Logic Regression

Ingo Ruczinski

Department of Biostatistics
Johns Hopkins University

Email: ingo@jhu.edu

http://biosun01.biostat.jhsph.edu/∼iruczins

With Charles Kooperberg and Michael LeBlanc, FHCRC

Introduction and Motivation

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

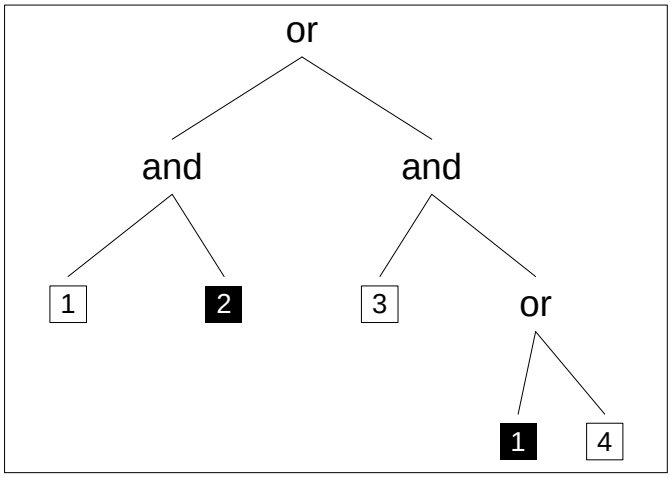
Y is a response variable.

Fit a model $g(F(Y)) = \beta_0 + \sum_{j=1}^{t} \beta_j \times L_j,$ 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

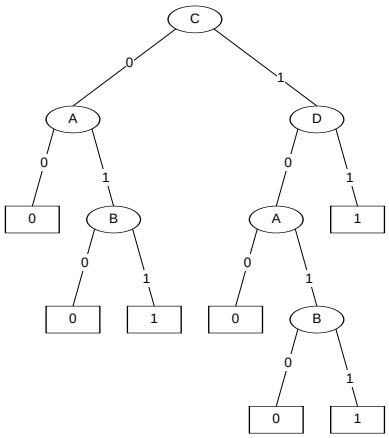
An equivalent representation of $(X_1 \wedge X_2^c) \vee (X_3 \wedge (X_1^c \vee X_4))$ is the following:



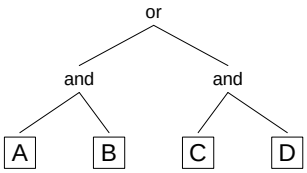
This is a Logic Tree!

