Logic Regression and Interactions in High Dimensional Genomic Data

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 disease (including cancer and cardiovascular disease). Particularly, gene-environment and gene-gene interactions associated with disease are of interest. Currently, SNP data from 51 genes are available.

Motivation

[With Brian Caffo, Steve Goodman, and Giovanni Parmigiani]

Associations between chromosomal deletions and stages of bladder cancer:

STAGE →	2	2	2	2	2	2	2	3	3	3	3	3	3	4	4	4	4	4	4	4	
P4	0	0	0	0	1	0	1	0	0	0	1	1	0	1	0	0	1	0	0	1	0	
Q5	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	1	1	0	
P8	1	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	1	1	0	
P9	0	1	0	1	1	1	1	1	0	1	1	1	1	1	0	1	1	0	1	1	0	
Q9	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	...
P11	0	0	0	1	0	1	0	1	0	0	1	1	1	1	0	0	1	0	1	1	0	
Q13	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	1	0	0	1	1	
Q14	0	0	0	0	0	0	0	1	0	0	0	0	1	1	1	0	1	0	0	0	0	
P17	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	1	0	1	1	0	
Q18	0	0	0	0	0	1	1	0	0	0	1	1	1	1	0	0	1	0	1	1	0	

Motivation

[With Tony Alberg]

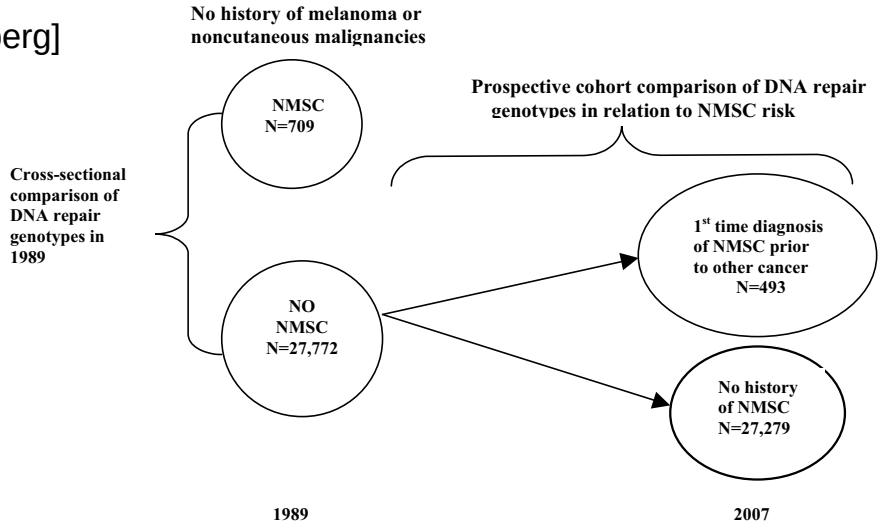


Table 1. Amino Acid Substitution Variants Identified in DNA Repair and Repair-Related Genes (Source: Mohrenweiser et al 2002, and Goode et al 2002)

Gene Name	Exon	Codon	Common Residue	Variant Residue	Allele Frequency
Base Excision Repair					
ADPRT	17	761	Val	Ala	0.18
APE1	5	148	Asp	Glu	0.33
OGG1	7	326	Ser	Cys	0.15-0.45
OGG1	Nucleotide 7143*		A	G	0.15
OGG1	Nucleotide 11657*		A	G	0.15
POLD1	1	19	Arg	His	0.12
POLD1	3	119	Arg	His	0.15
POLD1	4	173	Ser	Asn	0.05
XRCC1	6	194	Arg	Trp	0.13
XRCC1	10	399	Arg	Gln	0.24
Nucleotide Excision Repair					
ERCC2	10	312	Asp	Asn	0.4
ERCC2	23	751	Lys	Gln	0.32
ERCC4	8	415	Arg	Gln	0.06
ERCC5	15	1104	Asp	His	0.18
RAD23B	7	249	Ala	Val	0.10
XPC	8	499	Ala	Val	0.24
XPC	15	939	Lys	Gln	0.38
Double Strand Break/Recombination Repair					
NBS1	5	185	Gln	Glu	0.34
XRCC2	3	188	Arg	His	0.05
XRCC3	7	241	Thr	Met	0.43
XRCC4	5	247	Ala	Ser	0.08
Damage Recognition, Repair and Cell Cycle Check point					
CDKN2A	2	148	Ala	Thr	0.05
RAD52	8	287	Ser	Asn	0.05
MLH1	8	219	Ile	Val	0.12
Mismatch Repair					
MSH3	10	514	Glu	Lys	0.05
MSH3	21	940	Arg	Gln	0.1
MSH3	23	1036	Thr	Ala	0.3
MSH6	1	39	Gly	Glu	0.24

* Amino acid substitution variants of these SNPs have not been published. However, these nucleotide substitutions occur in a gene of particular interest (see section B.2.c.), and have been found to associate strongly with risk of prostate cancer. (Goode et al 2002)

Motivation

Lucek and Ott (1997):

"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."

