Logic Regression

Ingo Ruczinski

Department of Biostatistics, Johns Hopkins University

Email: ingo@jhu.edu

The slides and software used for this presentation are available at http://biostat.jhsph.edu/~iruczins

The Women's Health and Aging Study

[With Karen Bandeen-Roche]

The Women's Health and Aging Study (WHAS) began in 1992 to study the causes and the course of disability in moderately to severely disabled older women living in the community.

The WHAS is a population-based longitudinal study of women with at least mild disability, 65 years of age or older, living at home in eastern Baltimore city or county.

There is evidence that disability results from chronic diseases, and that interactions between diseases (comorbidities) are of importance in causing disability.

In this presentation we are concerned about relating chronic diseases and their interactions to death.

The Women's Health and Aging Study

Study subjects:

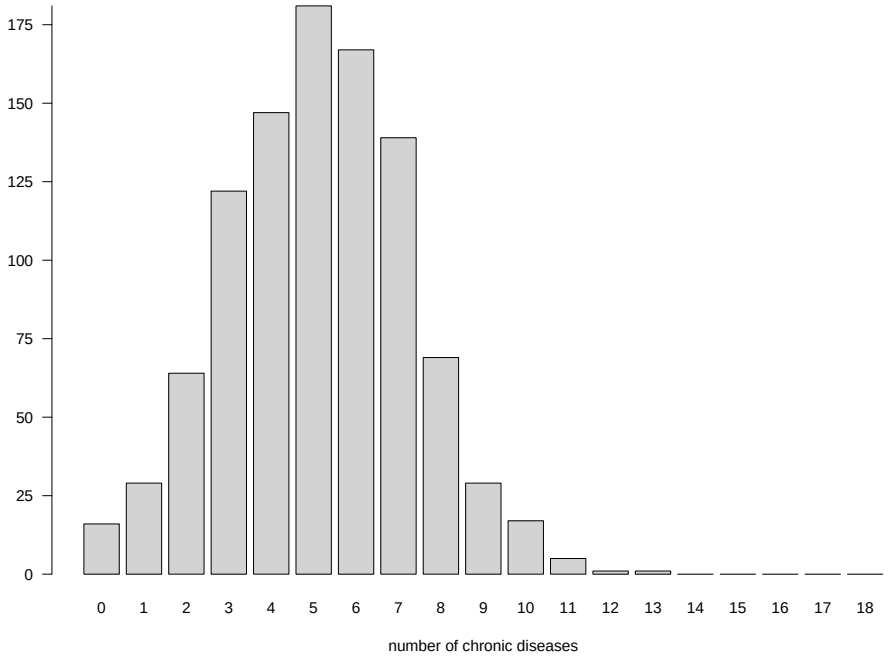
- 32538 women were identified by searching medicare enrollment files,
- 6521 women were sampled (age-stratified),
- 5316 women were alive and living at home,
- 4137 women participated in the home-based screening,
- 1409 women were eligible,
- 1002 women agreed to participate and provided written informed consent.

The major chronic diseases at baseline were ascertained by using complex algorithms. Follow-up evaluations were conducted every 6 months for 3 years.

The Women's Health and Aging Study

angina	heart pain
cancer	cancer
chf	congestion heart failure
diabetes	diabetes
disc	degenerative disc disease
hf	hip fracture
mi	myocardial infarction
oatot	osteo-arthritis at hand, knee or hip
oahand	osteo-arthritis at hand
oahip	osteo-arthritis at knee
oaknee	osteo-arthritis at hip
osteo	osteoporosis
pad	peripheral arterial disease
parkin	parkinson's disease
pulmonary	pulmonary disease
ra	rheumatoid arthritis
stenosis	spical stenosis
stroke	stroke

