medicine*. 2007; 26(15):2937–57. [PubMed: 17186501]
- [7]. Baldi D, Izzotti A, Bonica P, Pera P, Pulliero A. Degenerative periodontal-diseases and oral osteonecrosis: the role of gene-environment interactions. *Mutat Res*. 2009; 667(1-2):118–131. URL <http://dx.doi.org/10.1016/j.mrfmmm.2008.11.005>. [PubMed: 19056406]
- [8]. Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. *Classification and Regression Trees*. Chapman & Hall/CRC; Boca Raton: 1984.
- [9]. Cianfrocca M, Gradishar W. New molecular Classifications of breast cancer. *CA Cancer Journal for Clinicians*. 2009; 59(5):303–313. URL <http://dx.doi.org/10.3322/caac.20029>.
- [10]. Fernandes JK, Wiegand RE, Salinas CF, Grossi SG, Sanders JJ, Lopes-Virella MF, Slate EH. Periodontal disease status in gullah african americans with type 2 diabetes living in south carolina. *J Periodontol*. 2009; 80(7):1062–1068. URL <http://dx.doi.org/10.1902/jop.2009.080486>. [PubMed: 19563285]
- [11]. Fleisher H, Tavel M, Yeager J. Exclusive-OR representation of Boolean functions. *IBM Journal of Research and Development*. 1983; 27(4):412–416.
- [12]. Galimberti G, Soffritti G, Maso M. Classification trees for ordinal responses in R: The rpartScore package. *J. Stats. Software*. 2012; 47(10):1–25.
- [13]. Garte S. Metabolic susceptibility genes as cancer risk factors: time for a reassessment? *Cancer Epidemiology Biomarkers and Prevention*. 2001; 10(12):1233–1237.
- [14]. Gore EA, Sanders JJ, Pandey JP, Palesch Y, Galbraith GM. Interleukin-1 β ³⁹⁵³ allele 2: association with disease status in adult periodontitis. *Journal of Clinical Periodontology*. 1998; 25(10):781–785. [PubMed: 9797049]
- [15]. Hastie, T.; Tibshirani, R.; Friedman, J. Springer Science and Business Media, LLC; 2009. *The elements of statistical learning: data mining, inference, and prediction*, 2nd Edition.
- [16]. Kiyohara C, Otsu A, Shirakawa T, Fukuda S, Hopkin JM. Genetic polymorphisms and lung cancer susceptibility: a review. *Lung Cancer*. 2002; 37(3):241–256. [PubMed: 12234692]
- [17]. Kooperberg C, Ruczinski I. LogicReg: Logic Regression. 2007 R package version 1.4.7.
- [18]. Kornman KS, Crane A, Wang HY, di Giovine FS, Newman MG, Pirk FW, Wilson TG, Higginbottom FL, Duff GW. The interleukin-1 genotype as a severity factor in adult periodontal disease. *Journal of Clinical Periodontology*. 1997; 24(1):72–77. [PubMed: 9049801]
- [19]. Kotti S, Bickeboller H, Clerget-Darpoux F. Strategy for detecting susceptibility genes with weak or no marginal effects. *Human Heredity*. 2007; 63(2):85–92. [PubMed: 17283437]
- [20]. Kumar S, Mohan A, Guleria R. Biomarkers in cancer screening, research and detection: present and future: a review. *Biomarkers*. 2006; 11(5):385–405. [PubMed: 16966157]
- [21]. Liu Z, Sokka T, Maas K, Olsen NJ, Aune TM. Prediction of disease severity in patients with early rheumatoid arthritis by gene expression profiling. *Human Genomics and Proteomics*. 2009:1–10. 2009.
- [22]. Manne U, Srivastava RG, Srivastava S. Recent advances in biomarkers for cancer diagnosis and treatment. *Drug Discovery Today*. 2005; 10(14):965–76. [PubMed: 16023055]
- [23]. Morita I, Inagaki K, Nakamura F, Noguchi T, Matsubara T, Yoshii S, Nakagaki H, Mizuno K, Sheiham A, Sabbah W. Relationship between periodontal status and levels of glycated hemoglobin. *J Dent Res*. Feb; 2012 91(2):161–166. URL <http://dx.doi.org/10.1177/0022034511431583>. [PubMed: 22157098]
- [24]. Peto J, Houlston RS. Genetics and the common cancers. *European Journal of Cancer*. 2001; 37(Supplemental 8):S88–S96. [PubMed: 11602375]
- [25]. Piccarreta R. Classification trees for ordinal variables. *Computational Statistics*. 2008; 23(3) 407:427.
- [26]. Ruczinski I, Kooperberg C, LeBlanc M. Logic regression. *Journal of Computational and Graphical Statistics*. 2003; 12(3):475–511.