Comparison to Decision Trees

Decision Tree



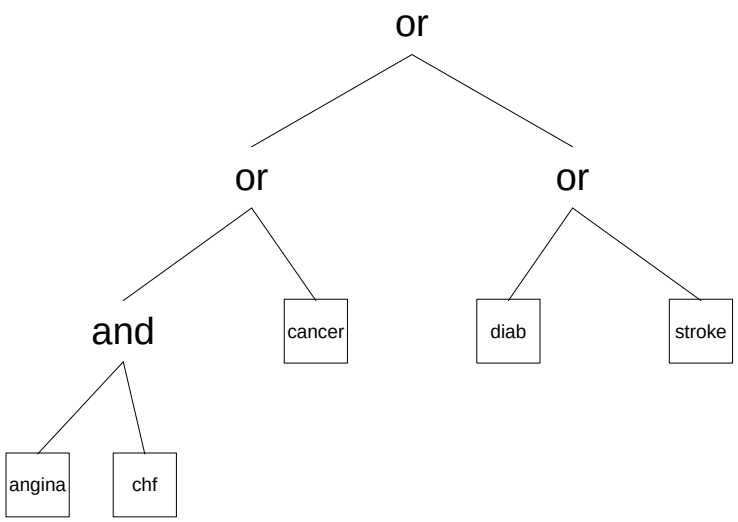
Logic Tree



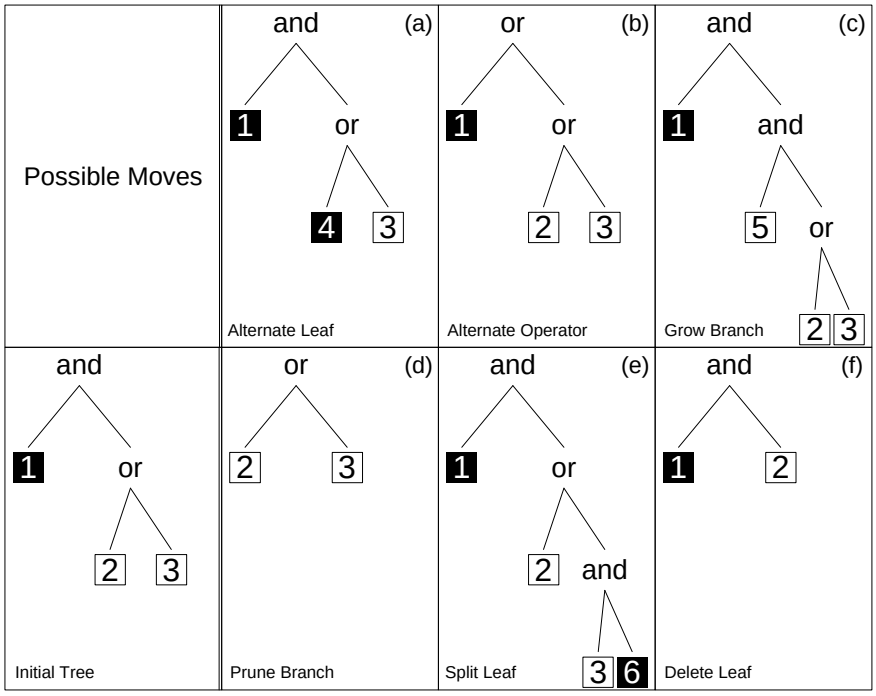
A Decision Tree (CART) is something different!

An Example

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



The Move Set



Simulated Annealing for Logic Regression

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

- 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.

Another Example

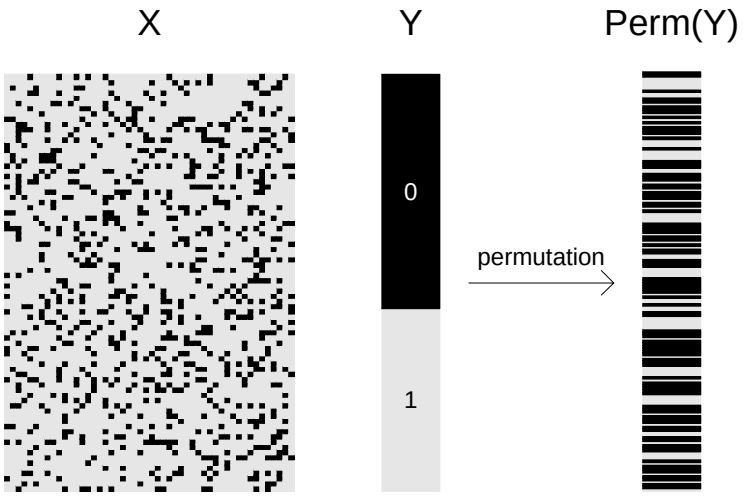
The Cardiovascular Health Study

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

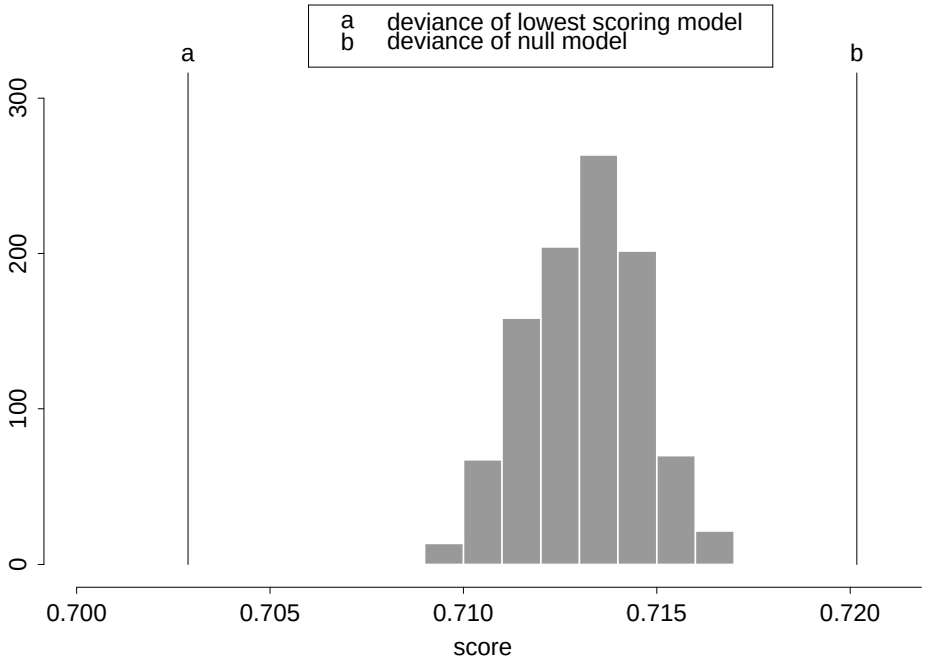
- The Cardiovascular Health Study (CHS) is a study of coronary heart disease and stroke in elderly people.
- Between 1989 and 1993, 5888 subjects over the age of 65 were recruited in four communities in the United States.
- During 1992 and 1994, a subset of these patients underwent an MRI scan.
- For 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 $Y = \beta_0 + \beta_1 \times L_1 + \cdots + \beta_p \times L_p + \epsilon$.