Trees versus Rules

Trees:	CART (Breiman et. al. 1984), ID3 (Quinlan 1986), M5 (Quinlan 1992), C4.5 (Quinlan 1993), SLIQ (Mehta, Agrawal, and Rissanen 1996).
Rules:	AQ (Michalski et. al. 1986), CN2 (Clark and Niblett 1989), SWAP1 (Weiss and Indurkhya 1993a; Weiss and Indurkhya 1993b; Weiss and Indurkhya 1995), RIPPER, SLIPPER (Cohen 1995; Cohen and Singer 1999), R2 (Torgo 1995; Torgo and Gama 1996), GRASP (Deshpande and Triantaphyllou 1998), CWS (Domingos 1996).

Classification versus Regression

Classification:	ID3, C4.5, CN2, SLIQ, RIPPER, SLIPPER, SWAP1, and a gazillion of derivations of the former methods (Apte, Damerau, and Weiss 1994).
Regression:	SWAP1R (Weiss and Indurkhya 1993b plus an extension by Torgo and Gama 1996), R2, M5, treed models (Chipman, George, and McCulloch 2002), CUBIST (Quinlan).
Both:	CART, MARS (Friedman 1991).

Objectives for Format of Solutions

Interpretation:	Clark and Niblett (1989), Quinlan (1993), Weiss and Indurkhya (1995), Cohen (1995).
Prediction:	Gray and Michel (1992), Cheng and Titterington (1994), Thimm and Fiesler (1996), Lucek and Ott (1997), Anthony (2001).

Worth having a look at: Wnek and Michalski (1994) compare a decision tree learning method (C4.5), a rule-learning method (AQ15), a neural net trained by a back-propagation algorithm (BpNet) and a classifier system using a genetic algorithm (CFS) with respect to their predictive accuracy and simplicity of solutions.

Alternatives to Straight Greedy Searches

Probablistic searches:

Genetic Algorithms:	Vafaie and DeJong (1991), Bala, DeJong, Pachowicz (1991), Giordana and Saitta (1993), Wnek and Michalski (1994).
Simulated Annealing:	Fleisher et. al. (1985), Sutton (1991), Lutsko and Kuijpers (1994), Frean (1990).

Alternatives to Straight Greedy Searches

Statistical approaches:

Bayesian CART (Chipman et. al. 1998, Denison et. al. 1998), EM algorithm (Jordan and Jabocs 1994), bagging (Breiman 1996), bumping (Tibshirani and Knight 1999), boosting (Freund and Schapire 1996), randomized decision trees (Amit and Geman 1997; Dietterich 1999), PRIM (Friedman and Fisher 1999).

Comparisons of some of those methods were for example carried out by Quinlan (1996), Dietterich (1999), and Breiman (1999).

Logic Regression

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

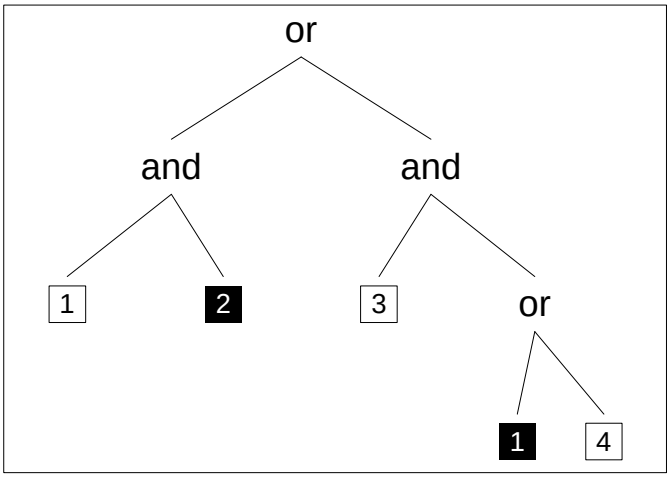
Y is a response variable.

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

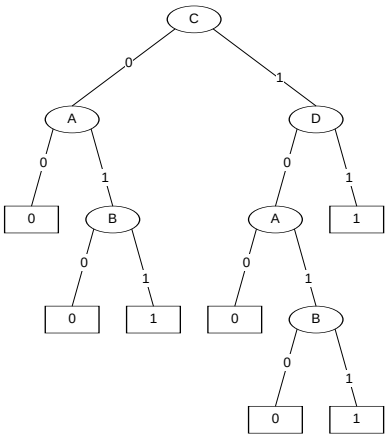
An equivalent representation of $(X_1 \wedge X_2^c) \vee (X_3 \wedge (X_1^c \vee X_4))$ is the following:



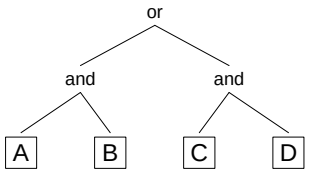
This is a Logic Tree!

