Regression Tree Methods for Precision Medicine

Wei-Yin Loh
Department of Statistics
University of Wisconsin–Madison
http://www.stat.wisc.edu/~loh/IMS17/nus17.pdf

About GUIDE

- GUIDE algorithm and software have been in development for 30 years
- GUIDE manual and compiled code for Linux, Mac OS X and Windows are available at `www.stat.wisc.edu/~loh/guide.html` free of charge
- GUIDE is not implemented in R but can be used in R (see manual)
- Key references in chronological order: Loh and Vanichsetakul (1988), Chaudhuri et al. (1994), Chaudhuri et al. (1995), Loh and Shih (1997), Kim and Loh (2001), Loh (2002), Loh (2009), Loh and Zheng (2013), Loh et al. (2015), and Loh et al. (2016)
- Research support at various stages provided by grants from the U.S. Army Research Office, National Science Foundation, National Institutes of Health, Bureau of Labor Statistics, IBM, Pfizer, and Eli Lilly

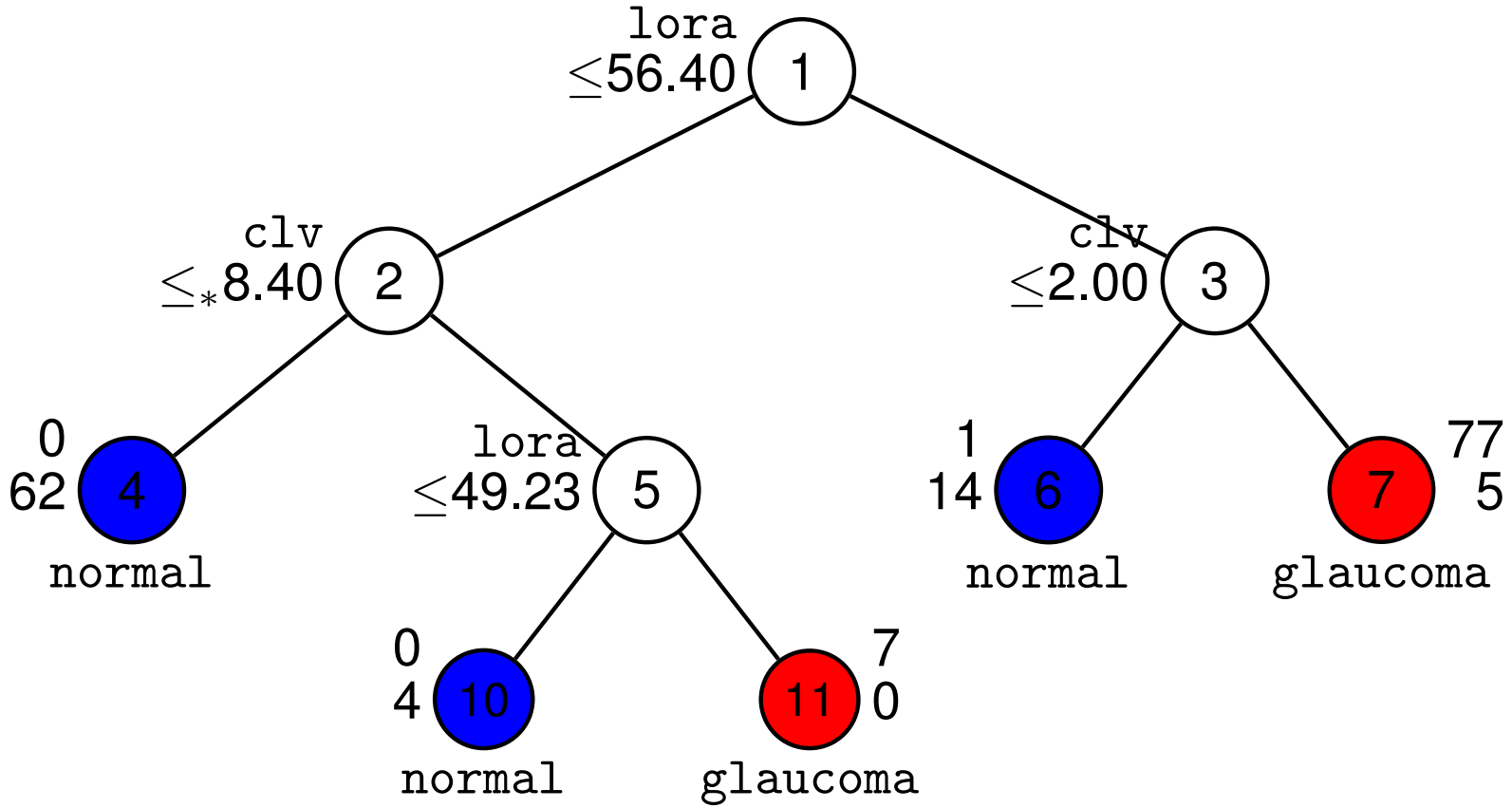
Outline

1. Glaucoma: classification tree with continuous predictors
2. Peptide binding: classification tree with categorical predictors
3. Key differences between GUIDE and CART: heart disease data
4. Alzheimer's disease: clustering longitudinal trajectories
5. Subgroup identification for differential treatment effects:
   (a) Breast cancer (no missing values)
   (b) Retrospective gene study (massive numbers of missing values)
   (c) Type II diabetes: longitudinal response
6. Bootstrap calibrated confidence intervals
7. Local linear control of prognostic variables
8. Comparison with Interaction Trees, Virtual Twins, and SIDES

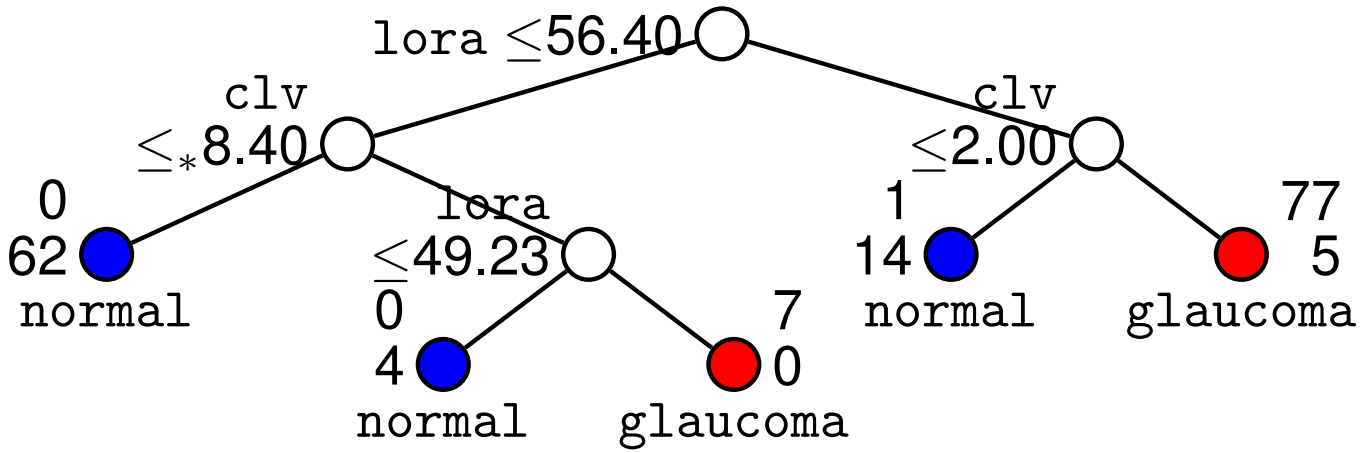
Glaucoma data

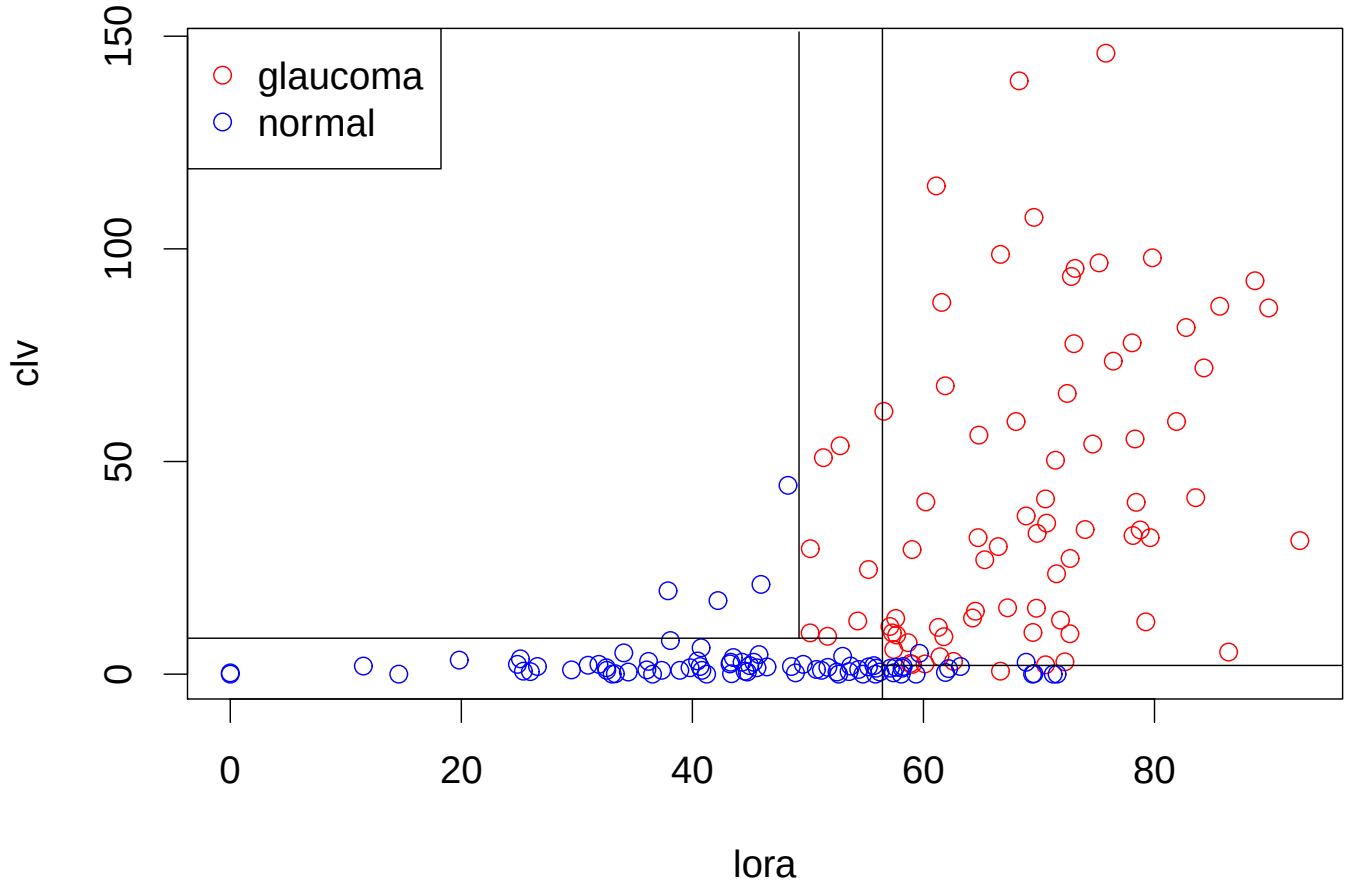
- From `TH.data` package; 170 observations, 17 with missing values
- Class variable takes two values: glaucoma and normal
- `lora` is loss of rim area, measured by fundus photography, no missing
- `clv` is corrected loss variance of the visual field, 12 missing
- 64 other predictors derived from a confocal laser scanning image of the optic nerve head, a visual field test, fundus photography and a measurement of the intra-occular pressure

