

Logic Regression and its applications to SNP association studies

Ingo Ruczinski

Department of Biostatistics
Johns Hopkins University

Email: ingo@jhu.edu

<http://biostat.jhsph.edu/~iruczins>

With Charles Kooperberg and Michael LeBlanc, FHCRC

Motivation

[With Kathy Helzlsouer and Han-Yao Huang]

The odyssey cohort study consists of 8,394 participants who donated blood samples in 1974 and 1989 in Washington County, Maryland. The cohort has been followed until 2001, and environmental factors such as smoking and dietary intake are available. The goals of the study include finding associations between polymorphisms in candidate genes and mortality (including cancer and cardiovascular death). Particularly, gene-environment and gene-gene interactions associated with mortality are of interest. Currently, SNP data from 51 genes are available.

Introduction

X_1, \dots, X_k are 0/1 (False/True) predictors.

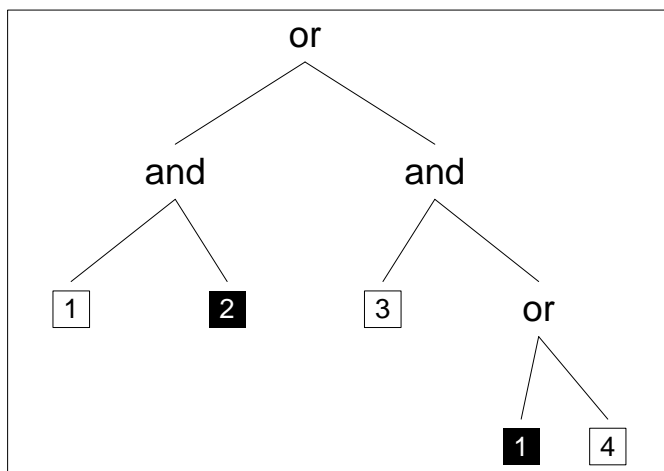
Y is a response variable.

Fit a model $g(E(Y)) = \beta_0 + \sum_{j=1}^t \beta_j \times L_j$, where L_j is a Boolean combination of the covariates, e.g. $L_j = (X_1 \vee X_2) \wedge X_4^c$.

Determine the logic terms L_j and estimate the β_j simultaneously.

Logic Trees

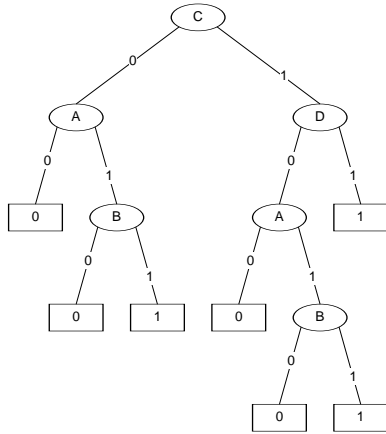
An equivalent representation of $(X_1 \wedge X_2^c) \vee (X_3 \wedge (X_1^c \vee X_4))$ is the following:



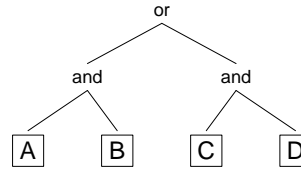
This is a Logic Tree!