Comparison to Decision Trees

Decision Tree

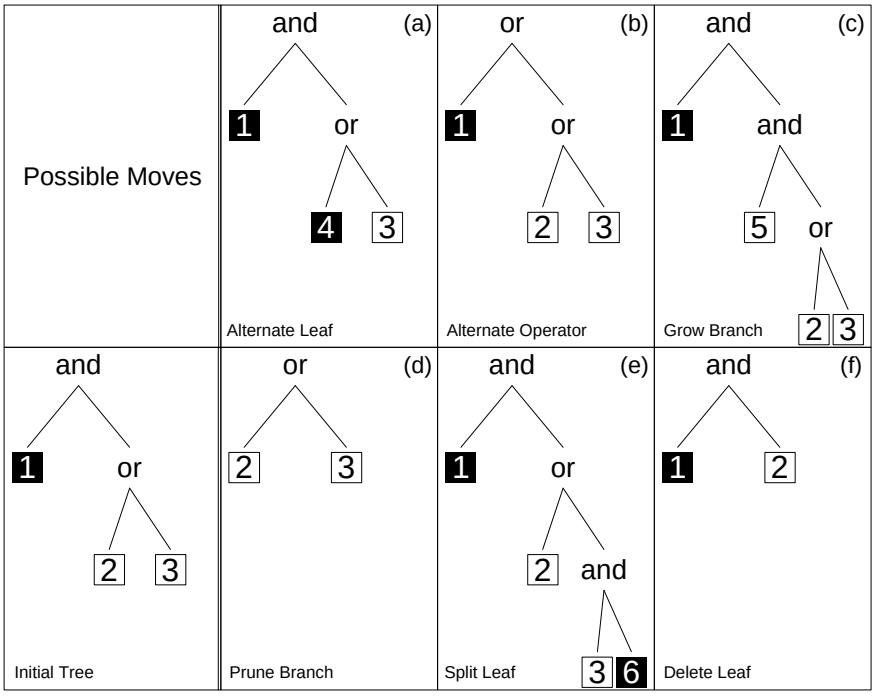


Logic Tree



A Decision Tree (CART) is something different!

The Move Set

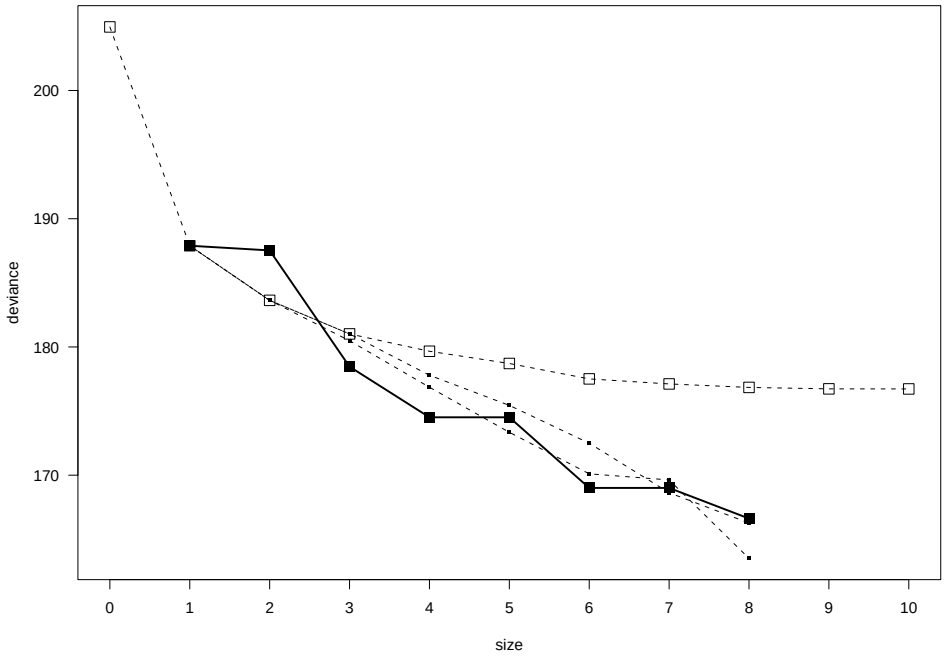


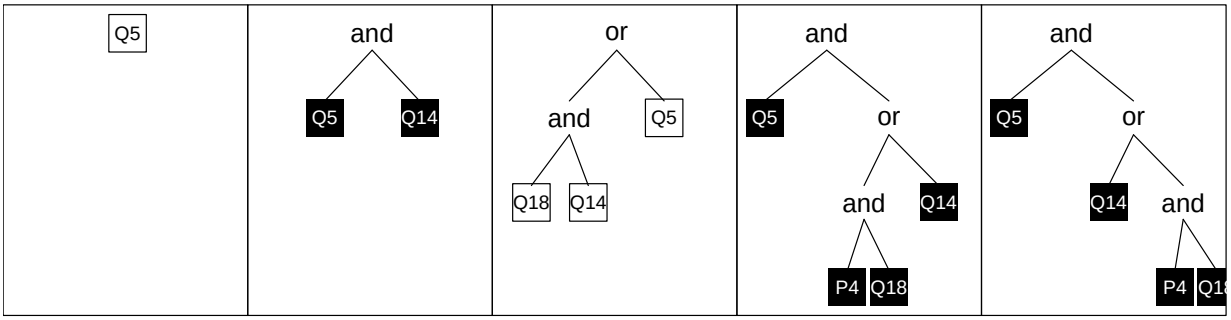
Simulated Annealing for Logic Regression

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

- Select a scoring function (RSS, log-likelihood, . . .).
- Pick the maximum number of Logic Trees.
- Pick the maximum number of 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.