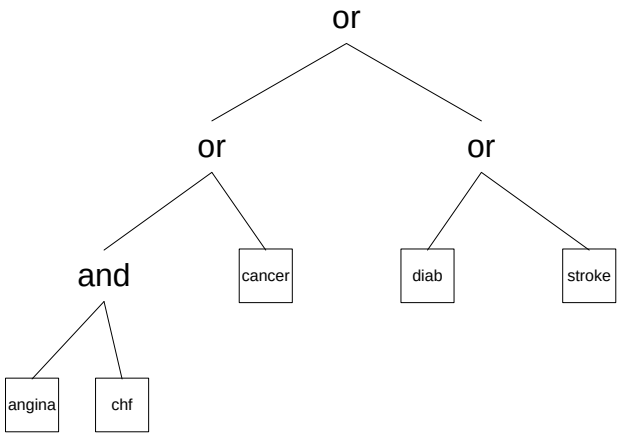
The Women's Health and Aging Study



The Women's Health and Aging Study

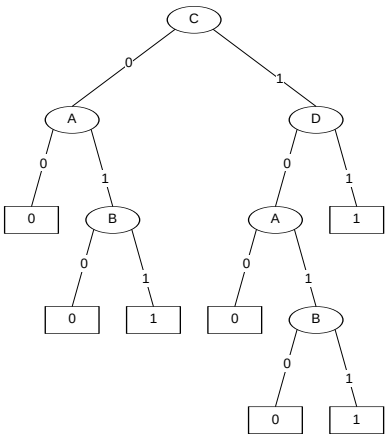
p = Pr(death in round j | survival to round j-1, X, age)

logit(p) = -9.01 + 0.06·age + 1.07·L(X)

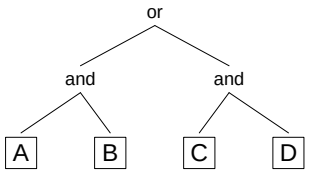


Comparison to Decision Trees

Decision Tree



Logic Tree



A Decision Tree (CART) is something different!

Logic Regression

[With Charles Kooperberg and Michael LeBlanc]

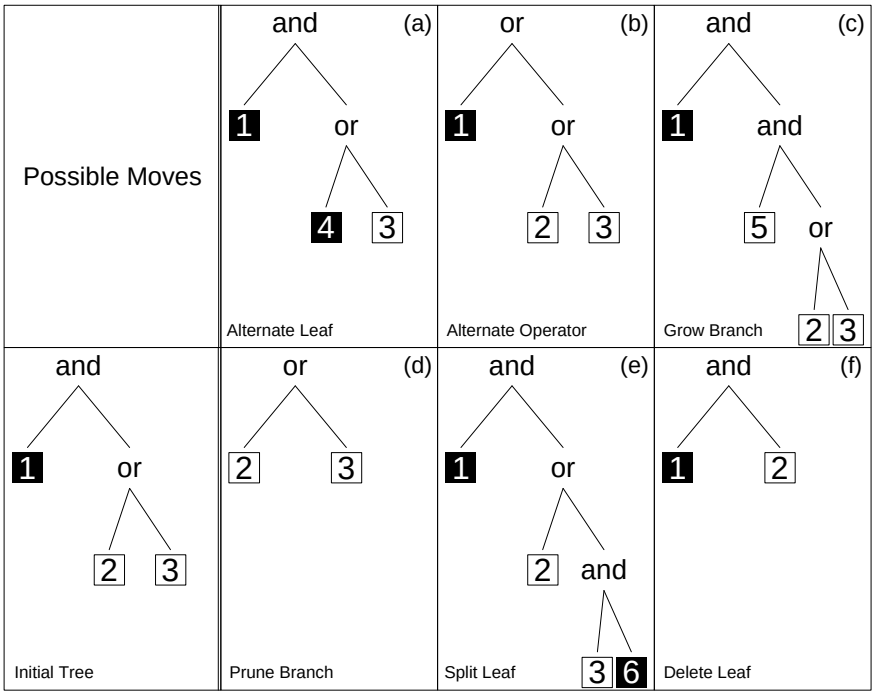
$X_1, \ldots, X_k$ are 0/1 (False/True) predictors.

Y is a response variable.

Fit a model $g(E(Y)) = b_0 + \sum_{j=1}^{t} b_j \cdot L_j$, 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 b_j simultaneously.