Comparison to Decision Trees

Decision Tree



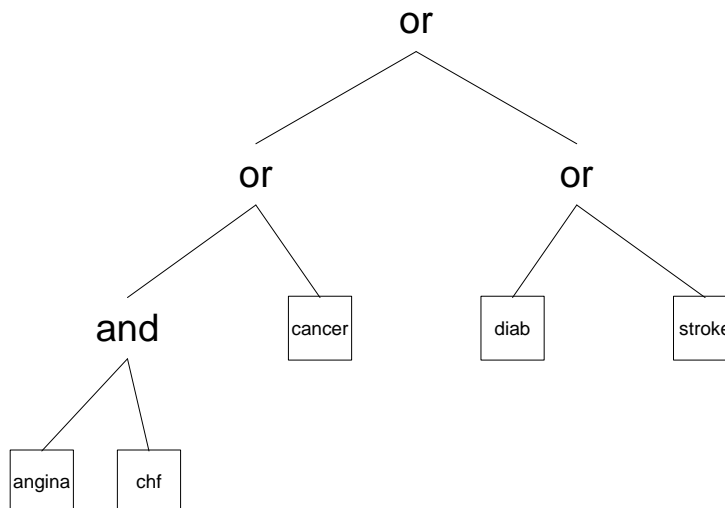
Logic Tree



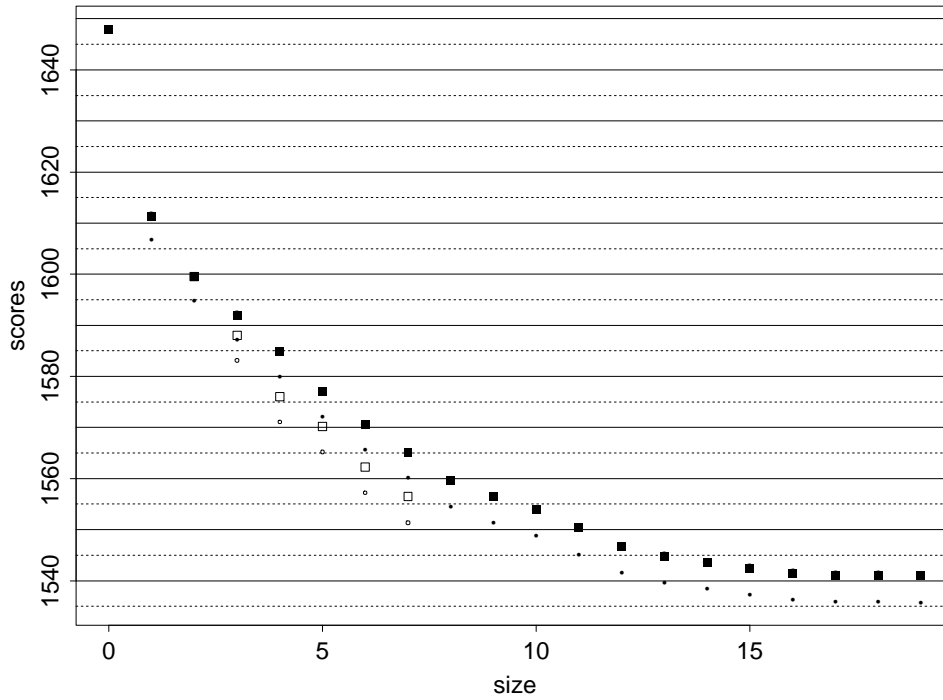
A Decision Tree (CART) is something different!

An Example

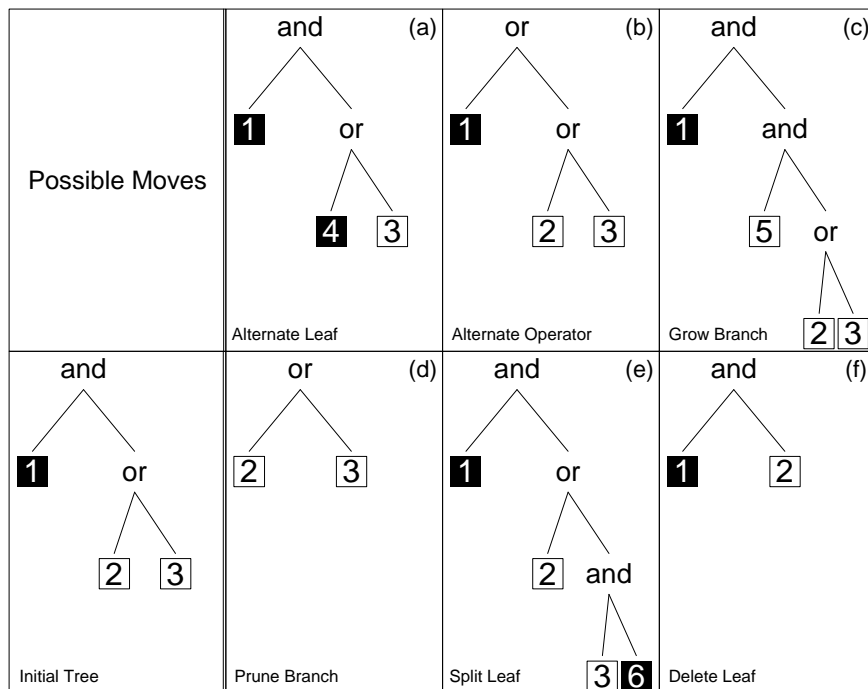
$$\text{logit}(\text{death}) = -9.01 + 0.06 \times \text{age} + 1.07 \times L$$



Example Cont.