Bladder Cancer Example

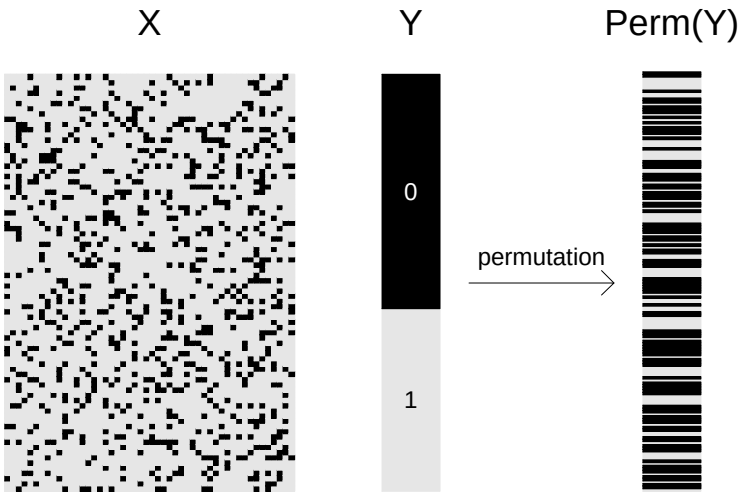




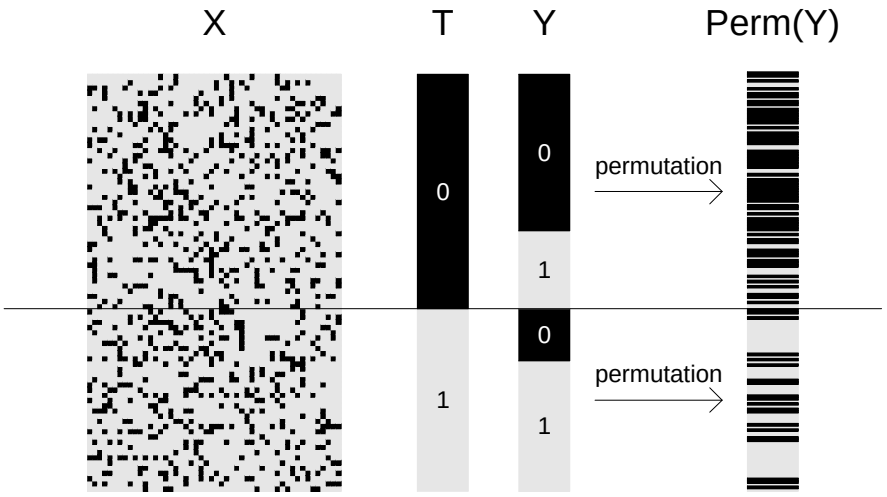
	Q18	and Q18 Q13	P4	and or Q5 Q14 and Q18 P4
	Q5	Q5	or and Q5 Q14 Q18	Q18

		Q18	and P9 P11	Q18
		Q5	Q18	Q5
		P4	Q5	and and Q9 Q14 Q13

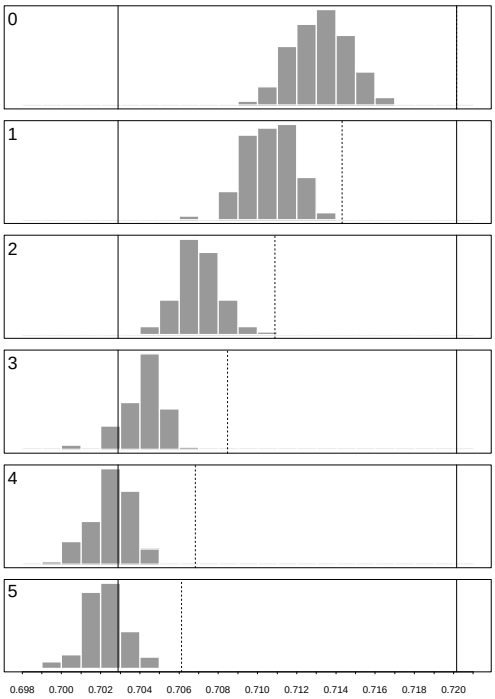
A Global Randomization Test of Association



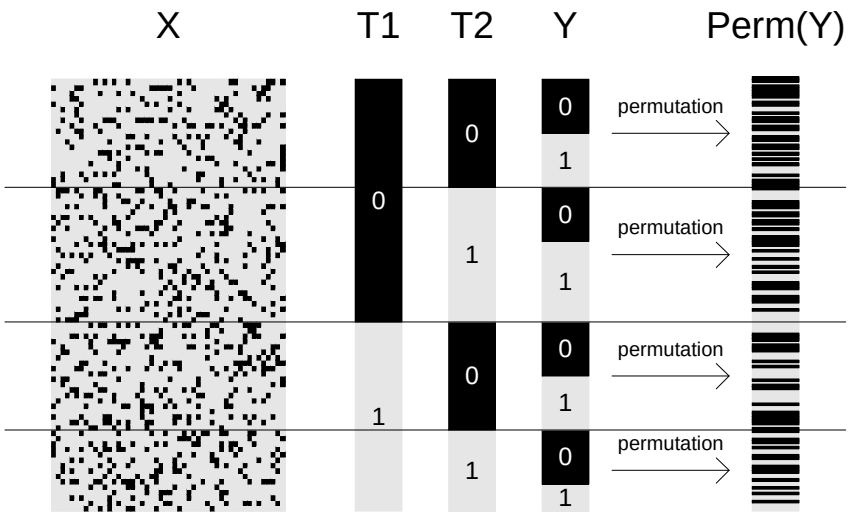
A Sequential Randomization Test for Model Size