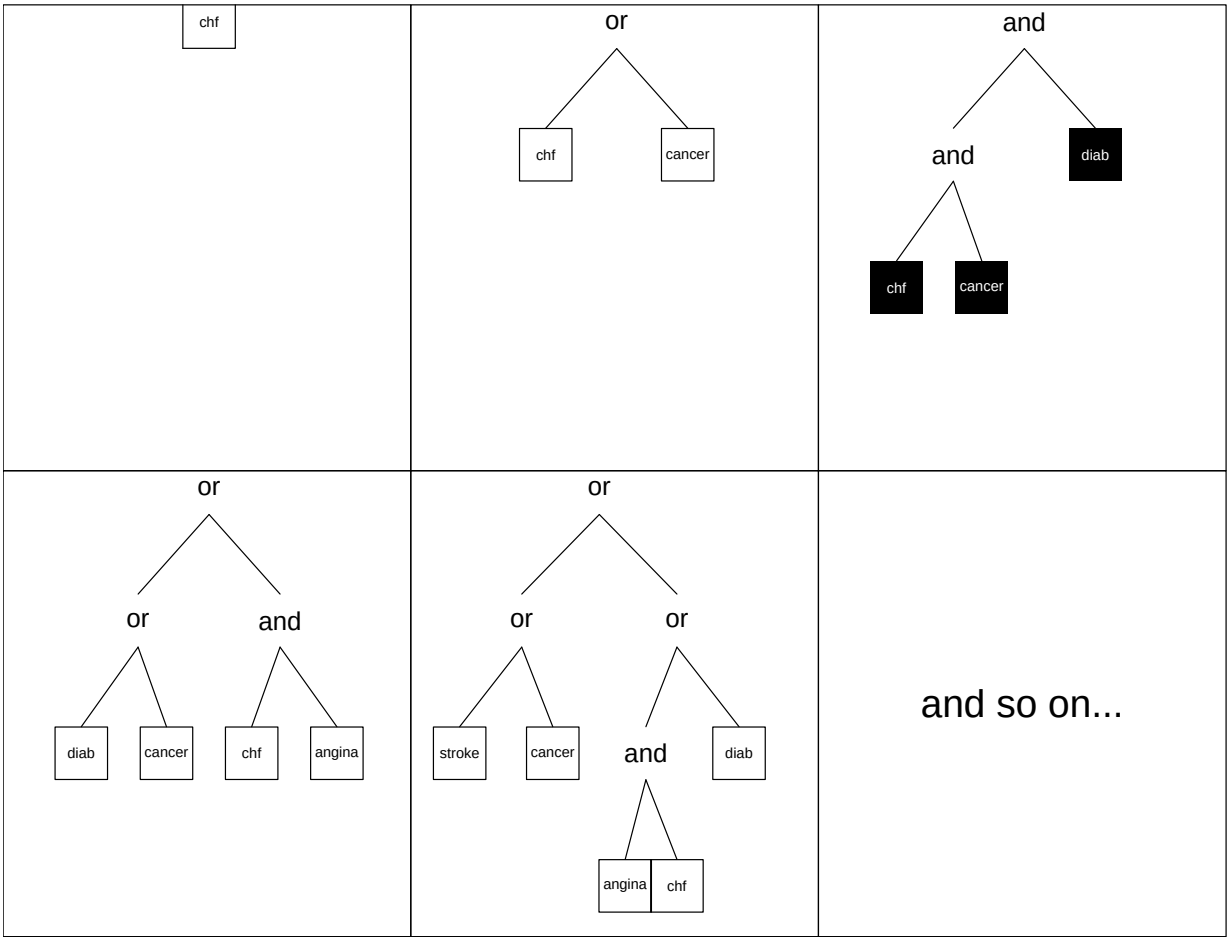
The Move Set for Logic Regression

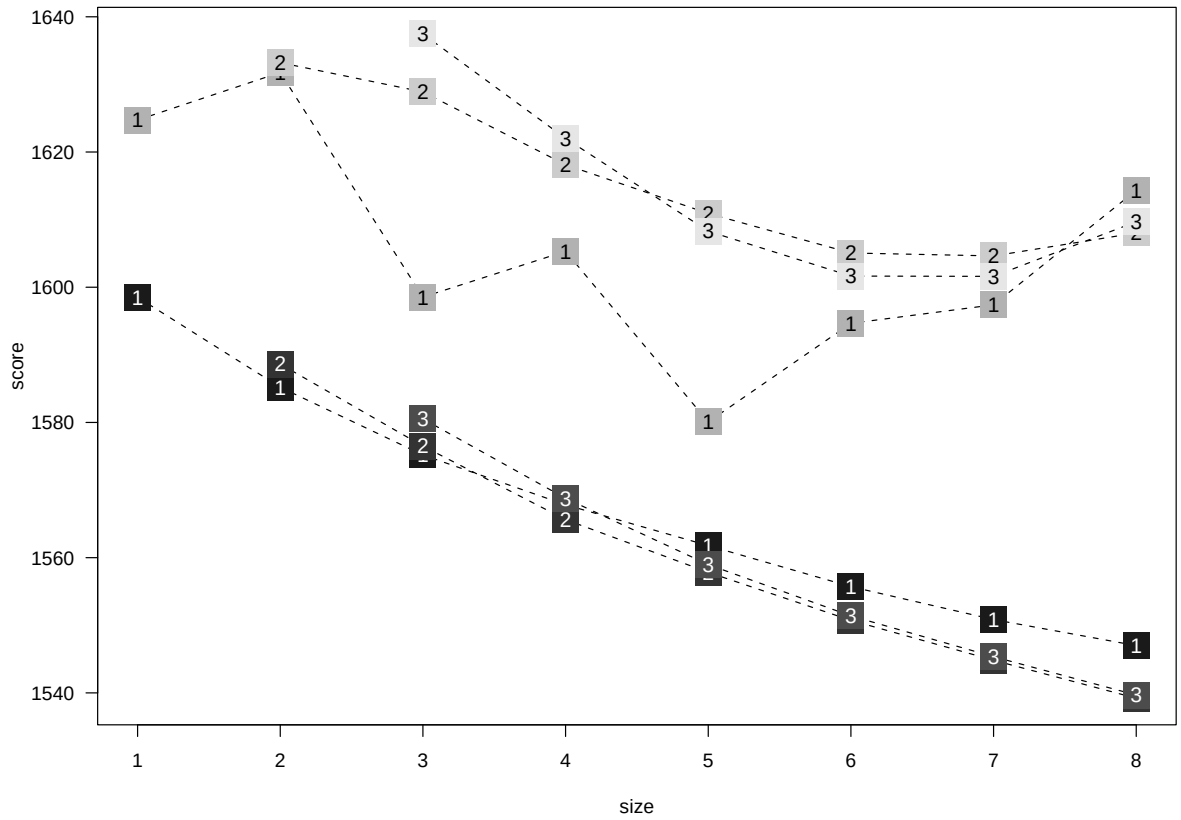