- [27]. Schwender H, Ickstadt K. Identification of SNP interactions using logic regression. *Biostatistics*. 2008; 9(1):187–198. [PubMed: 17578898]
- [28]. Srivastava S. Cancer biomarkers: an emerging means of detecting, diagnosing and treating cancer. *Cancer Biomarkers*. 2005; 1(1):1–2. [PubMed: 17192027]
- [29]. Ullah MF, Aatif M. The footprints of cancer development: Cancer biomarkers. *Cancer Treatment Reviews*. 2009; 35(3):193–200. URL <http://dx.doi.org/10.1016/j.ctrv.2008.10.004>. [PubMed: 19062197]

Figure 1.

Proportion of times each method recovers true PIs L^A as (A) exact and (B) subset matches in data with 10% error in the response. $N = 2500$ replications for each sample size (case 2, Table 1). The confidence intervals are 95% simulation intervals provided to convey the precision of the estimated proportion of times the methods recover each PI across simulation runs and are smaller than the point size. Note results for $X_{10} \wedge X_{45}$ and $X_{25} \wedge X_{45}$ are similar thus only $X_{10} \wedge X_{45}$ is included in the figure.

Figure 2. Proportion of times each method recovers true PIs L^A as (A) exact and (B) subset matches in data with 25% error in the response. $N = 2500$ replications for each sample size (case 2, Table 1). The confidence intervals are 95% simulation intervals provided to convey the precision of the estimated proportion of times the methods recover each PI across simulation runs and are smaller than the point size. Note results for $X_{10} \wedge X_{45}$ and $X_{25} \wedge X_{45}$ are similar thus only $X_{10} \wedge X_{45}$ is included in the figure.

Figure 3. Proportion of times each method recovers PIs in L^B as (A) exact and (B) subset matches. Since PI $X_5 \wedge !X_{20} \wedge X_{51}$ includes the latent predictor X_{51} , we report recovery of the observed component $X_5 \wedge !X_{20}$. $N = 2500$ replications for each sample size (case 2, Table 1). The confidence intervals are 95% simulation intervals provided to convey the precision of the estimated proportion of times the methods recover each PI across simulation runs and are smaller than the point size. Note results for $X_{10} \wedge X_{45}$ and $X_{25} \wedge X_{45}$ are similar thus only $X_{10} \wedge X_{45}$ is included in the figure.

Figure 4. Proportion of times each method recovers PIs L^C as (A) exact and (B) subset matches. Since PI $X_{10} \wedge X_{45} \wedge X_{51}$ includes the latent predictor X_{51} , we report recovery of the observed component $X_{10} \wedge X_{45}$. $N = 2500$ replications for each sample size (case 2, Table 1). The confidence intervals are 95% simulation intervals provided to convey the precision of the estimated proportion of times the methods recover each PI across simulation runs and are smaller than the point size.

Figure 5. Proportion of times each method recovers true PIs L^E as (A) exact and (B) subset matches in data with 25% error in the response. $N = 2500$ replications for each sample size (case 6, Table 1). The confidence intervals are 95% simulation intervals provided to convey the precision of the estimated proportion of times the methods recover

Figure 6.

Proportion of times OLR with and without cross-validation recovers PIs in L^D as (A) exact and (B) subset matches in data with 10% noise in all predictors. $N = 2500$ replications for each sample size (case 5, Table 1). The confidence intervals are 95% simulation intervals provided to convey the precision of the estimated proportion of times the methods recover each PI across simulation runs and are smaller than the point size.