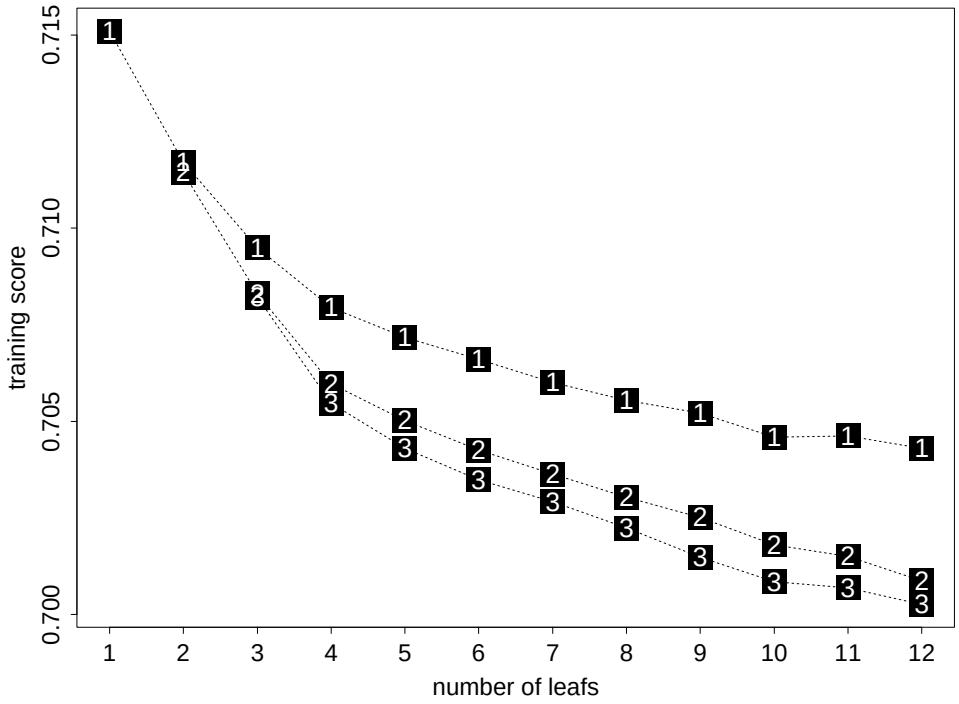
A Global Randomization Test of Association