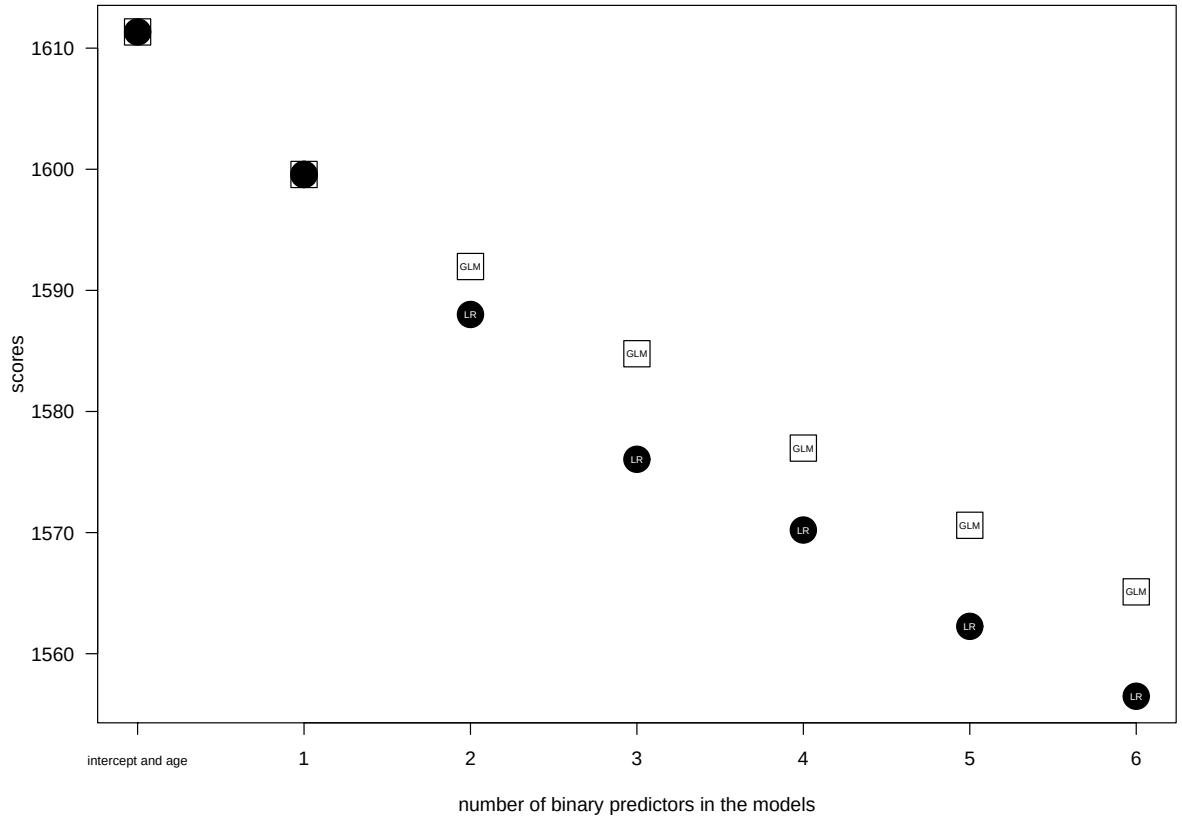
Simulated Annealing for Logic Regression