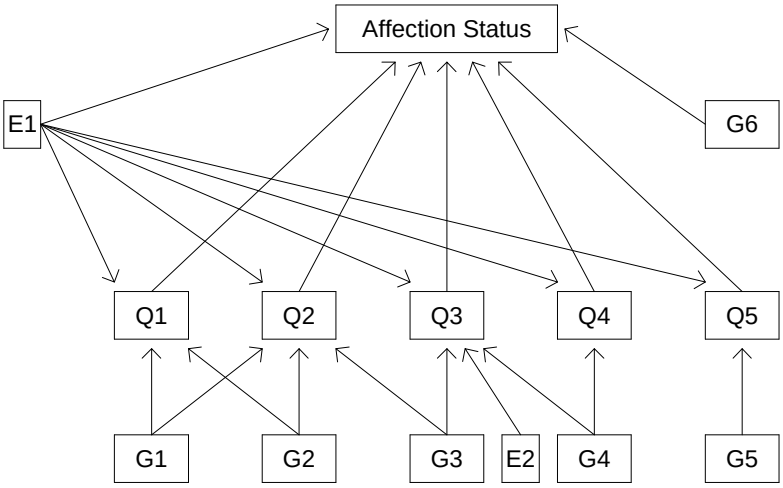
A Sequential Randomization Test for Model Size



Sequential Randomization Test for 2 Trees:



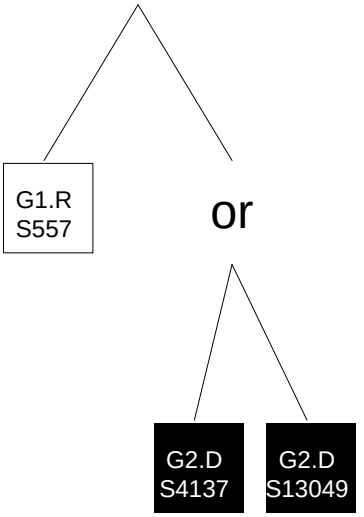
Genetic Analysis Workshop GAW 12



Genetic Analysis Workshop GAW 12

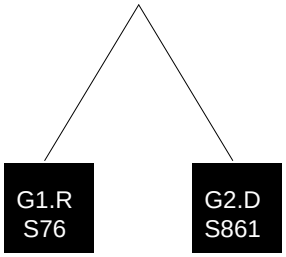
$$\text{logit(affected)} = \beta_0 + \beta_1 \times \text{ENV}_1 + \beta_2 \times \text{ENV}_2 + \beta_3 \times \text{GENDER} + \sum_{i=1}^{K} \beta_{i+3} \times L_i$$

L_1= and

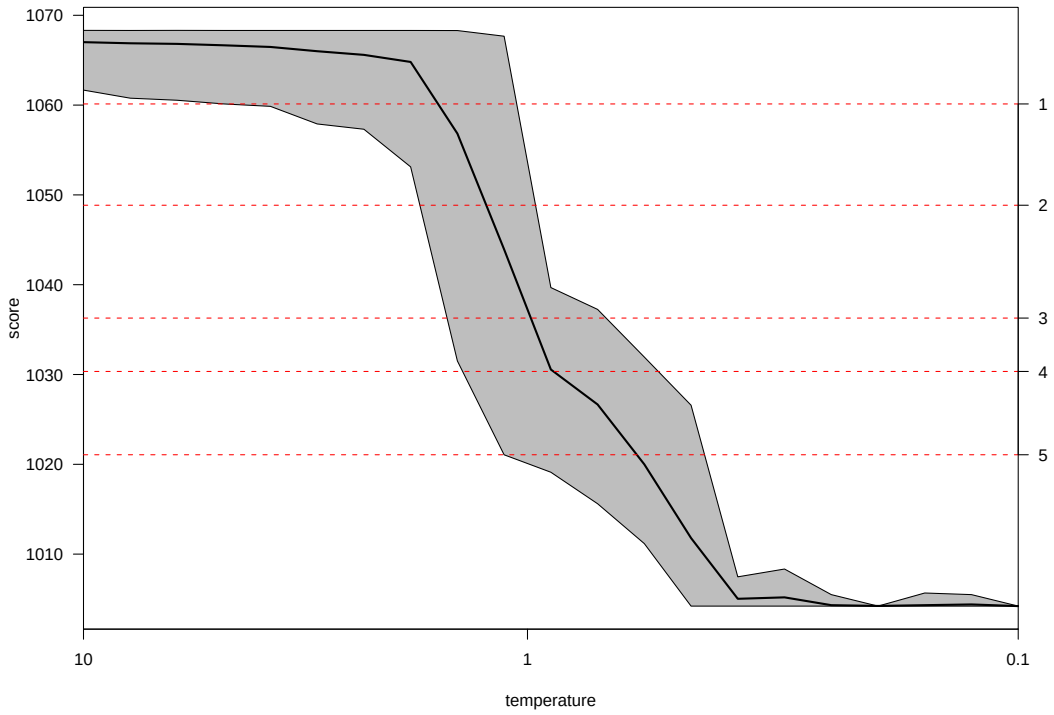


L_2= G6.D S5007

L_3= or



Multiple Models



Multiple Models

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 $\pi_t(S)$.

