Logic Regression

Ingo Ruczinski

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

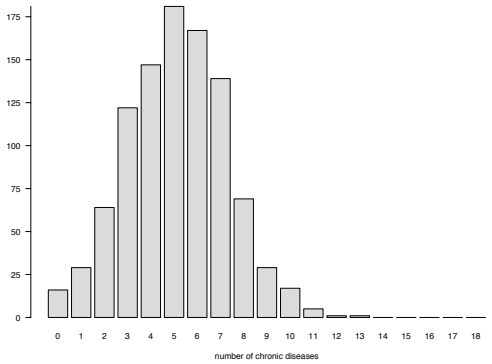
Email: ingo@jhu.edu. The slides and software used for this presentation are available at http://biostat.jhsph.edu/~iruczins

Example: The WHAS

- 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.
- 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.
- There is evidence that disability results from chronic diseases, and that interactions between diseases (comorbidities) are of importance in causing disability.
- The chronic diseases recorder included cancer, congestion heart failure, diabetes, degenerative disc disease, hip fracture, myocardial infarction, arthritis, osteoporosis, Parkinson's disease, pulmonary disease, stroke.

Reference: Fried LP et al (1999): *Association of Comorbidity with Disability in Older Women: The Women's Health and Aging Study*, Journal of Clinical Epidemiology, 52 (1), pp 27-37.

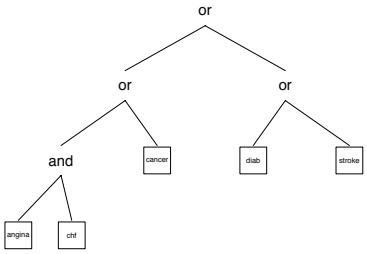
Example: The WHAS



Example: The WHAS

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

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



Motivation

[Lucek and Ott]

"Current methods for analyzing complex traits include analyzing and localizing disease loci one at a time. However, complex traits can be caused by the interaction of many loci, each with varying effect."

"... patterns of interactions between several loci, for example, disease phenotype caused by locus A and locus B, or A but not B, or A and (B or C), clearly make identification of the involved loci more difficult. While the simultaneous analysis of every single two-way pair of markers can be feasible, it becomes overwhelmingly computationally burdensome to analyze all 3-way, 4-way to N-way 'and' patterns, 'or' patterns, and combinations of loci."

Logic Regression

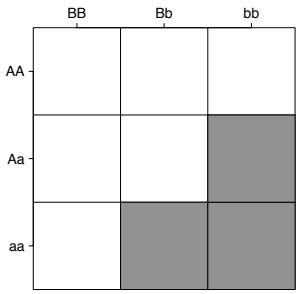
- $X_1, \ldots, X_k$ are 0/1 (False/True) predictors.
- Y is a response variable.
- Fit a model

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

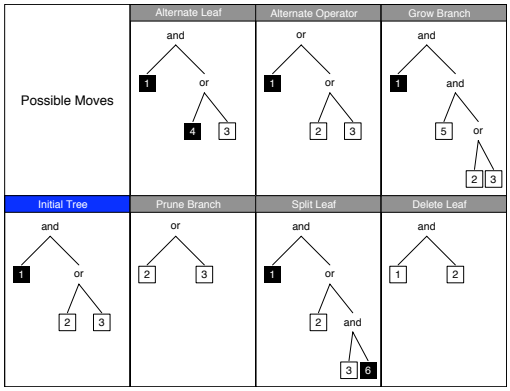
  where L_j is a Boolean combination of the covariates, e.g. $L_j = (X_1 \vee X_2) \wedge X_4^c$.
- Determine the logic terms L_j and estimate the b_j simultaneously.
- SNPs are usually coded as dominant and recessive:

SNP X	X.R	X.D
AA	0	0
AT	0	1
TT	1	1

Double Penetrance Models



The Move Set for Logic Regression



Simulated Annealing for Logic Regression

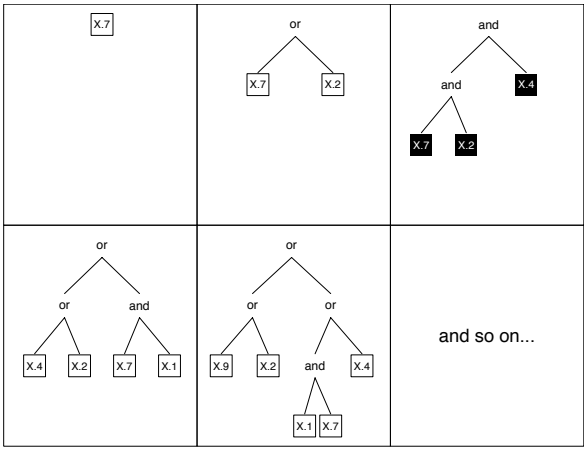
We try to fit the model $g(E(Y)) = b_0 + \sum_{j=1}^{t} b_j \cdot L_j$.

- 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:

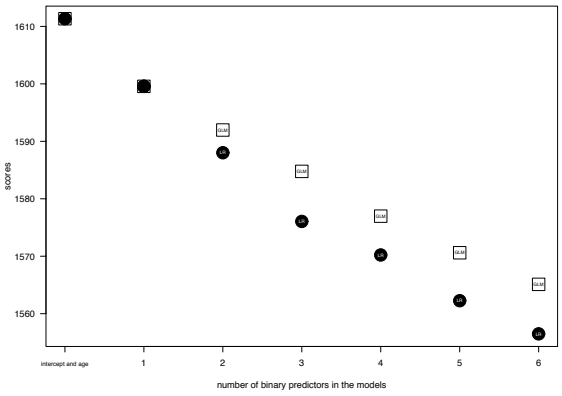
  → Propose a move.

  → Accept or reject the move, depending on the scores and the temperature.

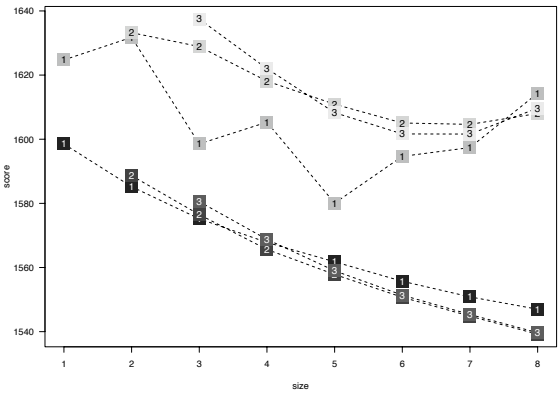
Growing Logic Models



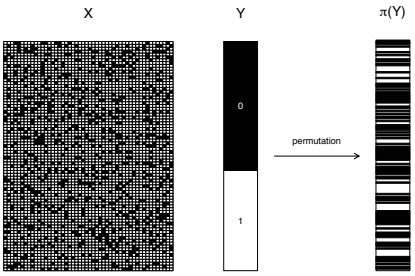
A Public Health Related Example



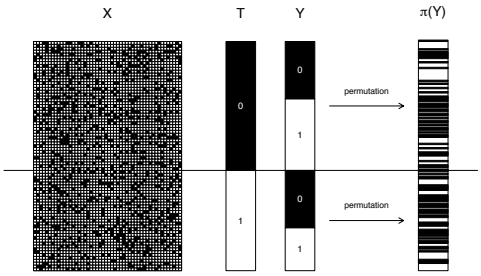
Model Selection 1 : Cross Validation



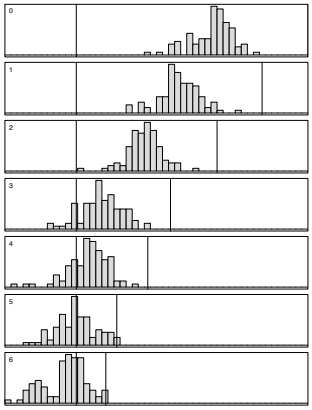
Model Selection 2 : Permutation Tests



Model Selection 2 : Permutation Tests



Model Selection 2 : Permutation Tests



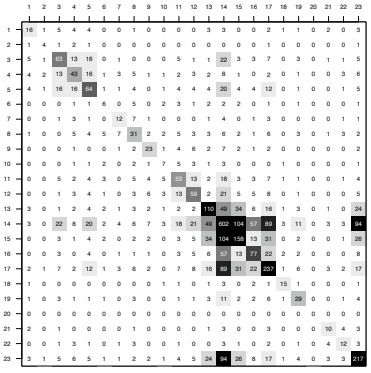
Example: The CHS

- The Cardiovascular Health Study 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.