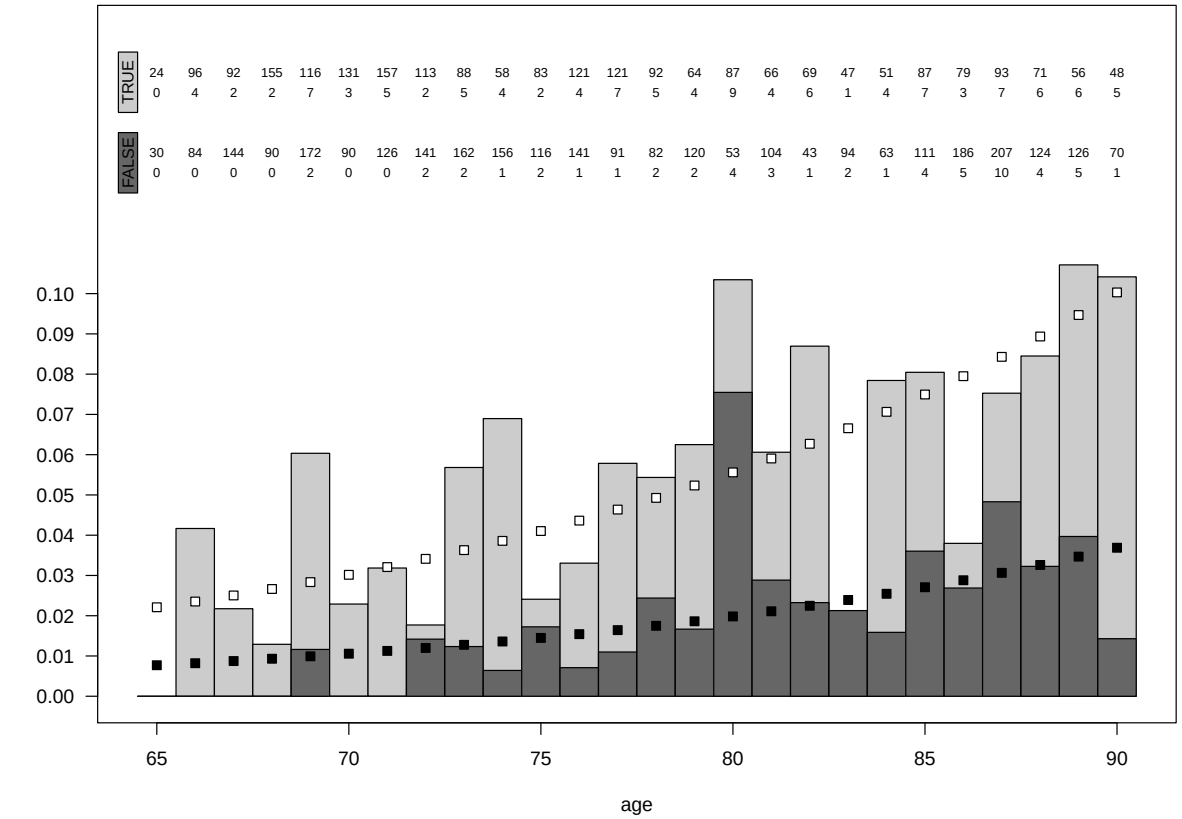
We try to fit the model $g(E(Y)) = b_0 + \sum_{j=1}^{t} b_j \cdot L_j$.

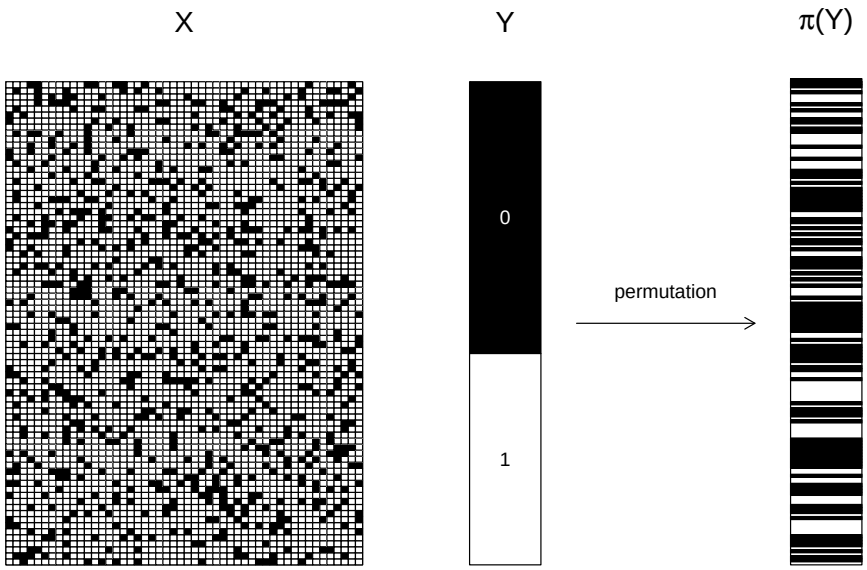
- Select a scoring function (RSS, log-likelihood, . . .).
- Pick the maximum number of Logic Trees.
- Pick the maximum number of leaves in a tree.
- Initialize the model with $L_j = 0$ for all j.
- Carry out the Simulated Annealing Algorithm:
  - Propose a move.
  - Accept or reject the move, depending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