GUIDE classification tree

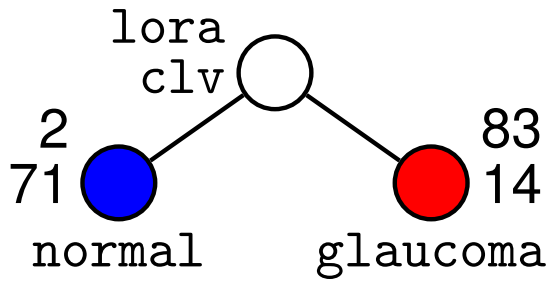


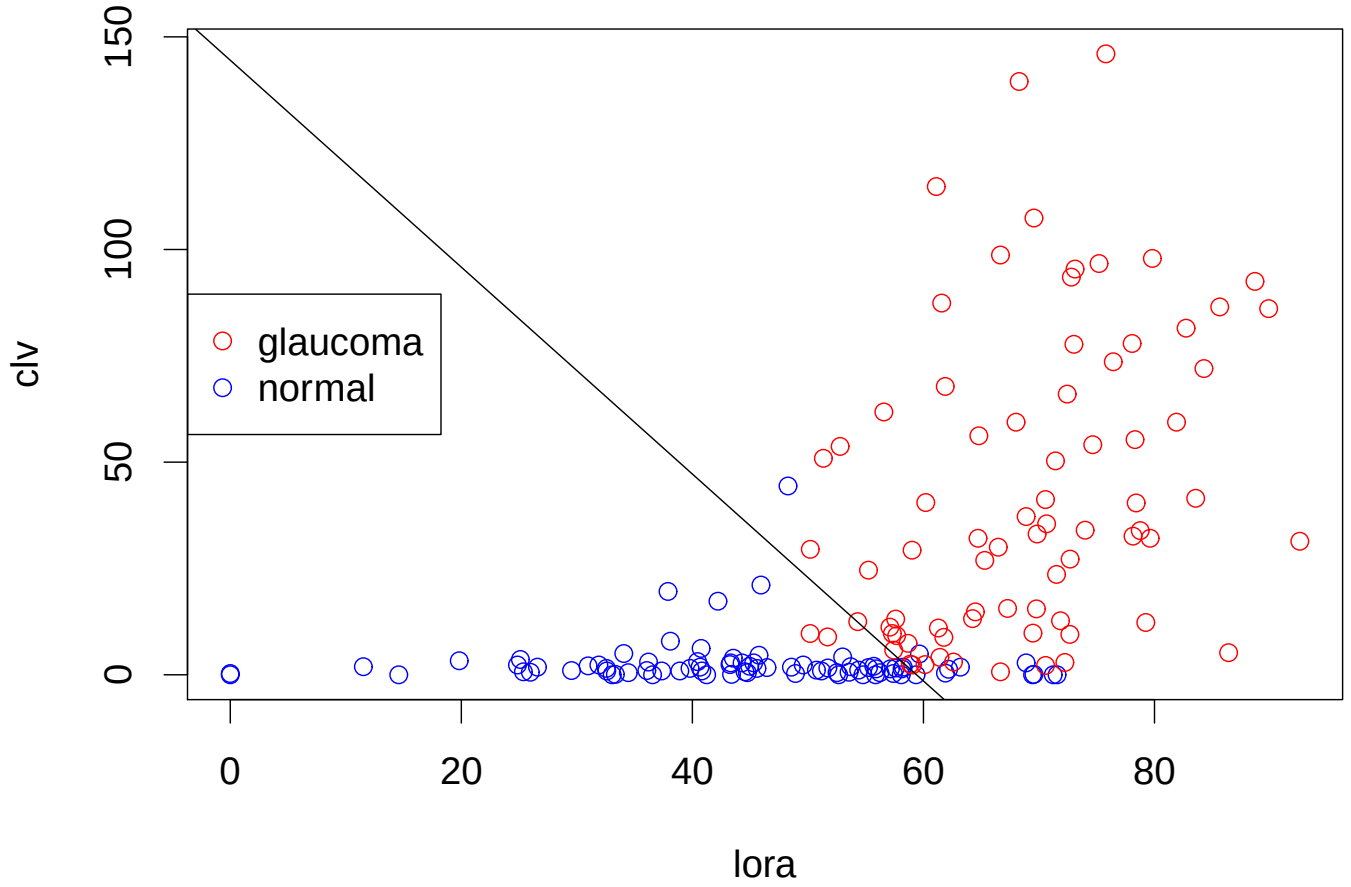
At each split, an observation goes to the left branch if and only if the condition is satisfied. Symbol '$\leq_*$' stands for '$\leq$ or missing'. Sample sizes for `Class = glaucoma` and `normal`, respectively, beside nodes.


lora ≤56.40
clv
≤* 8.40
clv
≤2.00
0
62
normal
lora
≤49.23
0
4
normal
7
0
glaucoma
1
14
normal
77
5
glaucoma


glaucoma
normal
clv
lora
0
50
100
150
0
20
40
60
80

Linear splits





Classification with categorical predictors: peptide-binding data

- 310 peptides; 181 bind to Class I MHC molecule, 129 do not
- Peptides are biologically occurring short chains of amino acid monomers linked by peptide (amide) bonds
- Class I molecules are cell surface proteins that present foreign peptides as targets for cytotoxic T lymphocytes that destroy the infected cell
- Each peptide is an amino acid sequence of length 8
- Each position in a sequence is one of 18–20 amino acids
- Problem: Which amino acids in which positions are predictive of binding?
- `http://repositories.cdlib.org/cbmb/peptide_binding`

Peptide-binding data

ID	Binder	Pos1	Pos2	Pos3	Pos4	Pos5	Pos6	Pos7	Pos8
1	1	S	S	P	S	H	P	G	M
2	1	S	M	I	T	F	T	P	L
3	1	S	M	V	A	P	P	H	L
4	1	Y	S	P	P	Y	S	S	I
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
307	0	S	P	S	N	P	S	V	F
308	0	T	P	Y	S	R	P	P	T
309	0	P	Y	S	R	P	P	T	P
310	0	Y	S	R	P	P	T	P	R
	#levels	18	20	20	20	20	20	19	20

Logistic regression

- 8 categorical predictors (18–20 levels each) need 149 dummy variables
- Model without interactions does not converge — quasi-complete separation
- Model with interactions is impossible

LASSO logistic regression coefficients

pos1C	-0.069	pos3D	-0.686	pos5D	-1.025	pos7D	-0.771
pos1D	-1.111	pos3P	0.876	pos5E	-0.301	pos7F	-4.452
pos1L	-0.623	pos3R	-1.303	pos5F	2.462	pos7G	-0.101
pos1P	-1.180	pos3T	-2.466	pos5H	0.634	pos7L	-0.045
pos1Q	0.120	pos4L	-0.668	pos5L	0.229	pos7S	0.209
pos1S	1.913	pos4M	-0.100	pos5M	1.615	pos8I	0.990
pos1T	0.821	pos4P	0.540	pos5N	-1.249	pos8L	1.676
pos1Y	0.042			pos5P	-0.549	pos8M	0.740
pos2D	-1.288			pos5R	-1.006	pos8N	-0.058
pos2E	-0.631			pos5S	-1.260	pos8P	-1.126
pos2M	1.025			pos5T	-1.536	pos8Q	-0.732
pos2N	1.962			pos5Y	2.052	pos8Y	1.957
pos2P	-0.430			pos6D	-0.378		
pos2S	0.156			pos6L	-0.030		
				pos6S	0.731		
				pos6W	-0.389		

Artificial neural networks (Milik et al., 1998)

Two necessary steps:

1. Encode the categorical variables into numerical variables
2. Randomly split data into training and validation samples

Three encoding schemes:

1. Convert each amino acid into a 20-dim 0-1 vector: **149** 0-1 variables total
— ANN over-fits due to too many variables

2. Classify each acid into 10 hierarchical clusters, then code each acid with a 10-dim 0-1 vector to indicate cluster membership: **80** 0-1 variables total

3. Represent each acid by 13 physico-chemical numerical properties, e.g., volume, bulkiness, flexibility, polarity, aromaticity, and charge: **104** variables total

Encoding schemes 2 and 3 employ information not in data

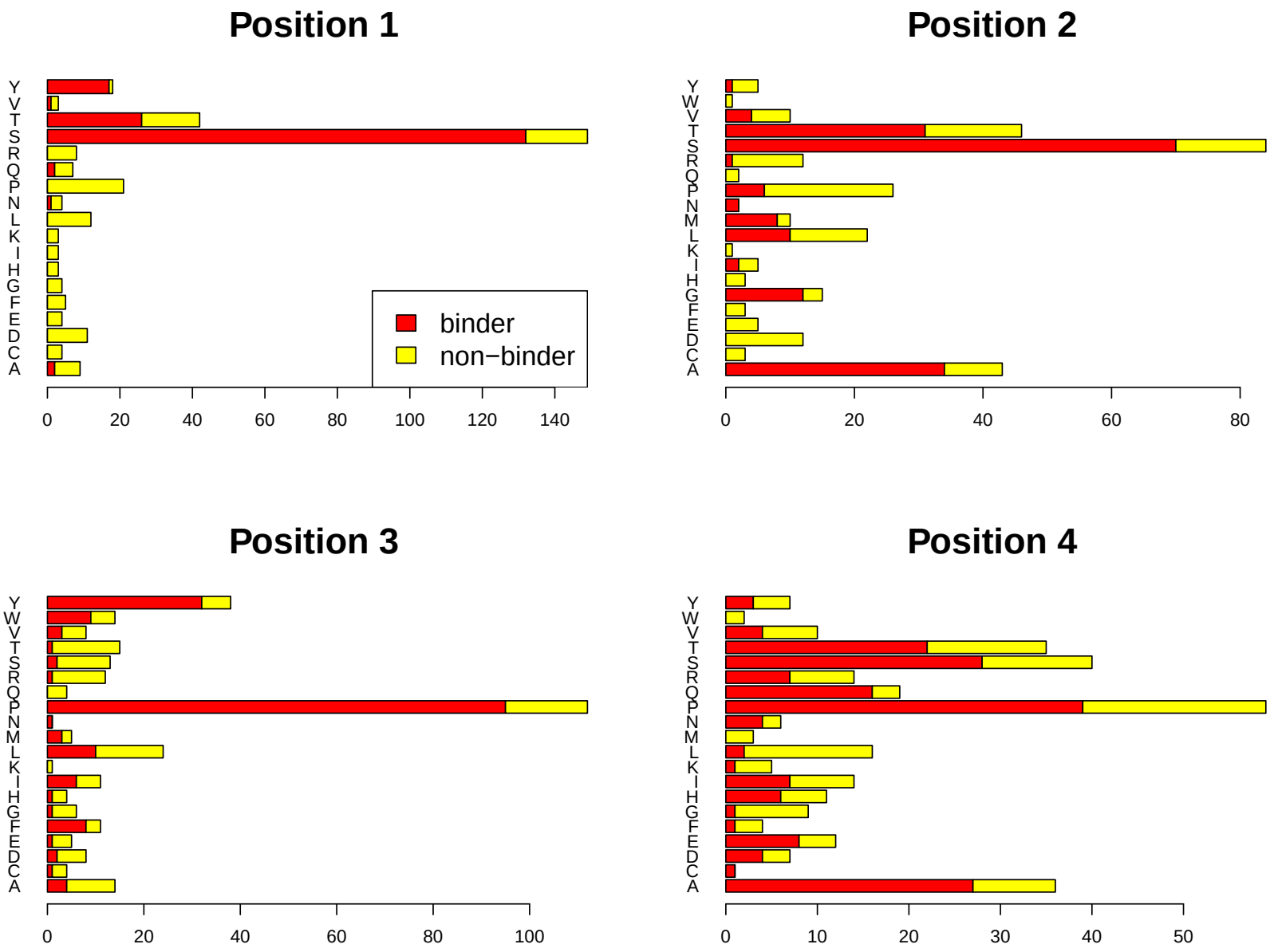
Clustering scheme 2

Cluster	Feature	Amino acid codes
0	hydrophobic	ACFGHIKLMTVWY
1	aliphatic	ILV
2	aromatic	FHWY
3	polar	DEHKNQRSTWY
4	charged	DEHKR
5	positive	HKR
6	small	ACDGNPSTV
7	tiny	AGS
8	glycine	G
9	proline	P

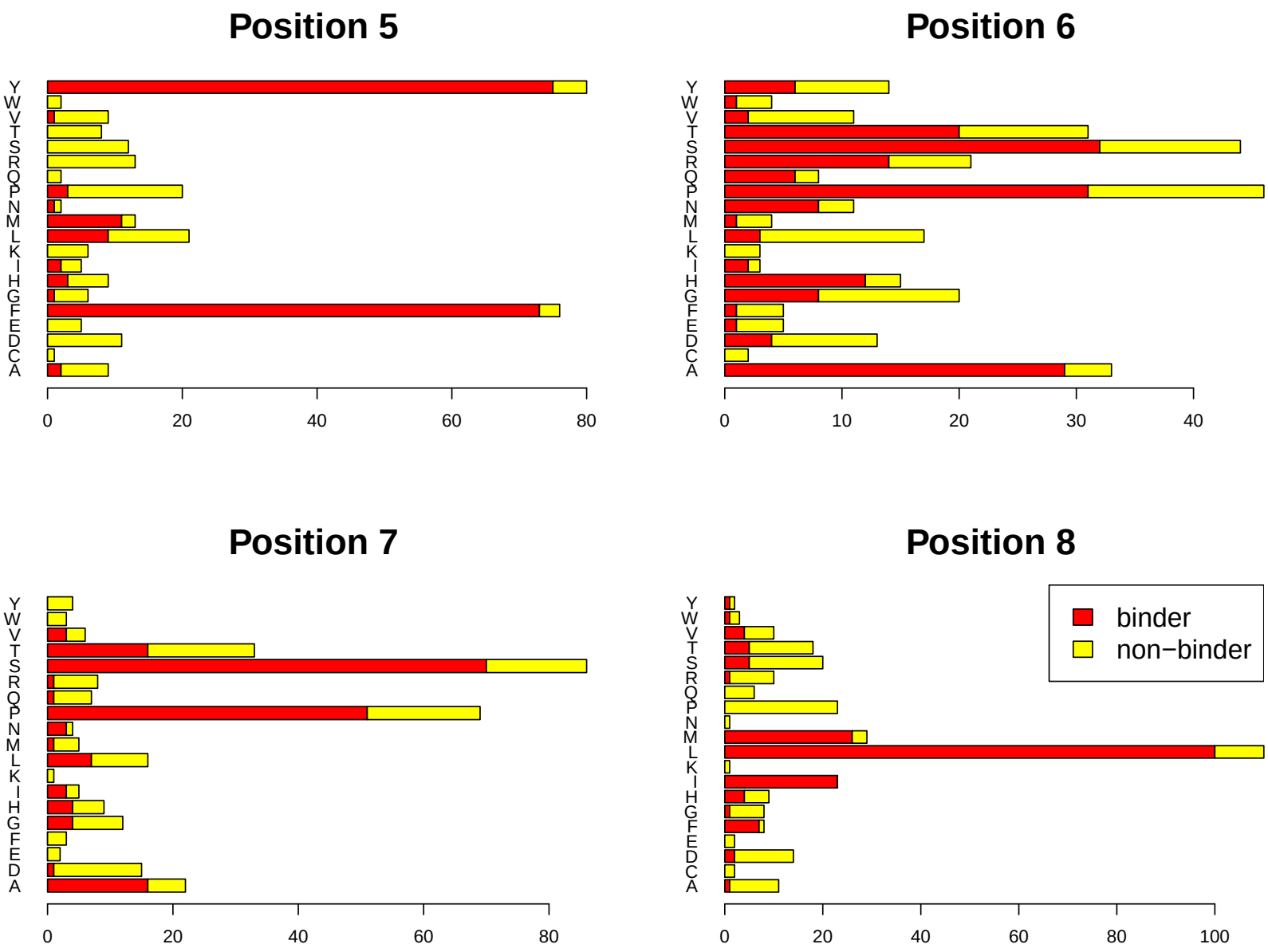
Representation scheme 3

	volume	bulkiness	flexibility	polarity	aromaticity	charge
A	0.1677	0.4433	0.2490	0.3951	0.0	0.5
C	0.3114	0.5506	0.2048	0.0741	0.0	0.5
D	0.3054	0.4532	0.8675	1.000	0.0	0.0
E	0.4970	0.5567	0.8112	0.9136	0.0	0.0
F	0.7725	0.8976	0.0763	0.0370	0.6667	0.5
G	0.0	0.0	1.0	0.5062	0.0	0.5
H	0.5569	0.5632	0.1124	0.6790	0.5556	0.5
I	0.6467	0.9852	0.6707	0.0370	0.0	0.5
K	0.6946	0.6738	0.6867	0.7901	0.0	1.0
L	0.6467	0.9852	0.2811	0.0	0.0	0.5
M	0.6108	0.7033	0.0	0.0988	0.0	0.5
N	0.3174	0.5156	0.6747	0.8272	0.0	0.5
P	0.1766	0.7679	0.8594	0.3827	0.0	0.5
Q	0.4910	0.6048	0.7952	0.6914	0.0	0.5
R	0.7246	0.5955	0.9398	0.6914	0.0	1.0
S	0.1737	0.3222	0.8514	0.5309	0.0	0.5
T	0.3473	0.6771	0.5984	0.4568	0.0	0.5
V	0.4850	0.9945	0.3655	0.1235	0.0	0.5
W	1.0	1.0	0.0402	0.0617	1.0	0.5
Y	0.7964	0.8008	0.5020	0.1605	0.6667	0.5

Distributions of peptide-binding data



Distributions of peptide-binding data (cont'd.)