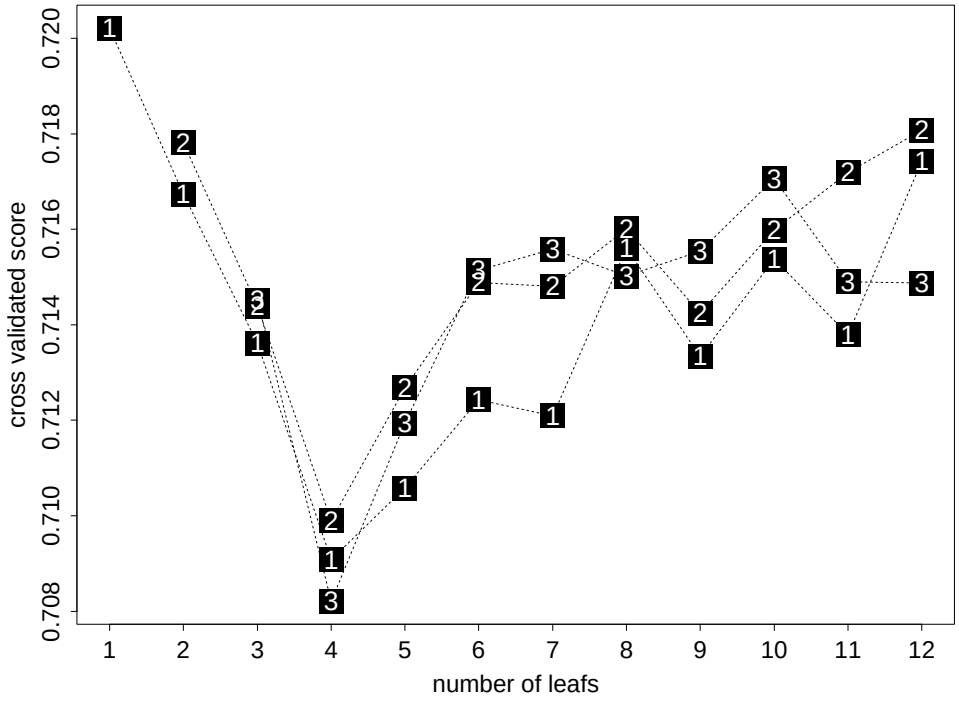
Example Cont.



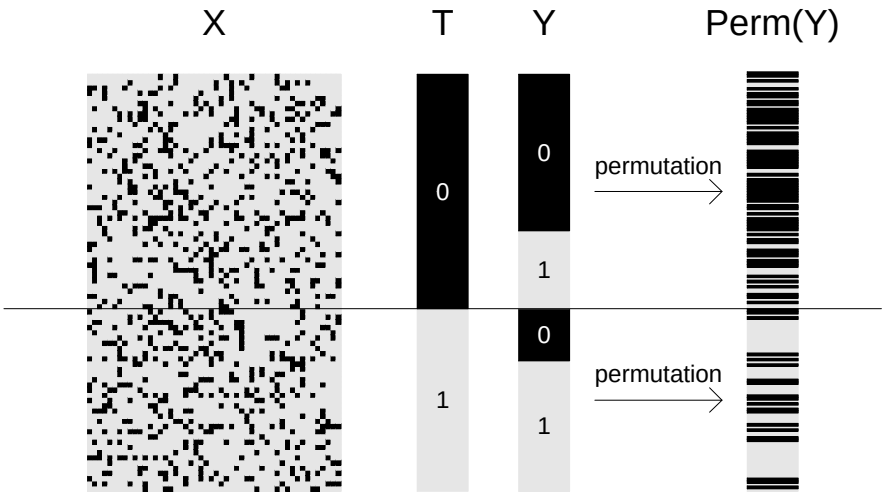
Cross-Validation



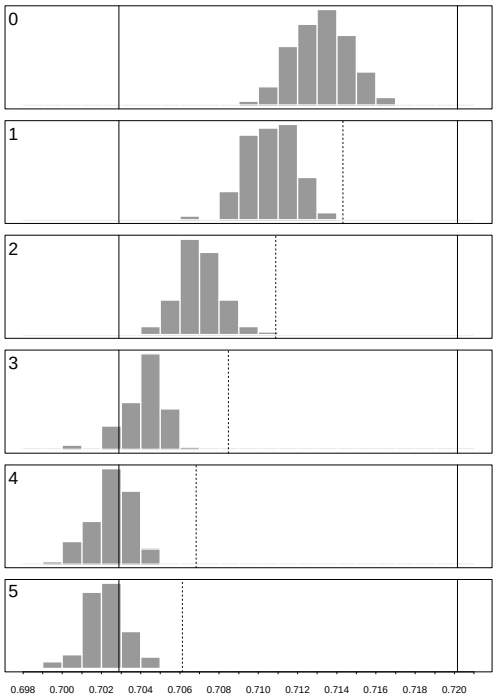
Cross-Validation Cont.