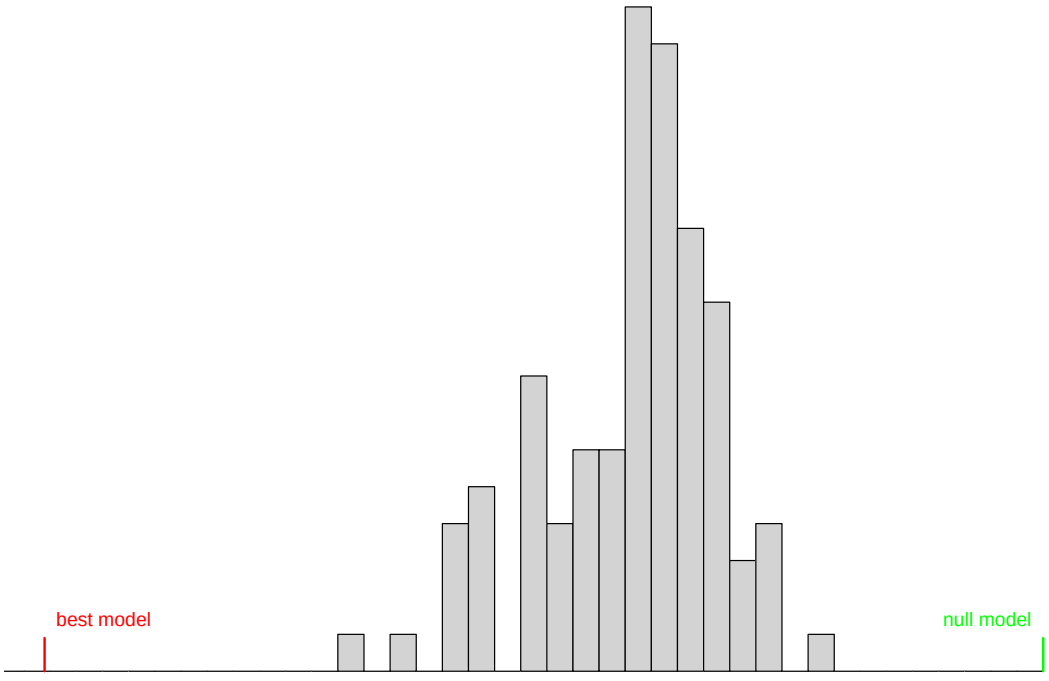

chf
or
chf
cancer
and
and
diab
chf
cancer
or
or
and
diab
cancer
chf
angina
or
or
or
stroke
cancer
and
diab
angina
chf
and so on...