Chi-squared tests

Pos1 (X^2 = 185.25, df = 17, p-value $<$ 2.2E-16)

bind	A	C	D	E	F	G	H	I	K	L	N	P	Q	R	S	T	V	Y
0	7	4	11	4	5	4	3	3	3	12	3	21	5	8	17	16	2	1
1	2	0	0	0	0	0	0	0	0	0	1	0	2	0	132	26	1	17

Pos2 (X^2 = 111.3, df = 19, p-value = 4.6E-15)

bind	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
0	9	3	12	5	3	3	3	3	1	12	2	0	20	2	11	14	15	6	1	4
1	34	0	0	0	0	12	0	2	0	10	8	2	6	0	1	70	31	4	0	1

Pos3 (X^2 = 114.35, df = 19, p-value = 1.3E-15)

bind	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
0	10	3	6	4	3	5	3	5	1	14	2	0	17	4	11	11	14	5	5	6
1	4	1	2	1	8	1	1	6	0	10	3	1	95	0	1	2	1	3	9	32

Pos4 (X^2 = 51.475, df = 19, p-value = 7.9E-05)

bind	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
0	9	0	3	4	3	8	5	7	4	14	3	2	20	3	7	12	13	6	2	4
1	27	1	4	8	1	1	6	7	1	2	0	4	39	16	7	28	22	4	0	3

Chi-squared tests (cont'd.)

Pos5 (X^2 = 211.5, df = 19, p-value $<$ 2.2E-16)

bind	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
0	7	1	11	5	3	5	6	3	6	12	2	1	17	2	13	12	8	8	2	5
1	2	0	0	0	73	1	3	2	0	9	11	1	3	0	0	0	0	1	0	75

Pos6 (X^2 = 66.888, df = 19, p-value = 3.0E-07)

bind	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
0	4	2	9	4	4	12	3	1	3	14	3	3	15	2	7	12	11	9	3	8
1	29	0	4	1	1	8	12	2	0	3	1	8	31	6	14	32	20	2	1	6

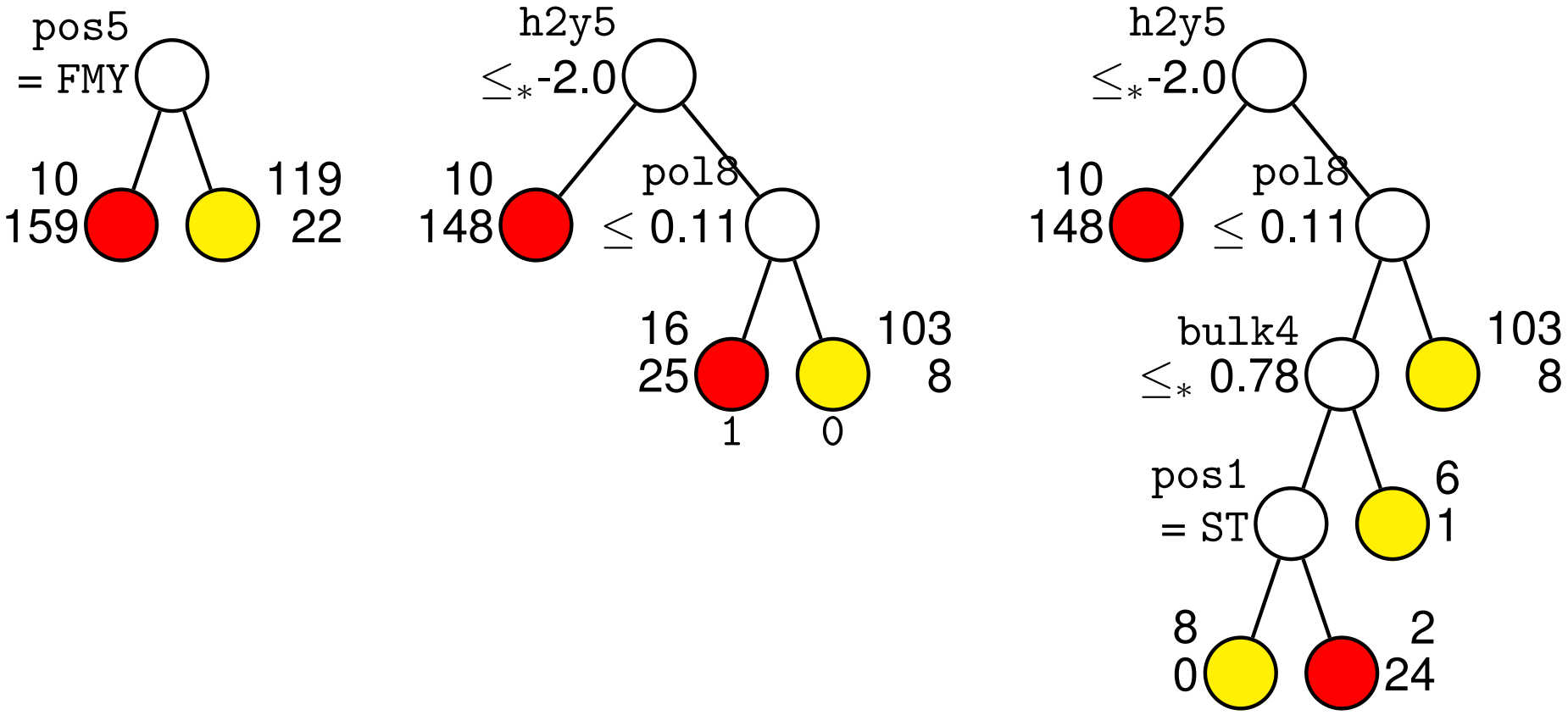
Pos7 (X^2 = 84.966, df = 18, p-value = 1.1E-10)

bind	A	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
0	6	14	2	3	8	5	2	1	9	4	1	18	6	7	16	17	3	3	4
1	16	1	0	0	4	4	3	0	7	1	3	51	1	1	70	16	3	0	0

Pos8 (X^2 = 185.69, df = 19, p-value $<$ 2.2E-16)

bind	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V
0	10	2	12	2	1	7	5	0	1	10	3	1	23	6	9	15	13	6
1	1	0	2	0	7	1	4	23	0	100	26	0	0	0	1	5	5	4

GUIDE classification trees for peptide data using 8 position vars (left), 104 physical vars (middle) and all 112 vars (right)



At each split, an obs. goes to left branch if and only if the condition is satisfied
Sample sizes for non-binder (yellow) and binder (red) beside nodes.

Sensitivity and specificity

	Classification trees		ANN	
	8 cat. vars	104 phy. vars	80 cluster vars	48 phys. vars
Sensitivity	0.878	0.950	0.700	0.450
Specificity	0.922	0.845	0.800	0.550

Basic components of tree construction

1. A set of **training data** consisting of Y and X variables
2. A **node impurity** function:
   (a) Gini index or entropy function (classification)
   (b) Sum of squared residuals or deviance (regression)

Recursive partitioning: Key steps in tree model construction

Given a measure of node impurity,

1. Find X_i and c_i or S_i to split the data in a node with "$X_i \leq c_i$" (if X_i is ordinal) or "$X_i \in S_i$" (if X_i is categorical). Two paradigms:
   (a) Greedy search: find best c_i or S_i for each X_i, then select "best of best" split — used by **CART** and many algorithms
   (b) Unbiased search: find most significant X_i, then find best split on selected X_i — used by **GUIDE** and a few others
2. Decide when to stop splitting. Two approaches:
   (a) Use a stopping rule
   (b) Grow a large tree, then prune
3. Predict Y for each observation in each terminal node

Some notation

Y: response variable

J: number of classes

$\mathcal{C} = \{1, \ldots, J\}$: set of classes

N: training sample size

K: number of predictor variables

$\mathbf{X} = (X_1, \ldots, X_K)$: vector of predictor variables

$\mathcal{X}$: Space of predictor variables

More notation

t denotes a node

J is the number of classes in training sample

J_t is the number of classes in t

$N(t)$ is the number of training samples in t

N_j is the number of class j training samples

$N_j(t)$ is the number of class j training samples in t

T denotes a tree

$\tilde{T}$ is the set of terminal nodes of T

$|\tilde{T}|$ is number of terminal nodes of T

T_t is a subtree of T with root node t

$\{t\}$ is a subtree of T_t containing only the root node t

Node impurity measures for classification

Let $p(j|t)$ be the proportion of class j learning samples in node t. Define the **node impurity measure**

$$i(t) = \phi(p(\cdot|t)) \geq 0$$

where ϕ is a symmetric function with maximum value $\phi(J^{-1}, J^{-1}, \ldots, J^{-1})$ and

$$\phi(1, 0, \ldots, 0) = \phi(0, 1, \ldots, 0) = \ldots = \phi(0, 0, \ldots, 0, 1) = 0$$

Entropy: $i(t) = -\sum_{j=1}^{J} p(j|t) \log p(j|t)$

Gini index: $i(t) = 1 - \sum_j p^2(j|t) = \sum_{j=1}^{J} p(j|t)(1 - p(j|t))$

- We use $g(t)$ to denote the Gini index
- If $J = 2$, then $g(t) = 2p(1|t)p(2|t)$, i.e., two times binomial variance

Split selection

1. Define the goodness of a split s as

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R)$$

   where t_L and t_R are the left and right subnodes of t and p_L and p_R are the probabilities of being in those subnodes.