Table 1

Simulation schemes for five simulations comparing OLR and CART

Scenario	Predictors in $L^\dagger$	L	Case	Response Error (π)
Response Error	$x_{10}, x_{25}, x_{45} \sim \text{Bern}(0.35)$ $x_5, !x_{20}, x_{30}, !x_{40} \sim \text{Bern}(0.45)$	$L_3^A = (X_{10} \wedge X_{45}) \vee (X_{25} \wedge X_{45})$ $L_2^A = L_3^A \vee (X_5 \wedge !X_{20}) \vee (!X_{20} \wedge !X_{30} \wedge !X_{40})$	1	0.10
			2	0.25
Latent Predictor	$x_{10}, x_{25}, x_{45} \sim \text{Bern}(0.3)$ $x_5, !x_{20}, x_{30}, !x_{40} \sim \text{Bern}(0.5)$ $x_{51} \sim \text{Bern}(0.5)$	$L_3^B = (X_{10} \wedge X_{45}) \vee (X_{25} \wedge X_{45})$ $L_2^B = L_3^B \vee (X_5 \wedge !X_{20} \wedge \mathbf{X}_{51}) \vee (!X_{20} \wedge !X_{30} \wedge !X_{40})$	3	0.10
	$x_{10}, x_{25}, x_{45} \sim \text{Bern}(0.3)$ $x_5, !x_{20}, x_{30}, !x_{40} \sim \text{Bern}(0.5)$ $x_{51} \sim \text{Bern}(0.75)$	$L_3^C = (X_{10} \wedge X_{45} \wedge \mathbf{X}_{51}) \vee (X_{25} \wedge X_{45})$ $L_2^C = L_3^C \vee (X_5 \wedge !X_{20}) \vee (!X_{20} \wedge !X_{30} \wedge !X_{40})$	4	0.10
	$x_{10}, x_{25}, x_{45} \sim \text{Bern}(0.3)$ $x_5, !x_{20}, x_{30}, !x_{40} \sim \text{Bern}(0.5)$ $x_{51} \sim \text{Bern}(0.5), x_{52} \sim \text{Bern}(0.75)$	$L_3^D = (X_{10} \wedge X_{45} \wedge \mathbf{X}_{52}) \vee (X_{25} \wedge X_{45})$ $L_2^D = L_3^D \vee (X_5 \wedge !X_{20} \wedge \mathbf{X}_{51}) \vee (!X_{20} \wedge !X_{30} \wedge !X_{40})$	5	0.10
Misspecified Response	$x_{10}, x_{25}, x_{45} \sim \text{Bern}(0.4)$	$L_3^E = (X_{10} \wedge X_{45}) \vee (X_{25} \wedge X_{45})$ $L_2^E = L_3^E$	6	0.25

$\ddagger$ $\mathbf{X}_{51}$ and $\mathbf{X}_{52}$ represent latent predictors and are not observed in the data used to construct OLR and CART models.

† All predictor not associated with the response: $x_i, i \in \{5, 10, 15, 20, 25, 30, 35, 40, 45\}$ $iid \text{ Bern}(0.5)$ for all simulation schemes.

Table 2

Mean model error rates for OLR and CART for all simulation cases (Table 1)

Simulation Case	Error in Y	Sample Size	OLR % Error	CART (nominal) % Error	CART (ordinal) % Error
1 (L^A)	10%	36	30.1	32.6	32.4
		100	7.10	14.8	13.6
		300	1.21	4.02	3.82
		1000	0.33	0.94	1.14
2 (L^A)	25%	36	39.6	37.1	37.3
		100	13.1	19.5	17.2
		300	2.88	5.61	4.77
		1000	0.97	1.14	1.08
3 (L^B)	10%	36	30.8	33.9	34.6
		100	9.83	17.5	15.7
		300	3.22	5.82	5.44
		1000	2.06	2.69	2.79
4 (L^C)	10%	36	35.6	36.3	35.9
		100	16.1	20.8	20.5
		300	8.14	10.0	9.92
		1000	5.81	6.65	6.98
5 (L^D)	10%	36	37.6	39.6	39.6
		100	19.8	23.6	23.3
		300	12.6	14.6	14.8
		1000	7.68	9.41	9.90
6 (L^E)	25%	36	20.6	19.8	15.7
		100	6.61	14.0	4.97
		300	6.11	3.24	3.21
		1000	6.49	0.97	0.86

Standard error of the mean model error rates ranged from 0.50 ($n = 36$) to 0.02 ($n = 1000$) in all cases and varied little by method.