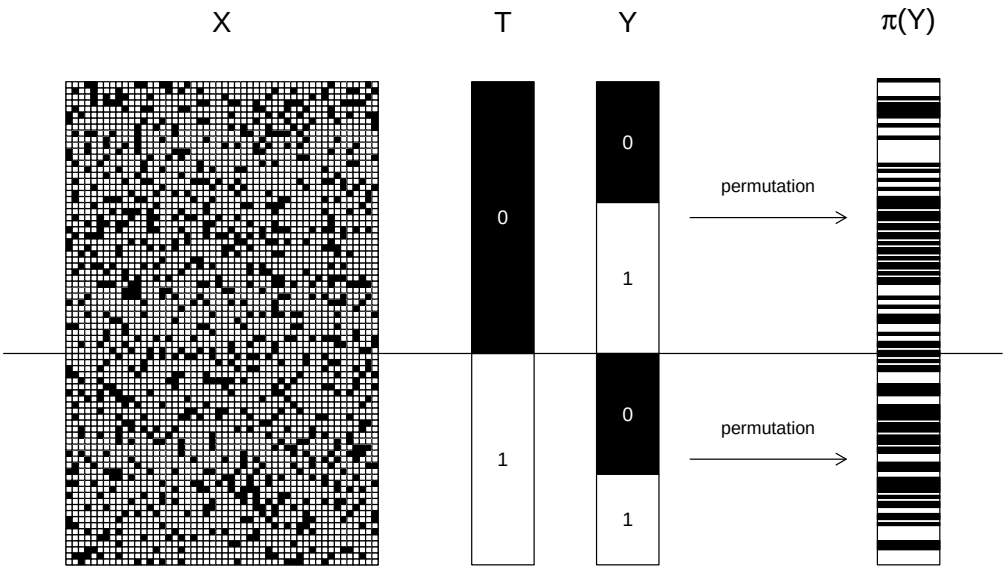

score
size
1640
1620
1600
1580
1560
1540
1
2
3
4
5
6
7
8

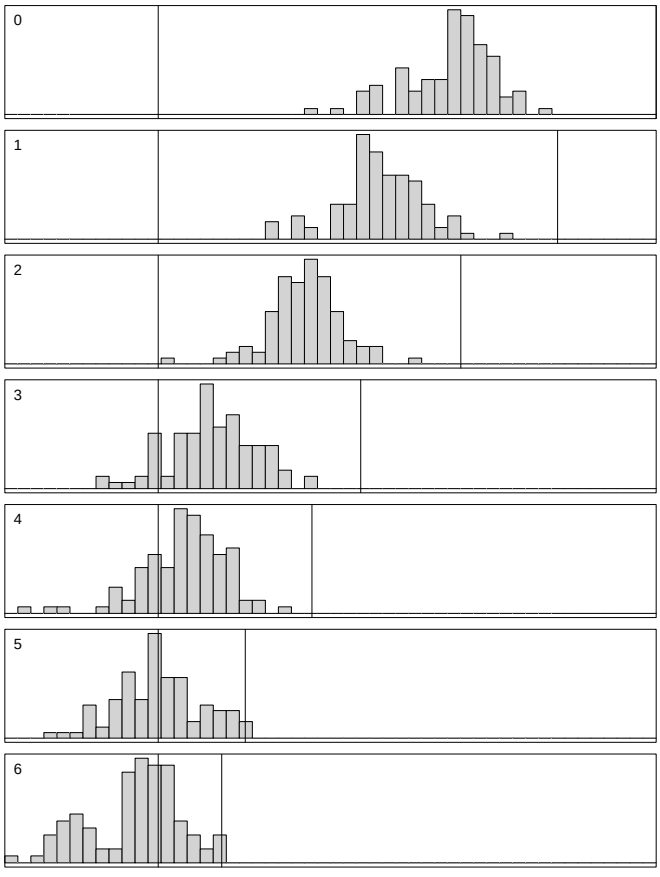

scores
1610
1600
1590
1580
1570
1560
GLM
LR
intercept and age
1
2
3
4
5
6
number of binary predictors in the models


TRUE
24 96 92 155 116 131 157 113 88 58 83 121 121 92 64 87 66 69 47 51 87 79 93 71 56 48
0 4 2 2 7 3 5 2 5 4 2 4 7 5 4 9 4 6 1 4 7 3 7 6 6 5
FALSE
30 84 144 90 172 90 126 141 162 156 116 141 91 82 120 53 104 43 94 63 111 186 207 124 126 70
0 0 0 0 2 0 0 2 2 1 2 1 1 2 2 4 3 1 2 1 4 5 10 4 5 1
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
65
70
75
80
85
90
age