The Move Set



Simulated Annealing for Logic Regression

We try to fit the model $g(E(Y)) = \beta_0 + \sum_{j=1}^t \beta_j \times L_j$.

- Select a scoring function (RSS, log-likelihood, ...).
- Pick the maximum number of Logic Trees.
- Pick the maximum number of leafs in a tree.
- Initialize the model with $L_j = 0$ for all j .
- Carry out Simulated Annealing Algorithm:
 - Propose a move.
 - Accept or reject the move, depending on scores and temperature scheme.

Simulated Annealing for Logic Regression (2)

- We use the acceptance function $\alpha(\epsilon_{old}, \epsilon_{new}, t) = \min \{1, \exp([\epsilon_{old} - \epsilon_{new}] / t)\}$
- We run homogeneous Markov chains at constant temperatures.
- We constructed the move set to be irreducible and aperiodic, therefore each homogeneous Markov chain has a limiting distribution $\pi_t(S)$.
- The limit (as $t \rightarrow 0$) of those distributions assigns probability 1 to the optimal scoring states.
- With limited resources, we cannot guarantee to find an optimal scoring state.

Another Example

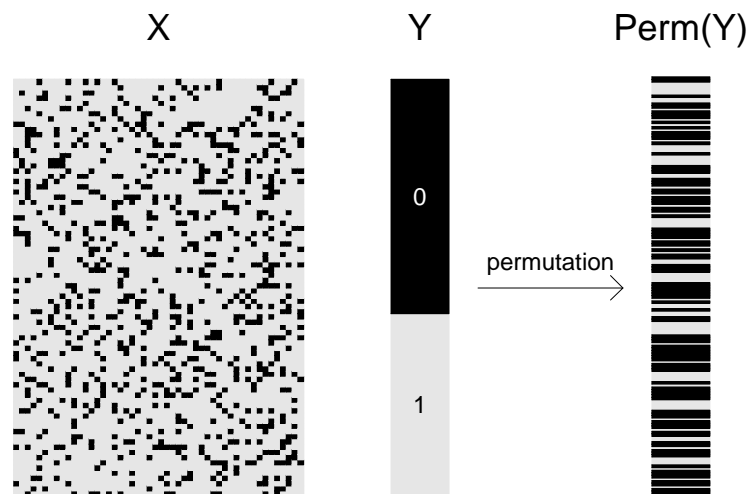
The Cardiovascular Health Study

(Fried et. al. , Annals of Epidemiology, 1991).

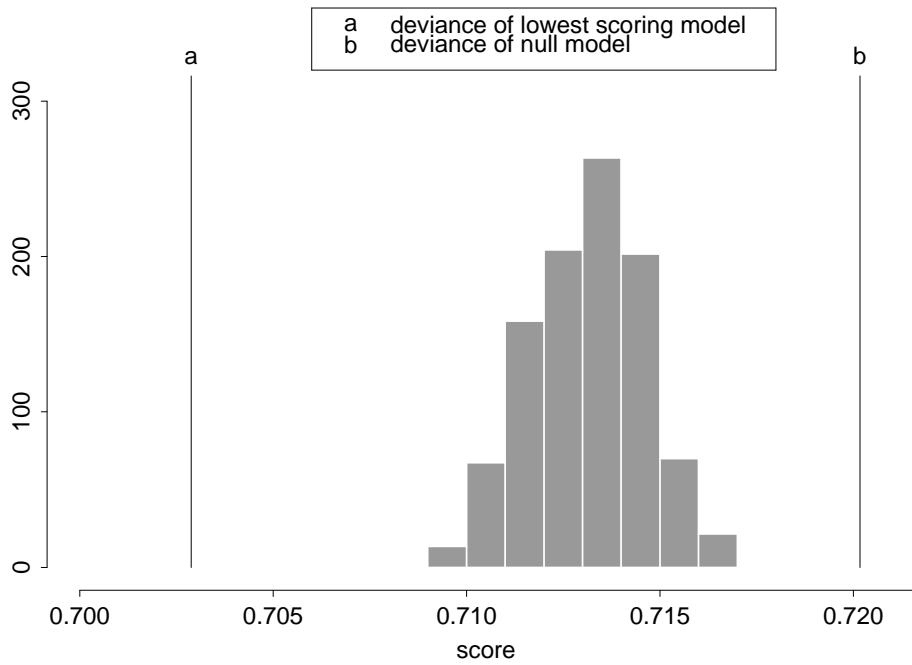
- The Cardiovascular Health Study (CHS) is a study of coronary heart disease and stroke in elderly people.
- Between 1989 and 1993, 5888 subjects over the age of 65 were recruited in four communities in the United States.
- During 1992 and 1994, a subset of these patients underwent an MRI scan.
- For 3647 CHS participants, MRI detected strokes (infarcts bigger than 3mm that led to deficits in functioning) were recorded as entries into a 23 region atlas of the brain.
- The mini-mental state examination is a brief screening test for dementia. The response Y is a variable derived by transforming the mini-mental score.