2. Define a set $\mathcal{S}$ of binary splits of the form $X \in A$, where,

$$\begin{aligned} A &= (-\infty, c], && \text{if } X \text{ is non-categorical} \\ A &\subset \mathcal{X}, && \text{if } X \text{ is categorical} \end{aligned}$$

3. Find $s^* \in \mathcal{S}$ such that $\Delta i(s^*,t) = \max_{s \in \mathcal{S}} \Delta i(s,t)$.

CART approach (Breiman et al., 1984)

1. Find the pair (X, S) such that the split $X \in S$ maximizes the decrease in node impurity
2. Let $C(i|j)$ be the cost of misclassifying a class j case as class i. Assign terminal node t to class j^* if it minimizes the misclassification cost

$$\sum_j C(j^*|j)p(j|t) = \min_i \sum_j C(i|j)p(j|t)$$

3. Prune tree using test sample or cross-validation
4. Use surrogates splits to deal with missing values

Estimates of misclassification error

Resubstitution estimate:

$$R(d) = N^{-1} \sum_{n=1}^{N} I(d(\mathbf{x}_n) \neq j_n)$$

This is usually overly optimistic

Test sample estimate: Divide $\mathcal{L}$ into $\mathcal{L}_1$ and $\mathcal{L}_2$. Let $N_2 = \#\mathcal{L}_2$. Construct d from $\mathcal{L}_1$. Then

$$R^{ts}(d) = N_2^{-1} \sum_{\mathcal{L}_2} I(d(\mathbf{x}_n) \neq j_n)$$

This is unbiased and computationally efficient

V-fold cross-validation estimate:

1. Divide $\mathcal{L}$ into V subsets $\mathcal{L}_1, \ldots, \mathcal{L}_V$
2. Let $d^{(v)}$ denote the classifier constructed from $\mathcal{L} - \mathcal{L}_v$
3. Define

$$R^{ts}(d^{(v)}) = N_v^{-1} \sum_{\mathcal{L}_v} I(d^{(v)}(\mathbf{x}_n) \neq j_n)$$

4. The V-fold cross-validation estimate is

$$R^{cv}(d) = V^{-1} \sum_{v=1}^{V} R^{ts}(d^{(v)})$$

CART pruning

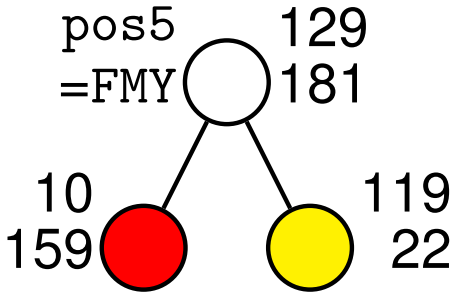
1. Given α and tree T, define the cost-complexity function

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|$$

2. For each α, there is a tree T that minimizes the cost-complexity

3. Let t be any node and T_t be the branch of T with root node t. Then

$$\begin{aligned} R_\alpha(\{t\}) &= R(t) + \alpha = 129/310 + \alpha \\ R_\alpha(T_t) &= R(T_t) + \alpha|\tilde{T}_t| = (10 + 22)/310 + 2\alpha \end{aligned}$$



4. Critical value of α for which $R_\alpha(T_t) = R_\alpha(\{t\})$ is $\alpha = u(t)$, where

$$u(t) = [R(t) - R(T_t)]/[|\tilde{T}_t| - 1] = (129 - 32)/(310(2 - 1)) = 97/310$$

5. Prune branches at nodes t_1 for which $u(t_1) = \min\{u(t) \, : \, t \in T - \tilde{T}\}$

6. Define $\alpha_1 = u(t_1)$ and iterate to obtain a nested sequence of trees

Subtree selection by test-sample estimation

- Estimate the misclassification cost for each subtree with the test sample
- Select the subtree with the smallest estimated cost

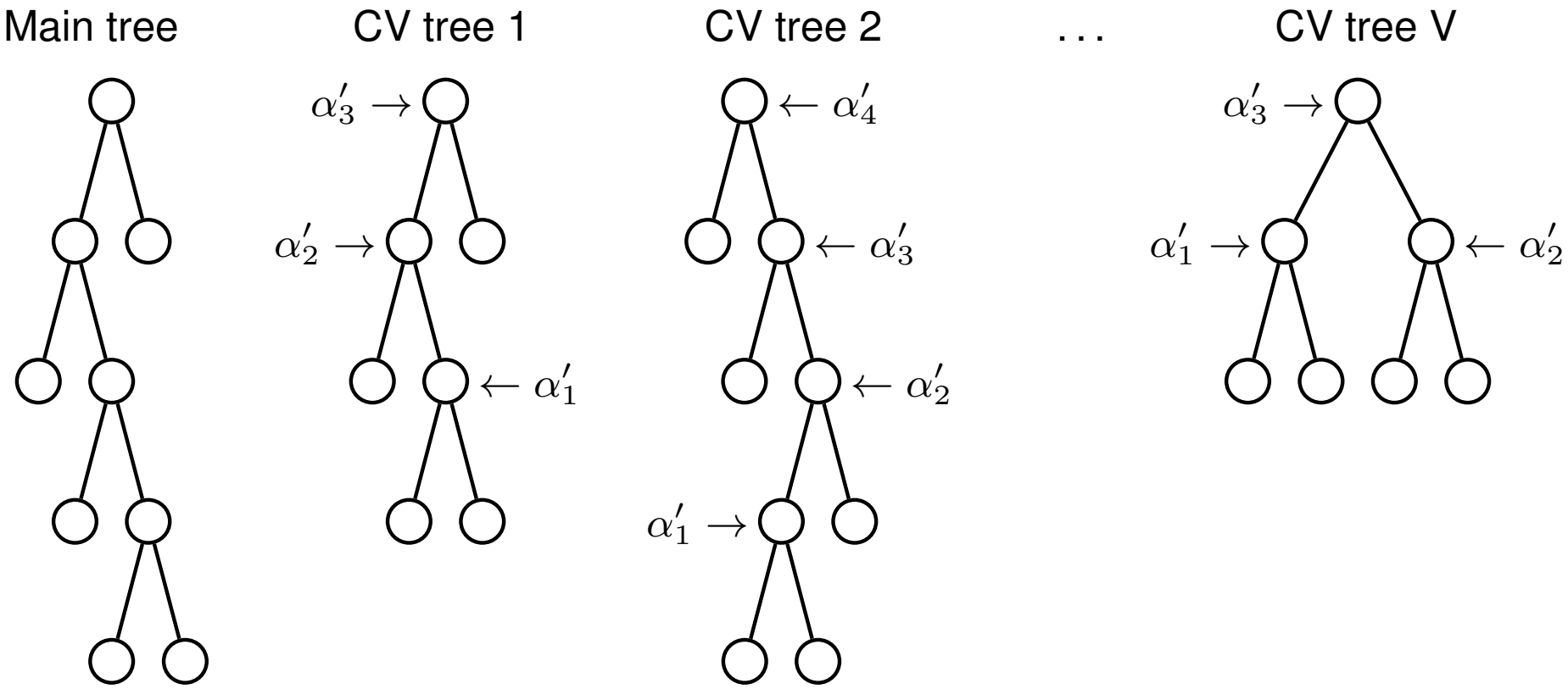
Subtree selection by V-fold cross-validation

1. Let $\alpha_1 < \alpha_2 < \ldots$ be the α-values associated with the pruned sequence of subtrees $T_1 \succ T_2 \succ \ldots$. Define $\alpha'_k = \sqrt{\alpha_k \alpha_{k+1}}$
2. Divide $\mathcal{L}$ into V subsets $\mathcal{L}_1, \ldots, \mathcal{L}_V$
3. Let $T^{(v)}(\alpha'_k)$ be the minimal cost-complexity tree grown from $\mathcal{L} - \mathcal{L}_v$, $v = 1, \ldots, V$
4. Let $R'(T^{(v)}(\alpha'_k))$ be the estimate of the misclassification cost of $T^{(v)}(\alpha'_k)$ based on the test sample $\mathcal{L}_v$
5. The V-fold CV estimate for subtree T_k is