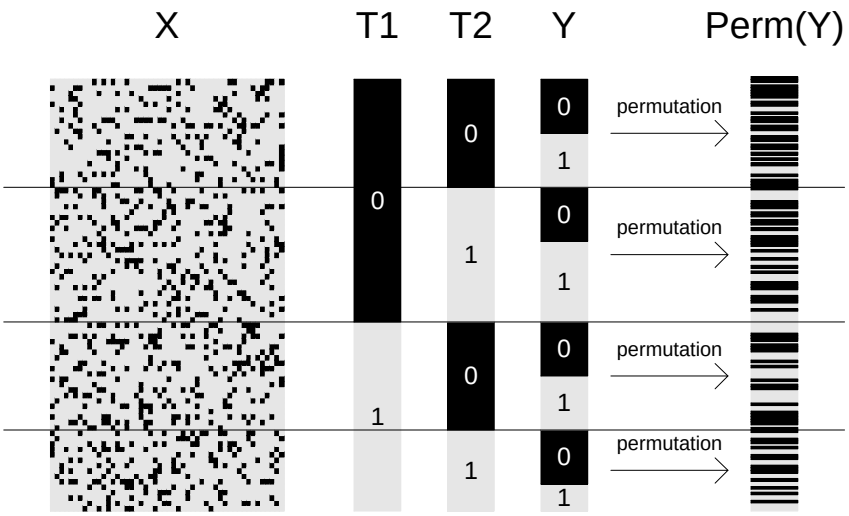
A Sequential Randomization Test for Model Size



Randomization Test Cont.

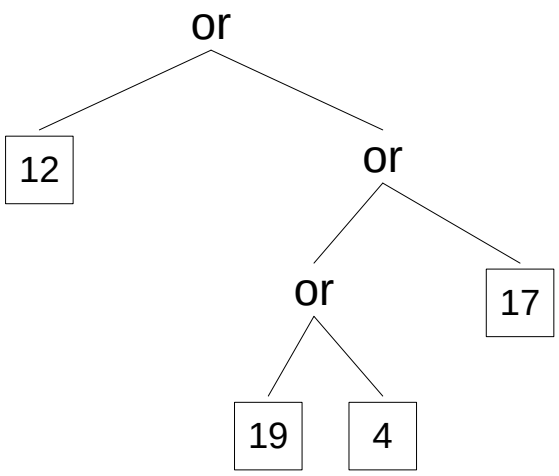


Sequential Randomization Test for 2 Trees:

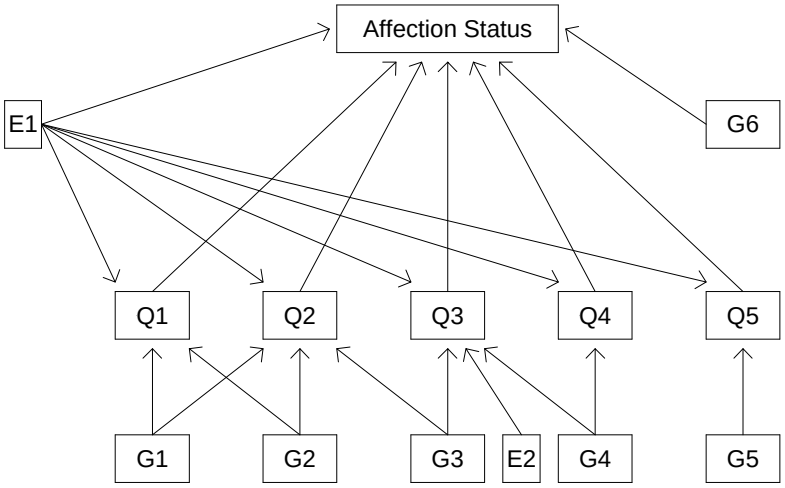


Results

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



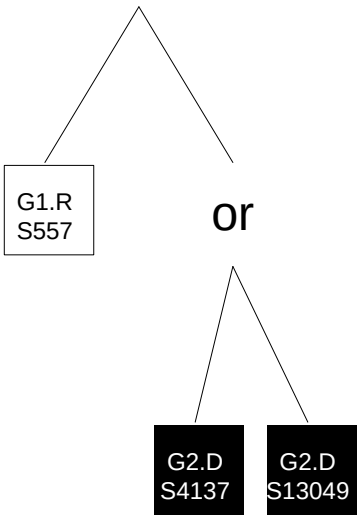
Genetic Analysis Workshop GAW 12



GAW12 Cont.

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

L_1= and



L_2= G6.D S5007

L_3= or