We investigated models of the form $Y = \beta_0 + \beta_1 \times L_1 + \dots + \beta_p \times L_p + \epsilon$.

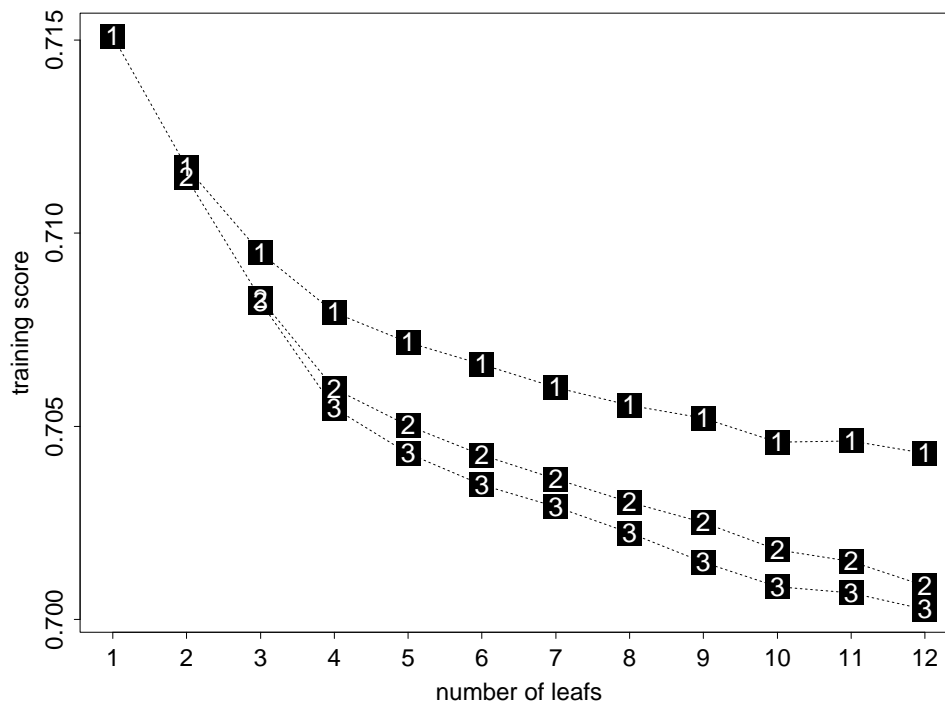
A Global Randomization Test of Association



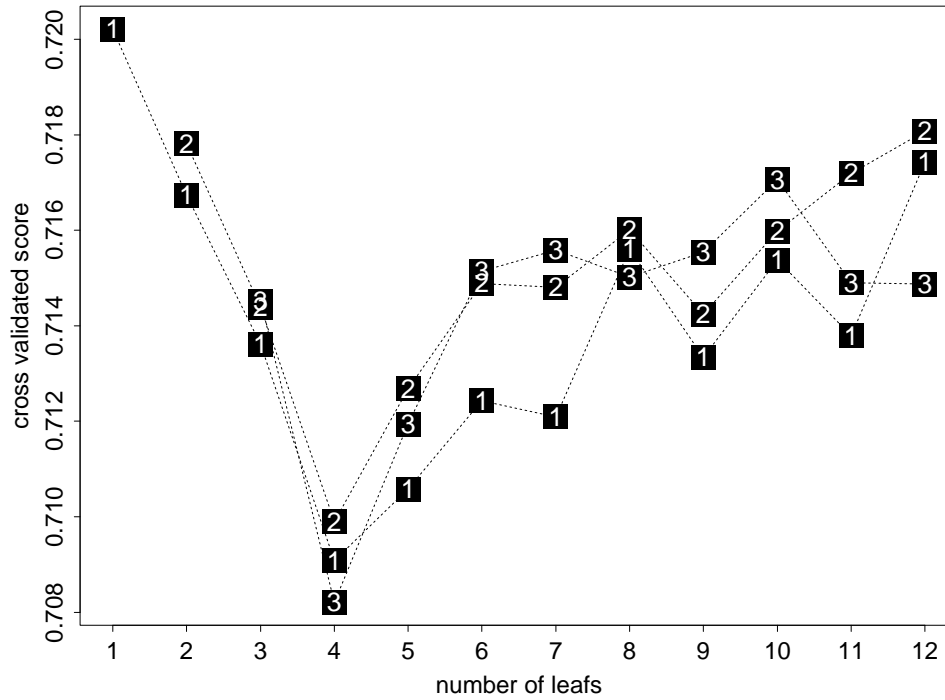
Example Cont.