$$R^{cv}(T_k) = V^{-1} \sum_{v=1}^{V} R'(T^{(v)}(\alpha'_k))$$

6. Select the subtree with the smallest CV cost

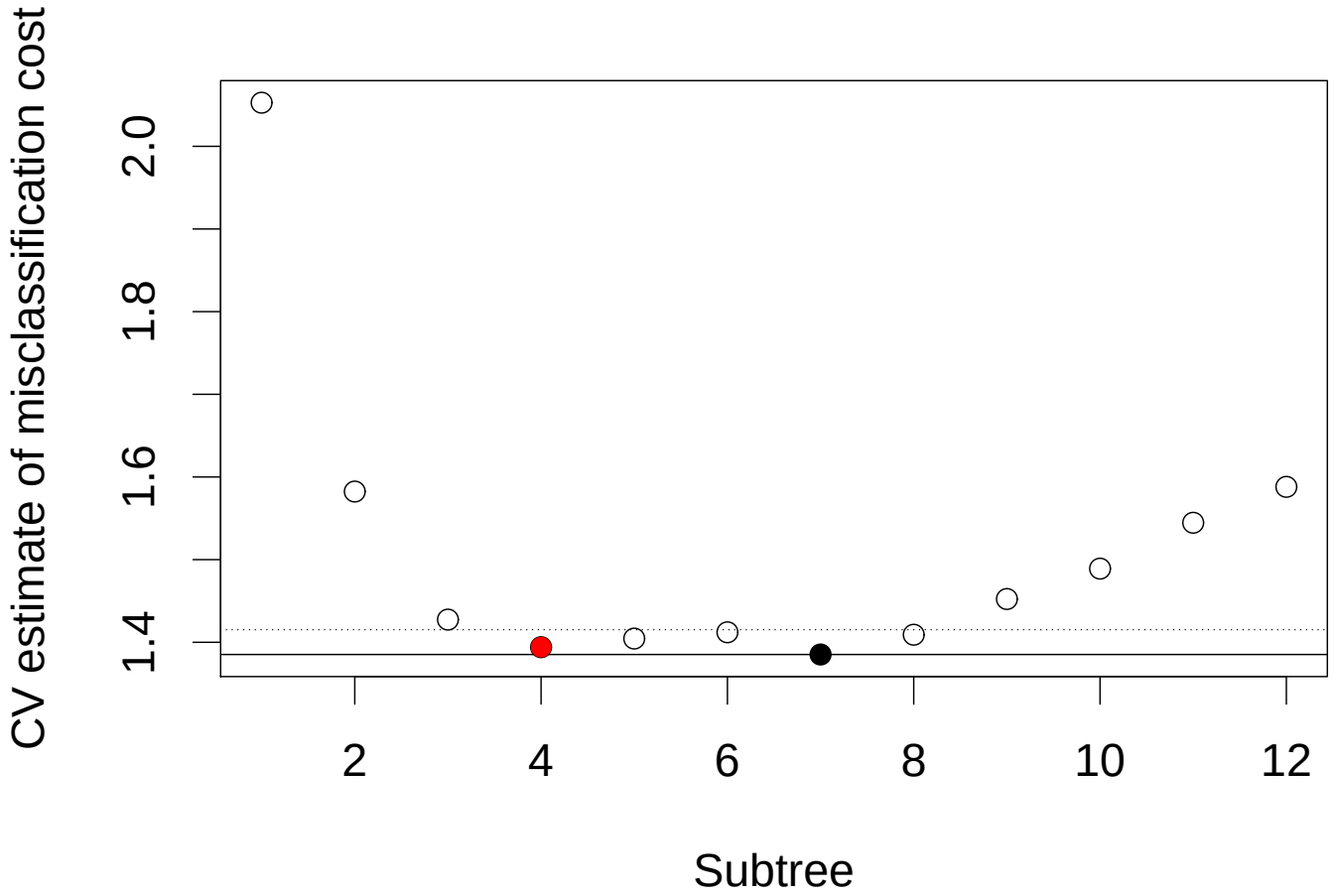
V-fold cross-validation



- Main tree is grown using all the data
- Each CV tree is grown using $(V-1)$ subsets

k-SE rule

1. Let $\hat{R}(T)$ = CV estimate of misclass. cost of T; let $\widehat{\mathrm{SE}}[\hat{R}(T)]$ be its SD
2. Let subtree T^* minimize $\hat{R}(T_k)$
3. k-SE tree T^{**} is smallest subtree s.t. $\hat{R}(T^{**}) \leq \hat{R}(T^*) + k \times \widehat{\mathrm{SE}}[\hat{R}(T^*)]$



Missing values: CART surrogate splits

Suppose X^* is selected to split node t. Let s^* denote the split on X^*.

1. For each $X_i \neq X^*$, find split $\tilde{s}_i = \{X_i \in S_i\}$ that best predicts s^* in terms of maximizing the number of observations m_i going to the corresponding subnodes
2. Order $\tilde{s}_i$ in decreasing m_i to form a preferential set of surrogate splits

E.g., for peptide data, best surrogate split for s^* = {Pos5 = FMY} is $\tilde{s}_8$ = {Pos8 = FILM} because m_i is maximized at m_8 = 142 + 113 = 255

	Pos5 = FMY	Pos5 $\neq$ FMY
Pos8 = FILM	142	28
Pos8 $\neq$ FILM	27	113

CART surrogate splits for classification

- Let s^* be the best split of t into t_L and t_R
- For each k, let $\mathcal{S}_k$ be the set of all splits on X_k
- For any $s' \in \mathcal{S}_k$, let $t_{L'}$ and $t_{R'}$ denote the subnodes of s'
- Let $N_j(LL')$, $N_j(RR')$ be the number of class j cases in $t_L \cap t_{L'}$, $t_R \cap t_{R'}$
- Let

$$p(t_L \cap t_{L'}) = \sum_j \pi(j) N_j(LL')/N_j, \quad p(t_R \cap t_{R'}) = \sum_j \pi(j) N_j(RR')/N_j$$

[Note: If there are missing values, $N_j(LL')$ and $N_j(RR')$ are numbers of cases with non-missing values in X_k and the variable involved in s.]

CART surrogate splits (cont'd.)

- Let $p(j,t) = \pi(j)N_j(t)/N_j$ and $p(t) = \sum_j p(j,t)$
- Let $p_{LL'}(s^*, s')$ be an estimate of $P(\text{both } s^* \text{ and } s' \text{ send a case left})$:

$$p_{LL'}(s^*, s') = p(t_L \cap t_{L'})/p(t)$$

- Similarly, define $p_{RR'}(s^*, s') = p(t_R \cap t_{R'})/p(t)$
- Estimate $P(s' \text{ predicts } s^*)$ by

$$p(s^*, s') = p_{LL}(s^*, s') + p_{RR}(s^*, s')$$

- $\tilde{s}_k$ is called a **surrogate split** on X_k for s^* if

$$p(s^*, \tilde{s}_k) = \max\{p(s^*, s') \ : \ s' \in \mathcal{S}_k\}$$

Measure of association for surrogate splits

- Let p_L and p_R be the probabilities that s^* sends a case to t_L and t_R, resp.
- The naive predictor sends every case to t_L if $p_L \geq p_R$ and to t_R otherwise
- Error probability of the naive predictor is $\min(p_L, p_R)$
- Define the measure of association between s^* and s as the relative reduction in error:

$$\lambda(s^*, s) = \frac{\min(p_L, p_R) - [1 - p(s^*, s)]}{\min(p_L, p_R)}$$

- Rank the surrogate splits according to their $\lambda(s^*, \tilde{s}_k)$ values
- If $\lambda(s^*, \tilde{s}_k) \leq 0$, $\tilde{s}_k$ is not used as a surrogate split

Uses of surrogate splits in CART

1. Enable tree construction when there are missing values in the learning sample
2. Enable classification of new cases with missing values
3. Rank variables by their order of importance
4. Detect masking of variables

Weaknesses and limitations of CART (and RPART)

CART searches for the "best" split for each X, with number depending on X:

Ordinal X with n unique values. $(n-1)$ splits of form "$X \leq a$"

Categorical X with c levels. $(2^{c-1}-1)$ splits of form "$X \in A$"

Consequently CART has selection bias:

1. Biased toward selecting X that have more splits — Breiman et al. (1984, p.42), Loh and Shih (1997)
2. Biased toward selecting X with **more** missing values (Kim and Loh, 2001)
3. Biased toward selecting surrogate variables with **fewer** missing values (Kim and Loh, 2001)

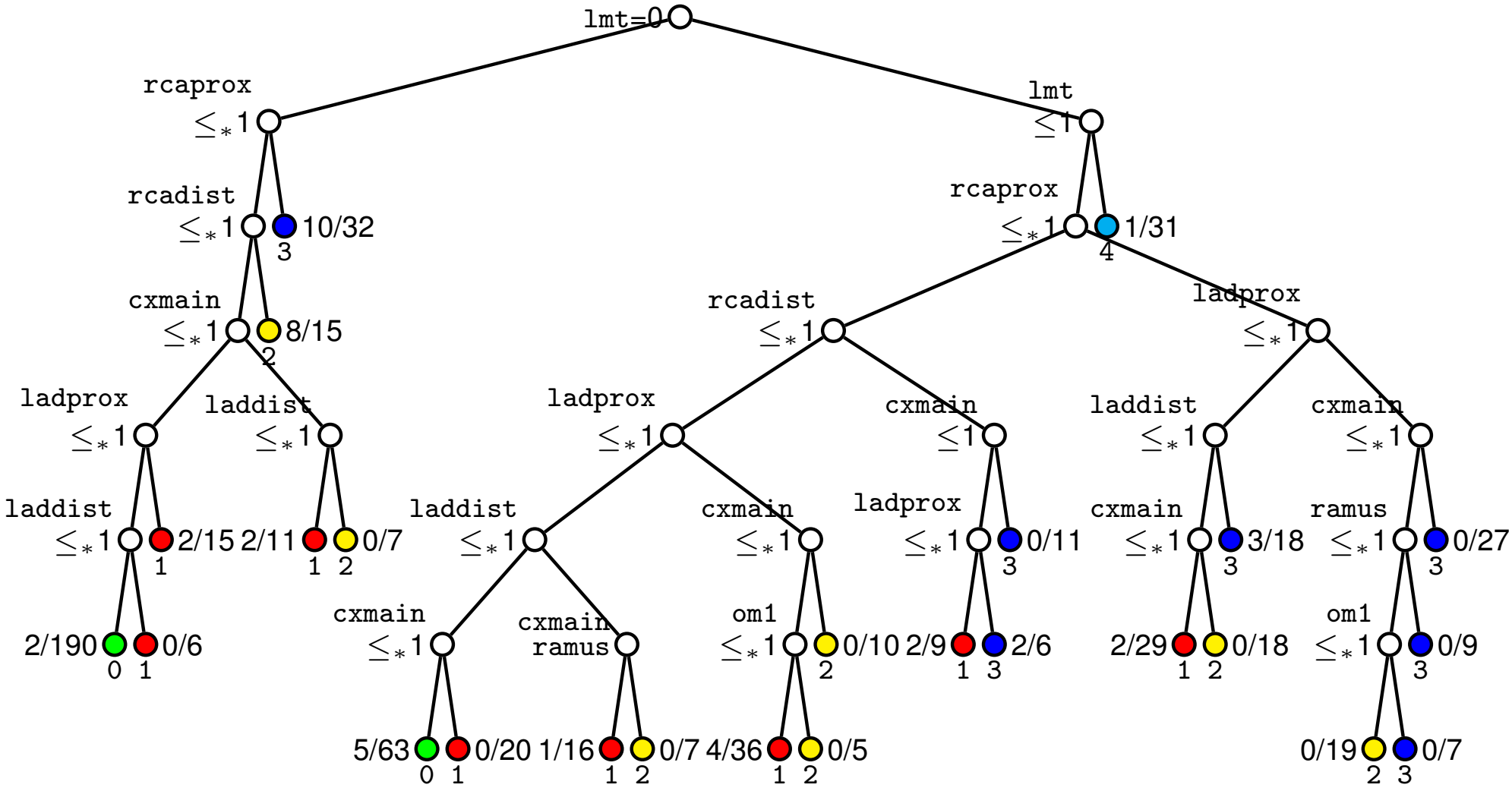
and these operational constraints:

1. Number of splits **increases linearly** in n and **exponentially** in c for ordinal and categorical, resp., X
2. CART can fit only **piecewise constant** regression trees

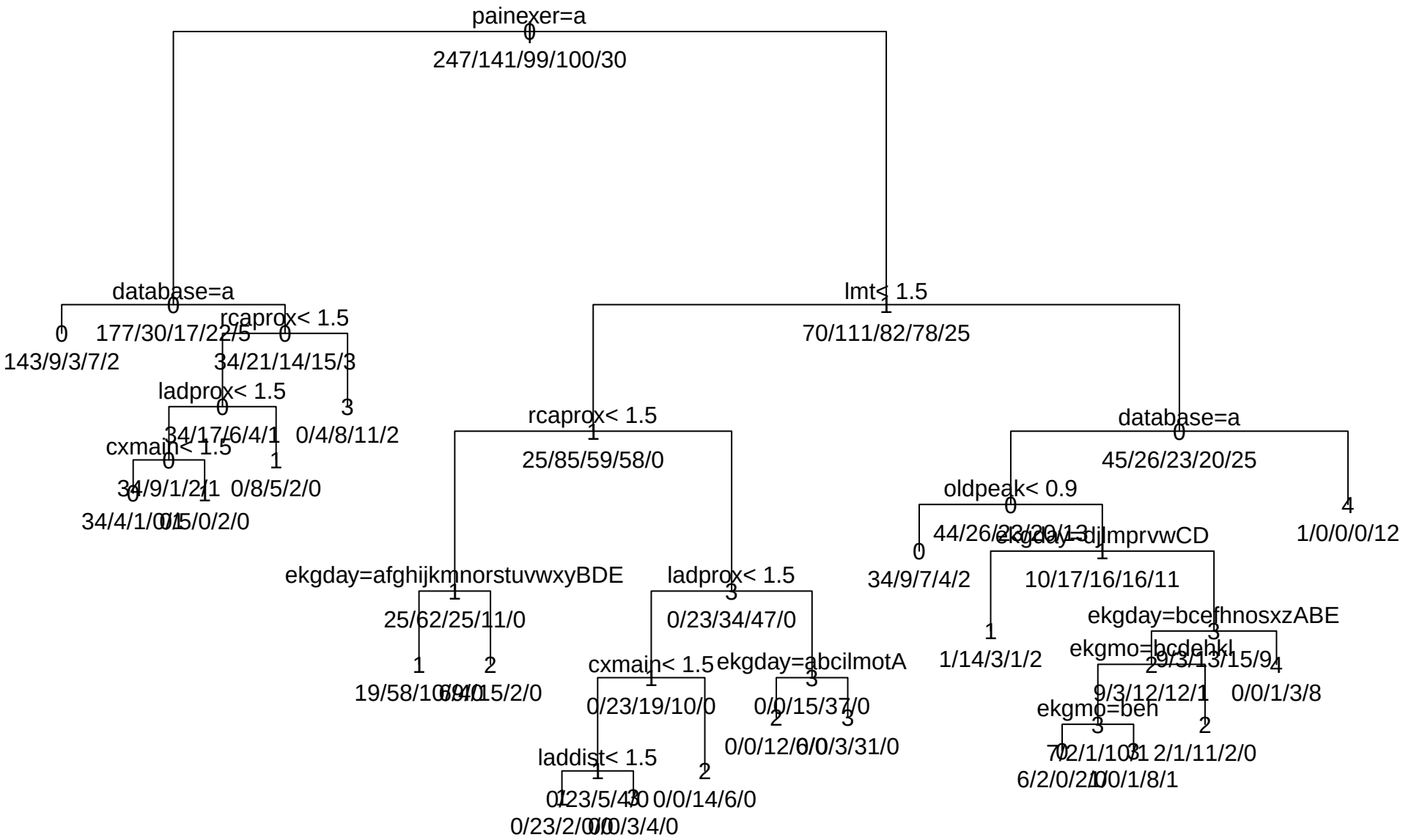
More than 2 classes and categorical variables: heart disease data

- 617 observations, no missing values
- Response is diagnosis of heart disease (5 levels)
- 52 predictor variables (29 ordinal, 23 categorical)
- Variables `ekgmo` and `ekgday` have 12 and 31 categorical levels

GUIDE tree for heart disease data (3 sec.)



RPART tree for heart disease data (3.6 hrs)



CART importance scoring of predictor variables

- The importance of variable x_k is measured by

$$M(x_k) = \sum_{t \in T} \Delta i(\tilde{s}_k, t)$$

- CART reports the standardized values

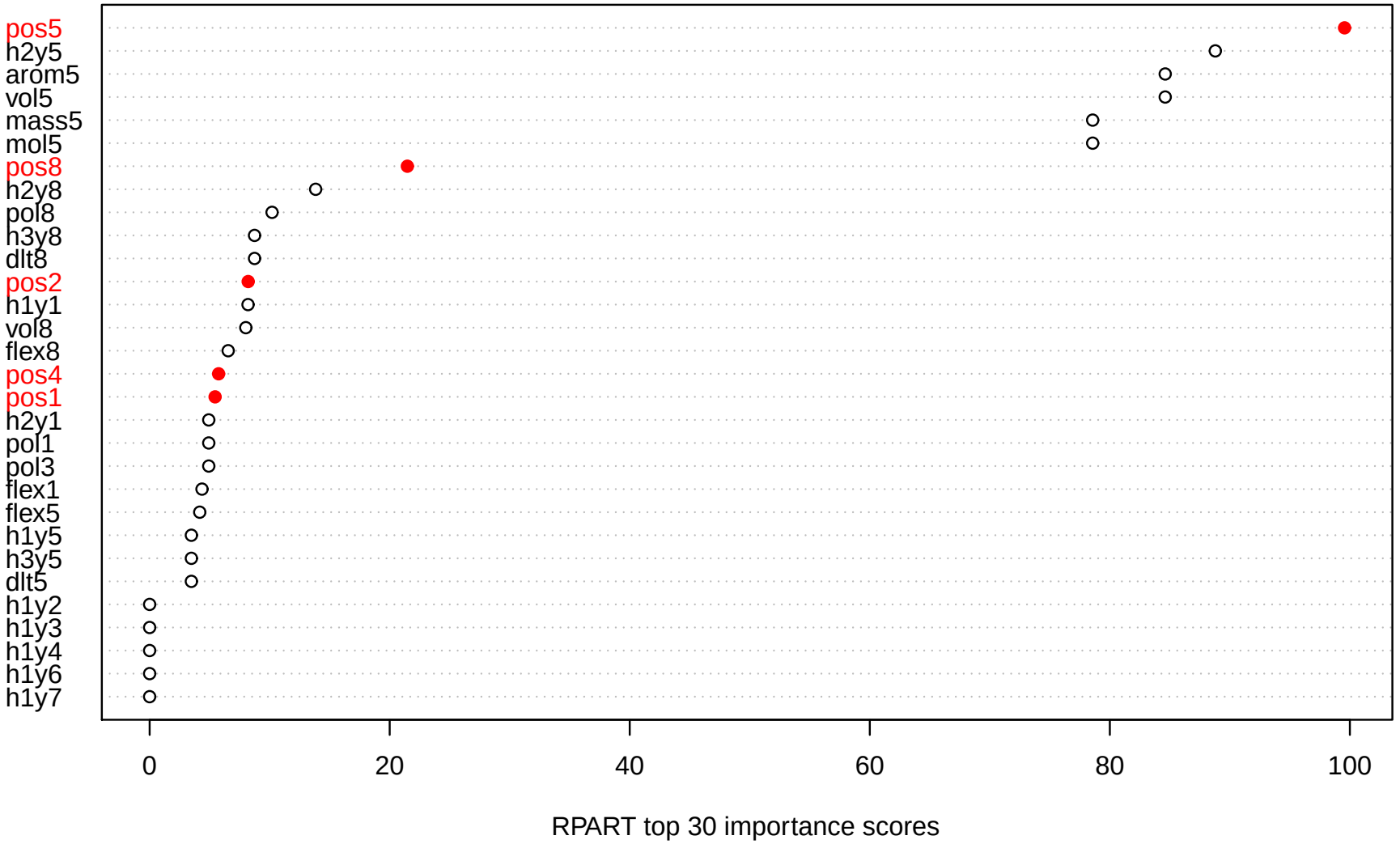
$$100 M(x_k) / \max_m M(x_m)$$

- The more obvious alternative measure

$$\sum_{t \in T} \Delta i(s_k^*, t)$$

  is not used because it was found to be inferior

RPART top 30 importance scores for peptide data



GUIDE classification

1. Select the most significant X variable to split a node
2. Find the split point or split set for X to minimize the Gini index
3. Recursively repeat steps 1 and 2 until too few observations in each node
4. Use the CART method to prune the tree to minimize CV estimate of misclassification cost

GUIDE hierarchical split variable selection

Level 1: Marginal tests. Cross-tab each X with Y, adding a "missing" level if any. Select most significant X if its p-value is below Bonferroni threshold. Otherwise, go to level 2.