Multiple Models

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

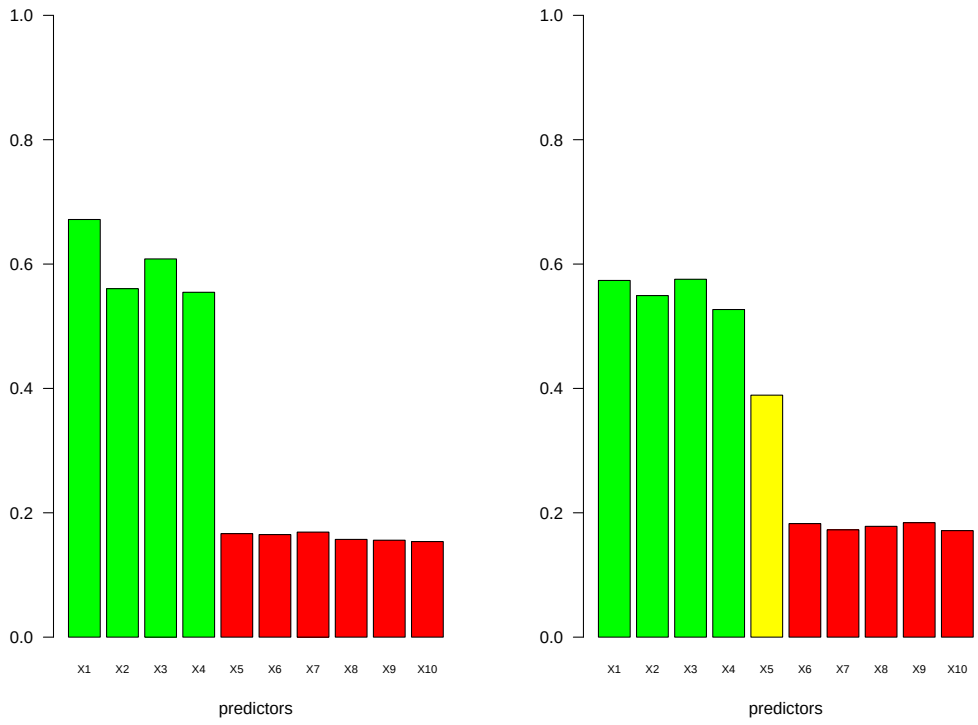
Let $Y = 5 + 1 \times L(X_1, X_2, X_3, X_4) + \epsilon, \quad \epsilon \sim \text{N}(0,1)$.

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

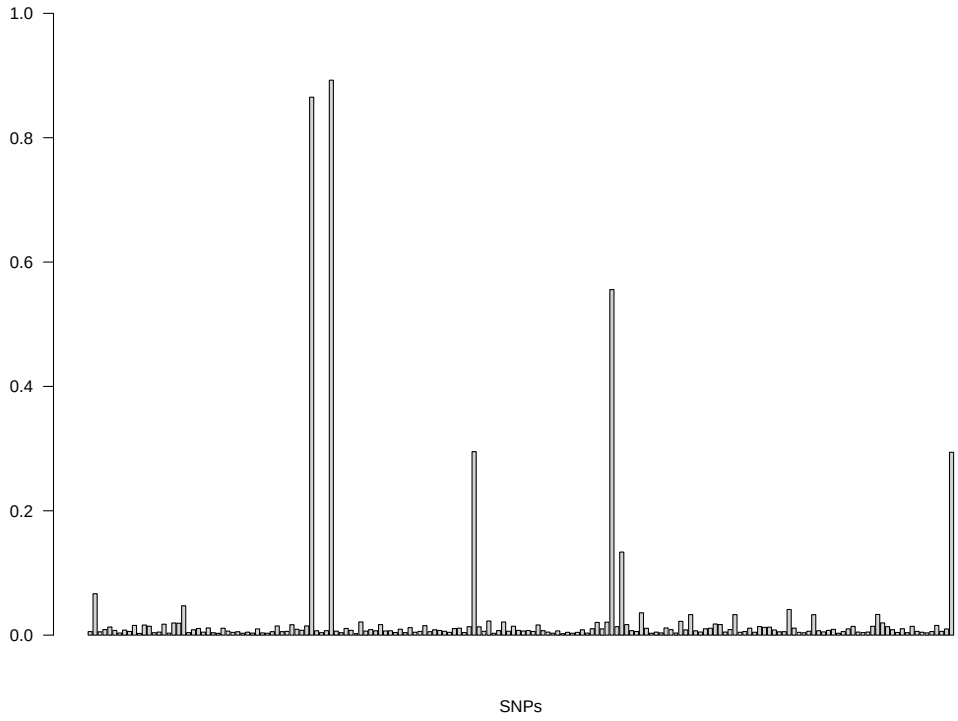
Case 1: All X are independent.