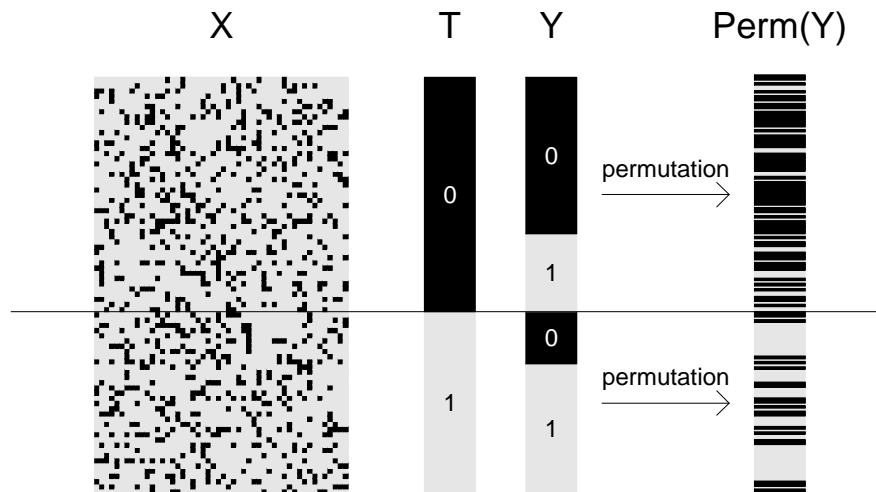
Cross-Validation



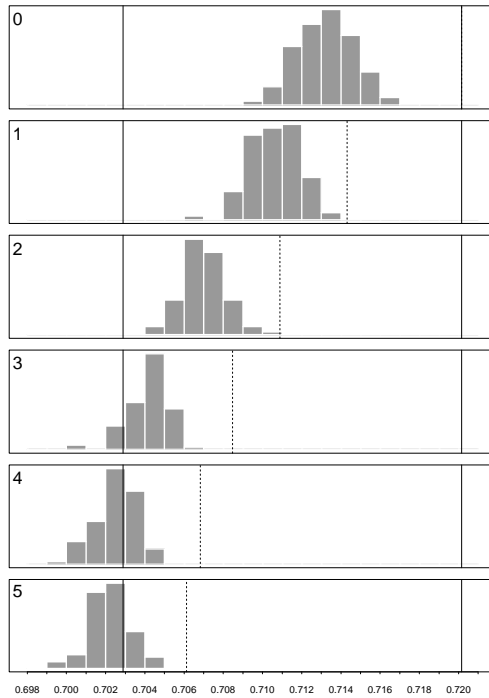
Cross-Validation Cont.



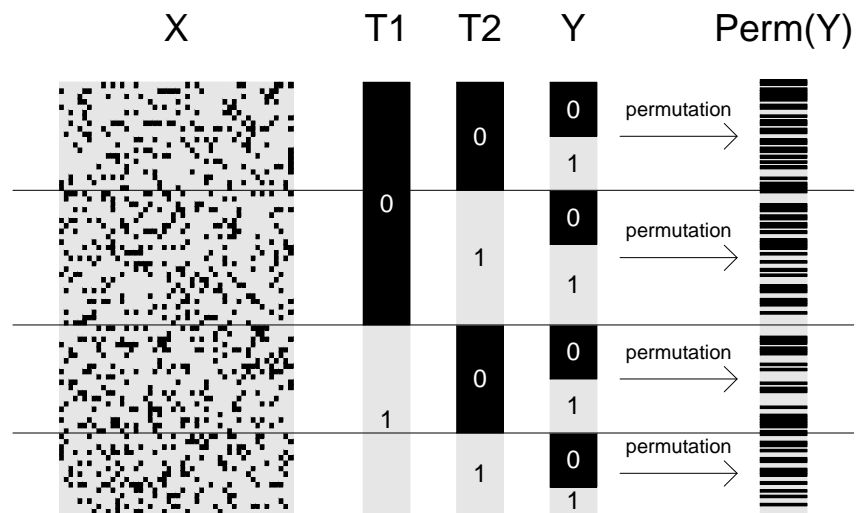
A Sequential Randomization Test for Model Size



Randomization Test Cont.