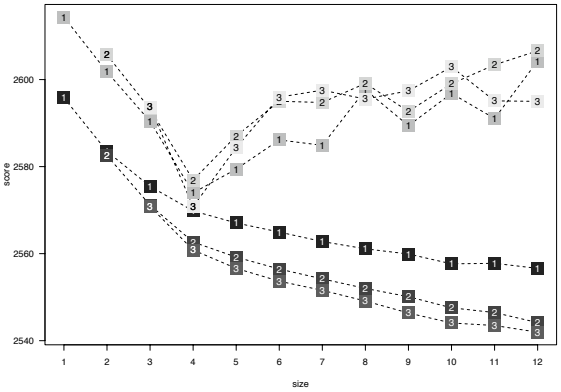
We investigated models of the form $Y = \beta_0 + \beta_1 \times L_1 + \cdots + \beta_p \times L_p + \epsilon$.

Reference: Fried LP et al (1991): *The Cardiovascular Health Study: Design and Rationale*, Annals of Epidemiology 3, pp 263-276.

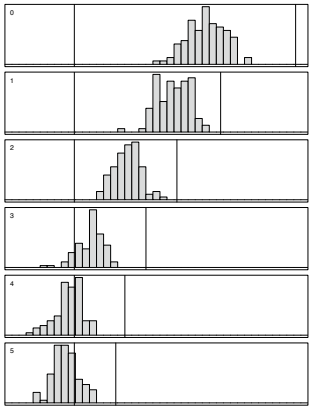
Example: The CHS



Example: The CHS

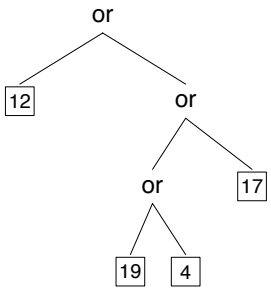


Example: The CHS



Example: The CHS

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



Example: The CHS

Linear model:

	$\hat{\beta}$	se($\hat{\beta}$)	t-value
Intercept	1.96	0.02	133.98
Region 4	0.52	0.13	4.06
Region 12	0.46	0.11	4.09
Region 17	0.24	0.06	4.17
Region 19	0.61	0.16	3.89

Logic model:

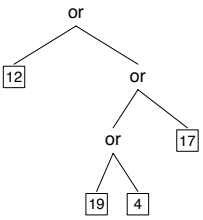
$$Y = 1.96 + 0.36 \times I_{\{X_4 \vee X_{12} \vee X_{17} \vee X_{19} \text{ is true}\}}$$

MARS:

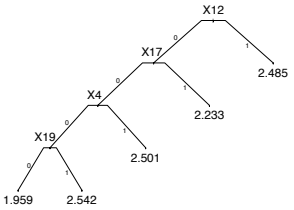
$$Y = 1.96 + 0.53\, X_4 + 0.37\, X_{12} + 0.24\, X_{17} + 0.61\, X_{19} + 1.05\, (X_{12} * X_{15})$$