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

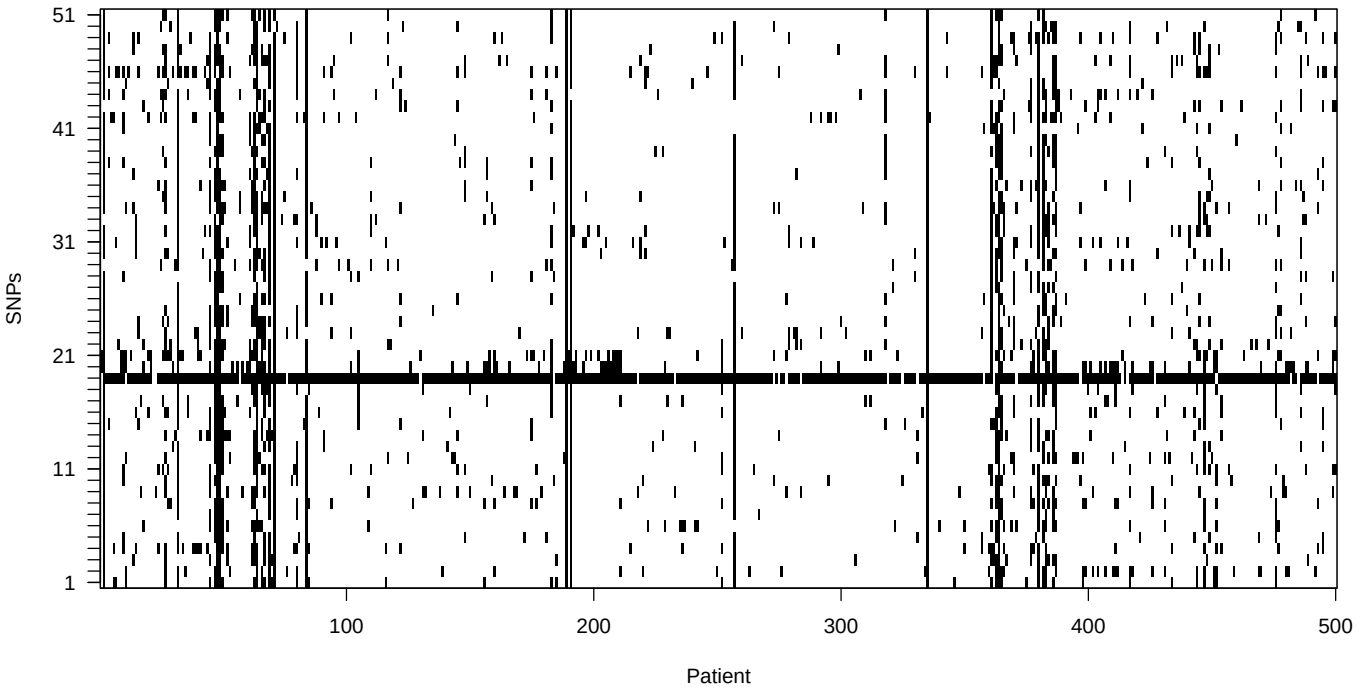
Multiple Models



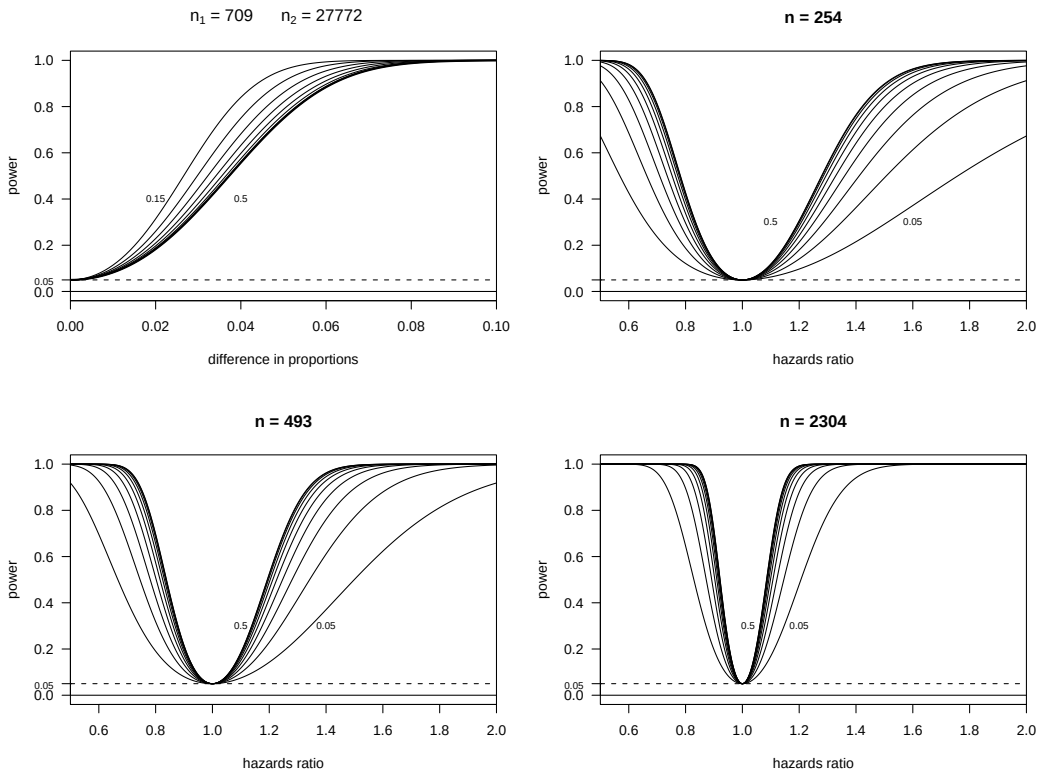
Multiple Models



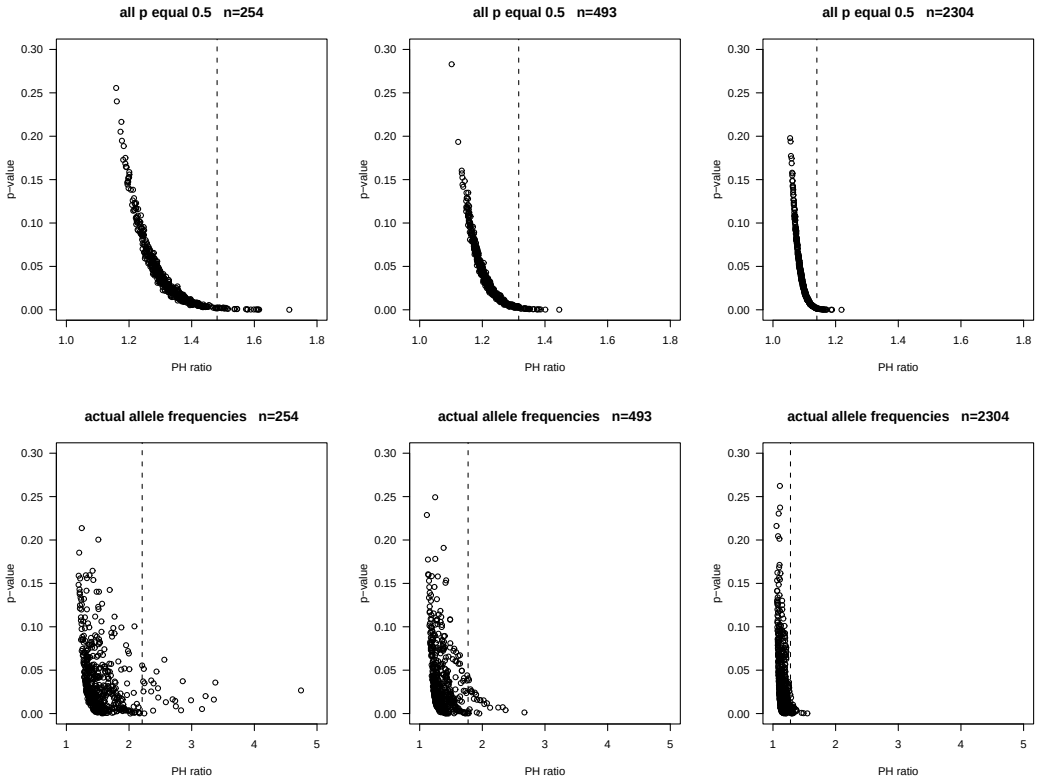
Statistical Issues: Missing Values



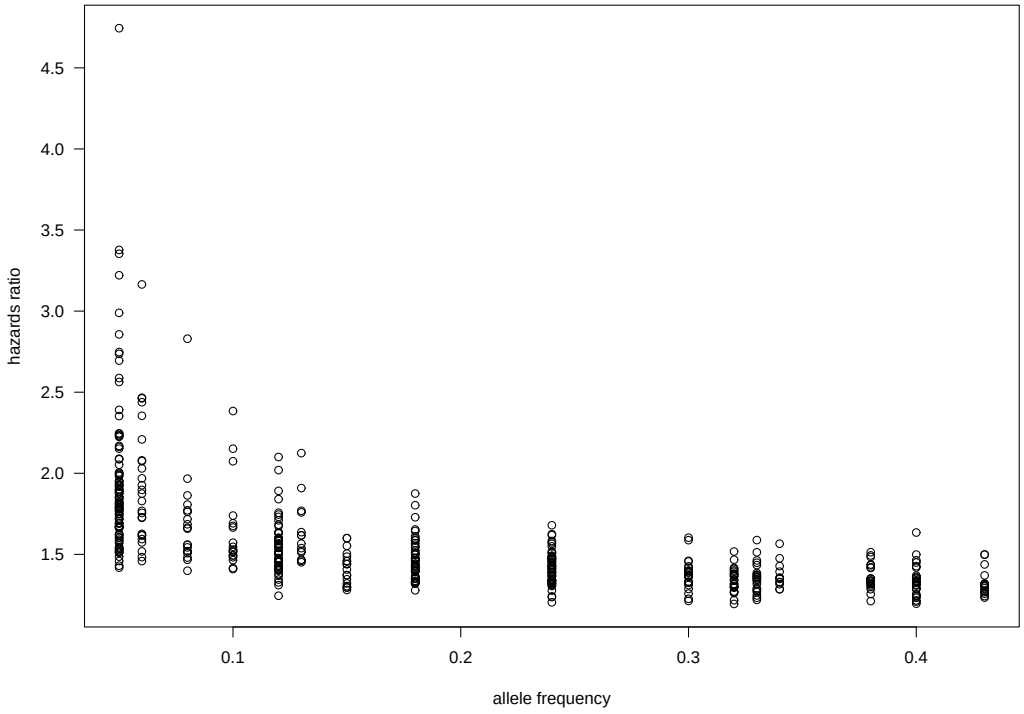
Statistical Issues: Power



Statistical Issues: Power



Statistical Issues: Power



References

- 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. (2003),
  Logic Regression
  Journal of Computational and Graphical Statistics (to appear).
  Available at: http://biostat.jhsph.edu/∼iruczins/

The Bibliography of this paper contains all the references in this presentation.