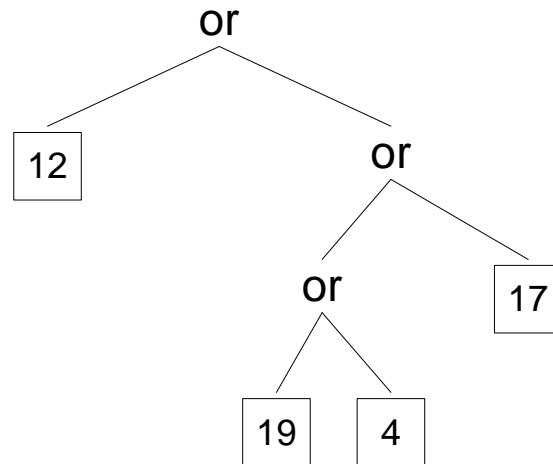


Sequential Randomization Test for 2 Trees:



Results

The model we found was $Y = 1.96 + 0.36 \times L$, with the following Logic Tree:



Logic Model vs Linear Model vs MARS

Linear model:

	$\hat{\beta}$	$\hat{\sigma}$	t-value
Intercept	1.961	0.015	133.98
Region 4	0.524	0.129	4.06
Region 12	0.460	0.112	4.09
Region 17	0.236	0.057	4.17
Region 19	0.611	0.157	3.89

Logic model:

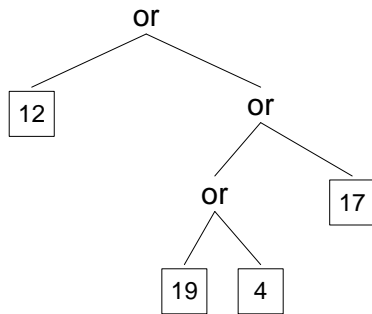
$$Y = 1.96 + 0.36 \times I_{\{X_4 \vee X_{12} \vee X_{17} \vee X_{19} \text{ is true}\}}$$

MARS:

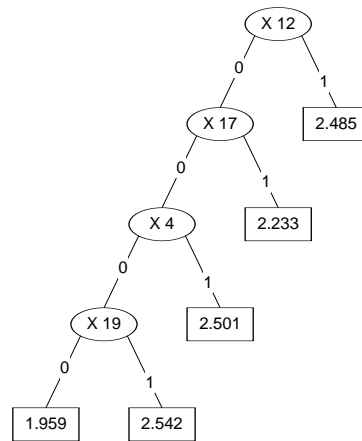
$$Y = 1.96 + 0.53 X_4 + 0.37 X_{12} + 0.24 X_{17} + 0.61 X_{19} + 1.05 (X_{12} * X_{15})$$