Example: The CHS

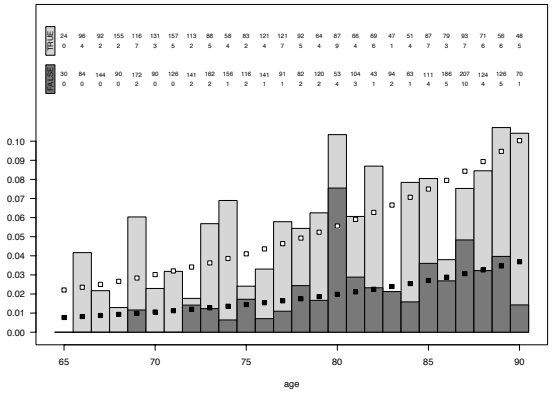
Logic Tree



CART Tree



Example: The WHAS



Logic Regression

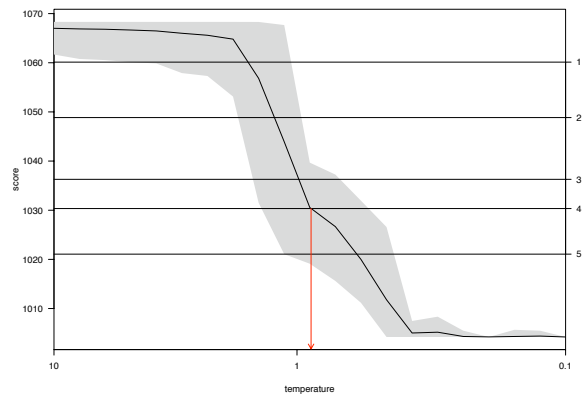
```
> install.packages("LogicReg")
> library(LogicReg)

> myanneal=logreg.anneal.control(start=-1,end=-4,
                                 iter=100000,update=1000)
> fit1=logreg(resp=y,bin=x,type=2,select=1,ntrees=2,
              anneal.control=myanneal)
> plot(fit1)

> myanneal2=myanneal
> myanneal2$update=0
> fit4=logreg(select=4,anneal.control=myanneal2,oldfit=fit1,nrep=100)
> plot(fit4)

> fit2=logreg(oldfit=fit1,select=2,ntrees=c(1,2),nleaves=c(1,7),
            anneal.control=myanneal2)

> fit3=logreg(select=3,oldfit=fit2)

> fit5=logreg(select=5,oldfit=fit2,nrep=100)
```

Multiple Models 1 : Monte Carlo LR

- Goal: identify all models and combinations of covariates that are potentially associated with the outcome.
- Use reversible jumps to implement an MCMC algorithm with priors on models and model size.
- The prior on model size does influence the total number of SNPs selected.
- The prior on model size has virtually no influence on the relative ordering of the SNPs or combinations thereof.

Reference: Kooperberg C, Ruczinski I. *Identifying Interacting SNPs using Monte Carlo Logic Regression*, Genet. Epidemiol, 28(2): 157-170, 2005.

Multiple Models 2 : Metropolis-Hastings



Multiple Models 2 : Metropolis-Hastings

Let γ_S be the score of a certain state S.

- We use the acceptance function

$$\alpha(\gamma_{\text{old}}, \gamma_{\text{new}}, t) = \min\{1, \exp([\gamma_{\text{old}} - \gamma_{\text{new}}]/t)\}$$

- If we keep the temperature constant, this defines a homogeneous Markov chain.
- We constructed the move set to be irreducible and aperiodic, therefore each homogeneous Markov chain has a limiting distribution $\pi_t(S)$.
- If we know the model size where the signal ends and the noise starts, we can read off the corresponding temperature from the diagnostic plot!

Multiple Models 2 : Metropolis-Hastings

Example: Simulate 10 binary predictors $X_1, \ldots, X_{10}$.

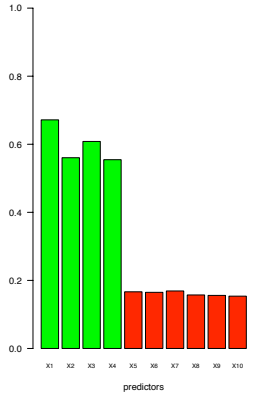
Let $Y = 5 + 1 \times L(X_1, X_2, X_3, X_4) + \epsilon, \quad \epsilon \sim \text{N}(0,1)$.

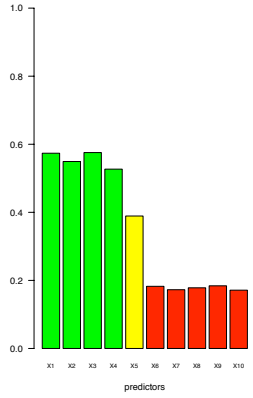
Run a homogeneous Markov chain during "crunch time" for two separate cases:

Case 1 All X are independent.

Case 2 All X are independent, except X_4 (in the signal) and X_5 (not in the signal), which are heavily correlated.

Multiple Models 2 : Metropolis-Hastings





Multiple Models 2 : Metropolis-Hastings