X
Y
π(Y)
0
1
permutation


best model
null model


X
T
Y
π(Y)
0
1
0
1
permutation
0
1
permutation


0
1
2
3
4
5
6

Another Example

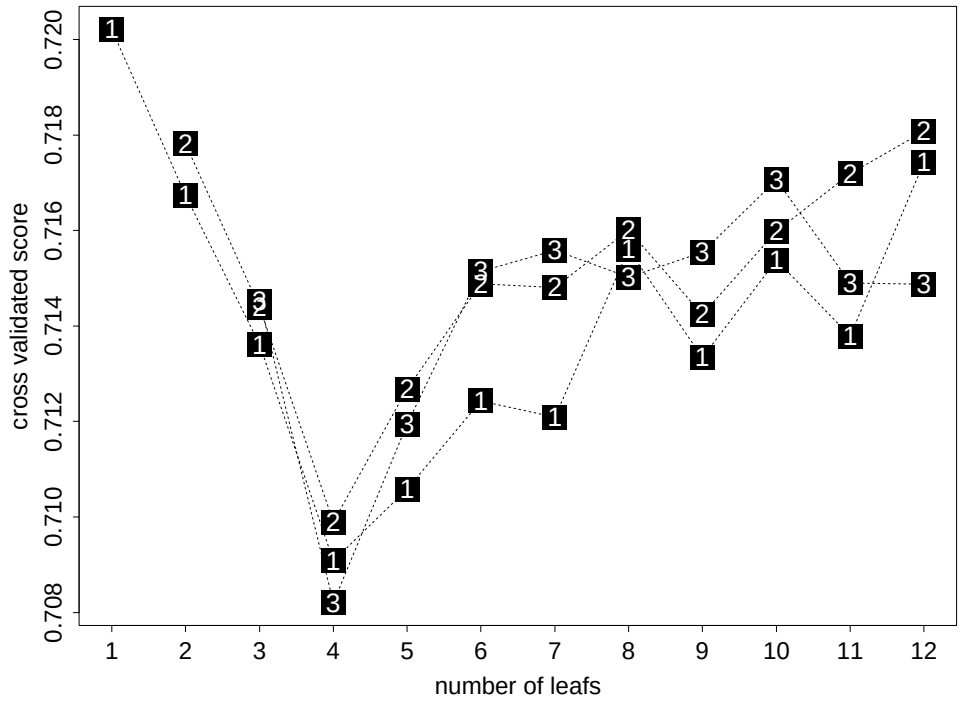
The Cardiovascular Health Study

(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.

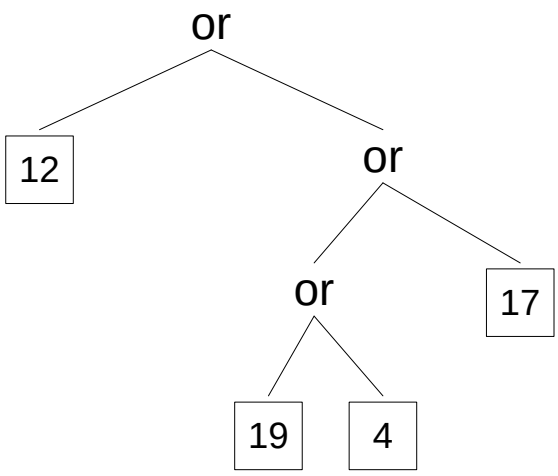
We investigated models of the form $Y = \beta_0 + \beta_1 \times L_1 + \cdots + \beta_p \times L_p + \epsilon$.

Cross-Validation Cont.



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:

	$\hat{\beta}$	s. e.	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:

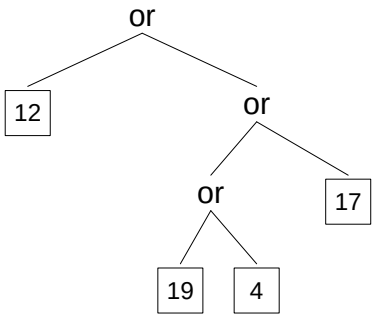
$$Y = 1.96 + 0.36 \times I_{\{X_4 \vee X_{12} \vee X_{17} \vee X_{19} \text{ is true}\}}$$

MARS:

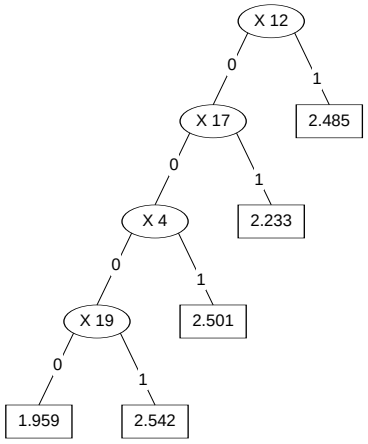
$$Y = 1.96 + 0.53\, X_4 + 0.37\, X_{12} + 0.24\, X_{17} + 0.61\, X_{19} + 1.05\, (X_{12} * X_{15})$$

Logic Model vs CART

Logic Tree



CART Tree



References

- Fried L. P., Bandeen-Roche K., Kasper J. D., and Guralnik, J. M. (1999). *Association of Comorbidity with Disability in Older Women: The Women's Health and Aging Study*. Journal of Clinical Epidemiology, 52 (1), 27-37.
- 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. (2003). *Logic Regression*. Journal of Computational and Graphical Statistics, 12 (3), 475-511.
- Ruczinski, I., Kooperberg, C., and LeBlanc, M. (2004). *Exploring Interactions in High Dimensional Genomic Data: An Overview of Logic Regression, With Applications*. Journal of Multivariate Analysis, in press.

Software and manuscripts available at: http://biostat.jhsph.edu/∼iruczins/