Logic Model vs CART

Logic Tree



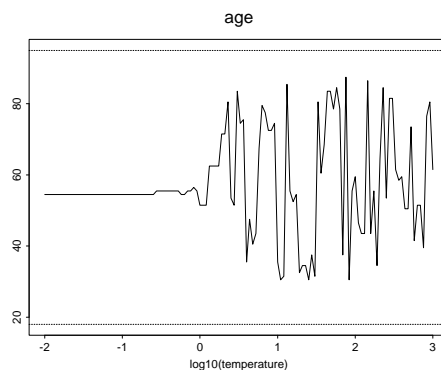
CART Tree



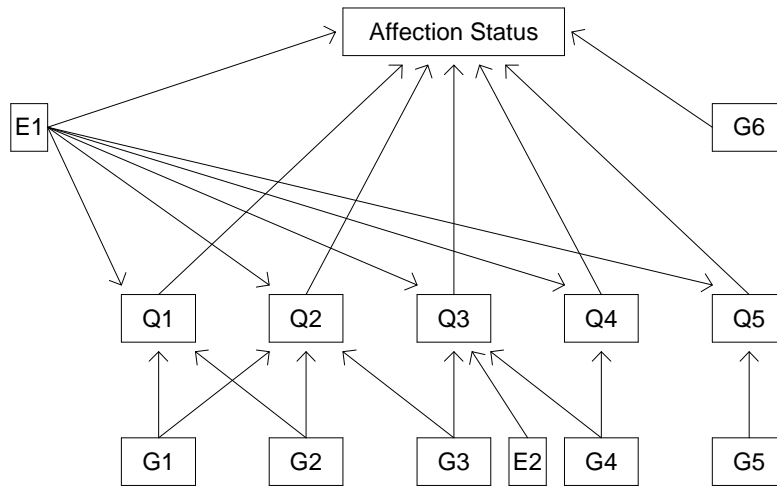
Continuous Predictors

What can we do about the continuous predictors?

- Incorporate continuous predictors in (generalized) linear model.
- Dichotomize continuous predictors before starting the simulated annealing chain.
- Find optimal split points during the simulated annealing run.



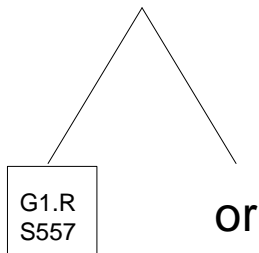
Genetic Analysis Workshop GAW 12



GAW12 Cont.

$$\text{logit}(\text{affected}) = \beta_0 + \beta_1 \times \text{ENV}_1 + \beta_2 \times \text{ENV}_2 + \beta_3 \times \text{GENDER} + \sum_{i=1}^K \beta_{i+3} \times L_i$$

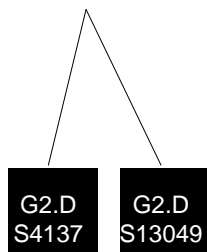
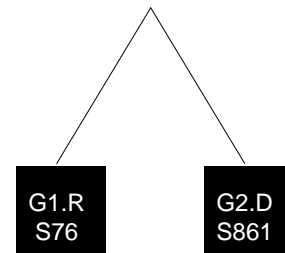
L₁= and



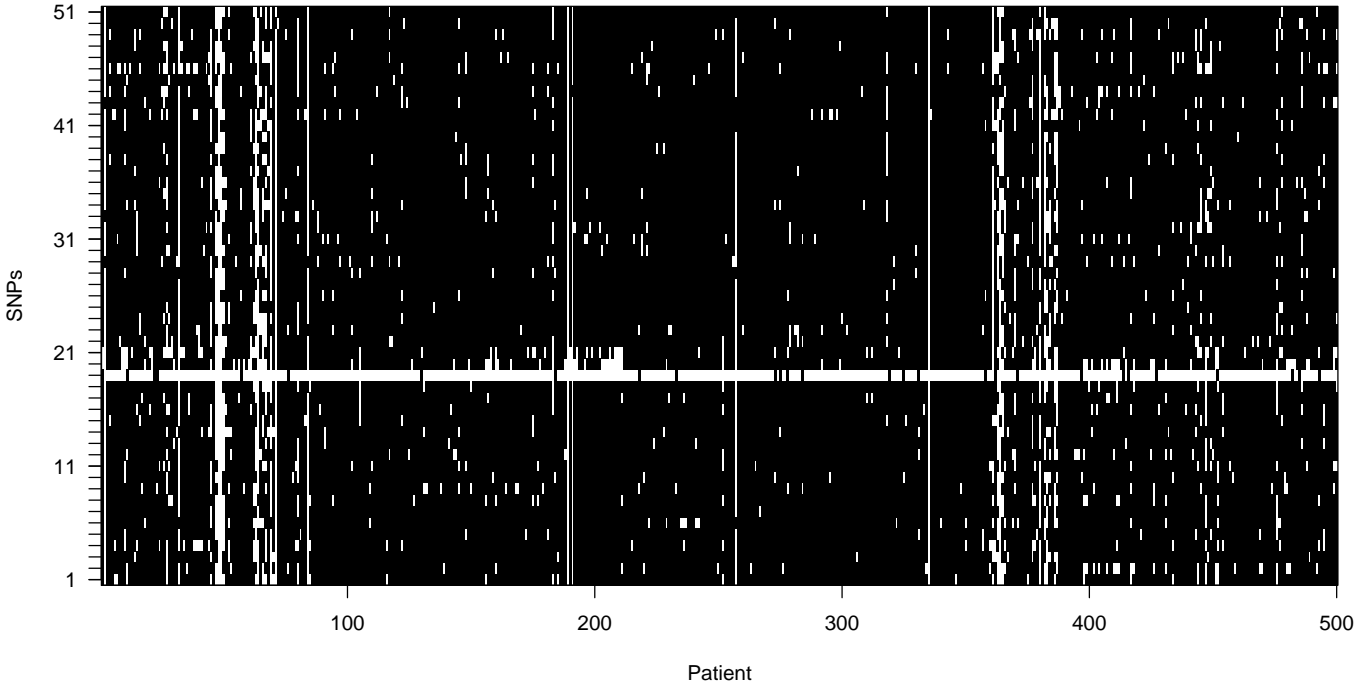
L₂=



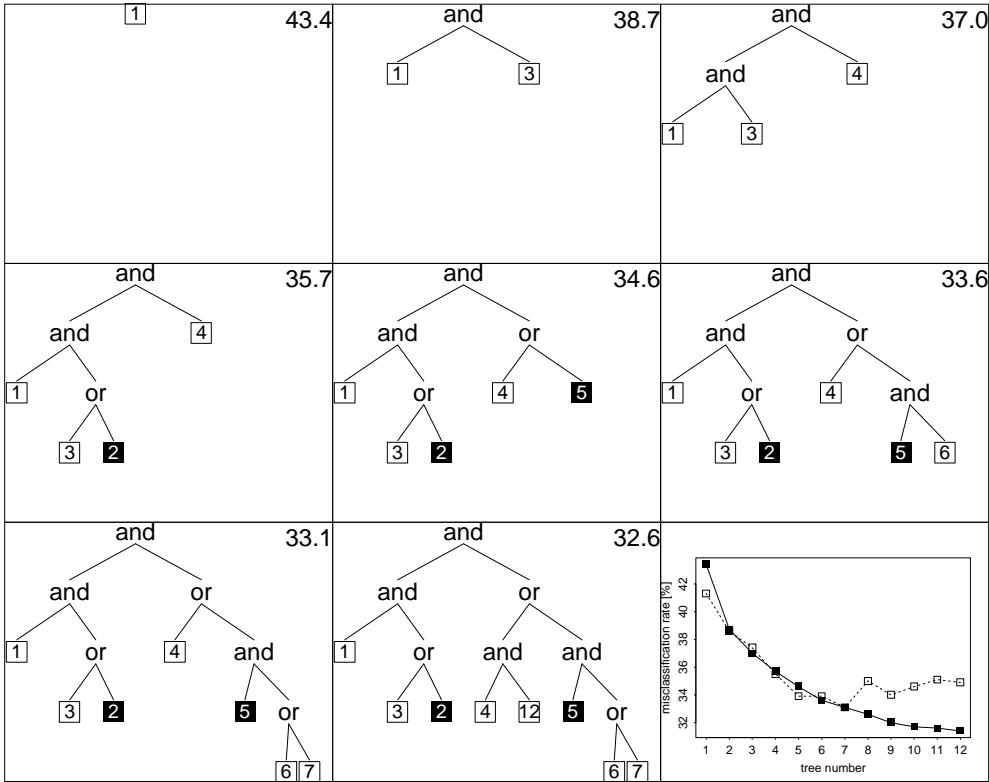
L₃= or



Missing Values



Greedy Search



References

- Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C.J. (1984), *Classification and Regression Trees*, Belmont, CA: Wadsworth.
- Friedman, J. H. (1991), *Multivariate Adaptive Regression Splines* (with discussion), *Annals of Statistics*, 19, 1-141.
- Kooperberg, C., Ruczinski, I., LeBlanc, M., and Hsu, L. (2001), *Sequence Analysis using Logic Regression*, *Genetic Epidemiology*, 21 (S1), 626-631.
- Ruczinski, I., Kooperberg, C., and LeBlanc, M. (2002), *Logic Regression - Methods and Software*, *Proceedings of the MSRI workshop on Nonlinear Estimation and Classification* (Eds: D. Denison, C. Holmes, M. Hansen, B. Mallick, B. Yu), Springer.
- Ruczinski, I., Kooperberg, C., and LeBlanc, M. (2002), *Logic Regression* (under review), <http://biostat.jhsph.edu/~iruczins/>