Level 2: Interaction tests. For each pair (i, j), divide (X_i, X_j)-space into several regions. Cross-tab regions with Y. Select most significant (X_i, X_j) if its p-value is below Bonferroni threshold. Otherwise go to level 3.

Level 3. Linear split. For each pair of ordinal (X_i, X_j), apply marginal test to its largest linear discriminant coord. Select most significant (X_i, X_j) if its p-value is below Bonferroni threshold. Otherwise, select most significant X from level 1.

Chi-squared tests for peptide data

Pos1 (X^2 = 185.25, df = 17, p-value $<$ 2.2E-16)

bind	A	C	D	E	F	G	H	I	K	L	N	P	Q	R	S	T	V	Y
0	7	4	11	4	5	4	3	3	3	12	3	21	5	8	17	16	2	1
1	2	0	0	0	0	0	0	0	0	0	1	0	2	0	132	26	1	17

Pos2 (X^2 = 111.3, df = 19, p-value = 4.6E-15)

bind	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
0	9	3	12	5	3	3	3	3	1	12	2	0	20	2	11	14	15	6	1	4
1	34	0	0	0	0	12	0	2	0	10	8	2	6	0	1	70	31	4	0	1

Pos3 (X^2 = 114.35, df = 19, p-value = 1.3E-15)

bind	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
0	10	3	6	4	3	5	3	5	1	14	2	0	17	4	11	11	14	5	5	6
1	4	1	2	1	8	1	1	6	0	10	3	1	95	0	1	2	1	3	9	32

Pos4 (X^2 = 51.475, df = 19, p-value = 7.9E-05)

bind	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
0	9	0	3	4	3	8	5	7	4	14	3	2	20	3	7	12	13	6	2	4
1	27	1	4	8	1	1	6	7	1	2	0	4	39	16	7	28	22	4	0	3

Chi-squared tests (cont'd.)

Pos5 (X^2 = 211.5, df = 19, p-value $<$ 2.2E-16)

bind	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
0	7	1	11	5	3	5	6	3	6	12	2	1	17	2	13	12	8	8	2	5
1	2	0	0	0	73	1	3	2	0	9	11	1	3	0	0	0	0	1	0	75

Pos6 (X^2 = 66.888, df = 19, p-value = 3.0E-07)

bind	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
0	4	2	9	4	4	12	3	1	3	14	3	3	15	2	7	12	11	9	3	8
1	29	0	4	1	1	8	12	2	0	3	1	8	31	6	14	32	20	2	1	6

Pos7 (X^2 = 84.966, df = 18, p-value = 1.1E-10)

bind	A	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
0	6	14	2	3	8	5	2	1	9	4	1	18	6	7	16	17	3	3	4
1	16	1	0	0	4	4	3	0	7	1	3	51	1	1	70	16	3	0	0

Pos8 (X^2 = 185.69, df = 19, p-value $<$ 2.2E-16)

bind	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V
0	10	2	12	2	1	7	5	0	1	10	3	1	23	6	9	15	13	6
1	1	0	2	0	7	1	4	23	0	100	26	0	0	0	1	5	5	4

Level 1 marginal tests for ordinal X

1. Define $k = 3$ if $N(t) < 20J_t$ and $k = 4$ otherwise.
2. Divide the range of X into k intervals at the i/k quantiles, $i = 1, 2, \ldots, k-1$,. Add one "interval" for missing values, if any.
3. Form a contingency table with class values as rows and intervals as columns.
4. Let ν be df of the table after deleting empty rows and columns. Compute the chi-squared statistic χ^2_ν for testing independence.
5. Use Wilson-Hilferty (1931) approximation to convert χ^2_ν to $W_M \sim \chi^2_1$.

Level 1 marginal tests for categorical X

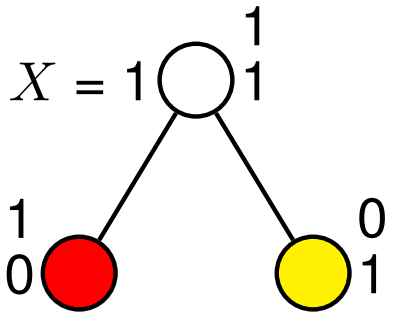
1. Let k be the number of distinct values of X.
2. Form a $J_t \times k$ contingency table with the class values as rows and the values of X as columns.
3. Proceed as for ordinal X.

Chi-squared tests with and without missing values (expected counts in parentheses)

	$X = 1$	$X = 2$	$X =$ NA
$Y = 1$	1 (0.5)	0 (0.5)	m (m)
$Y = 2$	0 (0.5)	1 (0.5)	m (m)
	χ^2 = 2 with 2 df		

	$X = 1$	$X = 2$
$Y = 1$	1 (0.5)	0 (0.5)
$Y = 2$	0 (0.5)	1 (0.5)
	χ^2 = 2 with 1 df	

CART will like this split because it reduces node impurity to 0:



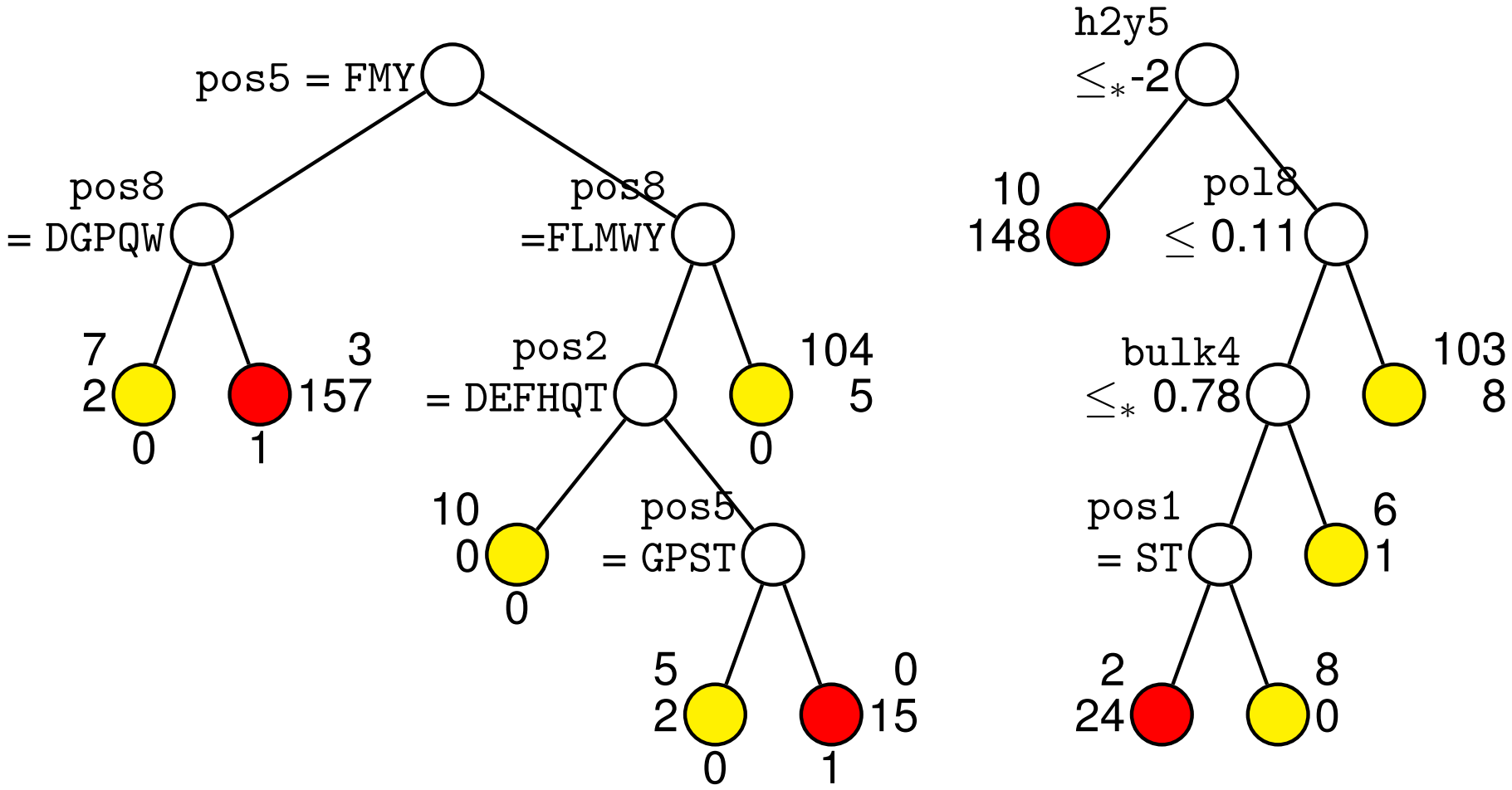
Modified Wilson-Hilferty (1931) approximation

- Given X^2 and $\nu > 1$, define

$$
\begin{aligned}
W_1 &= \left\{\sqrt{2X^2} - \sqrt{2\nu - 1} + 1\right\}^2 /2 \\
W_2 &= \max\left(0, \left[\frac{7}{9} + \sqrt{\nu}\left\{\left(\frac{X^2}{\nu}\right)^{1/3} - 1 + \frac{2}{9\nu}\right\}\right]^3\right) \\
W &= \begin{cases} W_2 & \text{if } X^2 < \nu + 10\sqrt{2\nu} \\ (W_1 + W_2)/2 & \text{if } X^2 \geq \nu + 10\sqrt{2\nu} \text{ and } W_2 < X^2 \\ W_1 & \text{otherwise.} \end{cases}
\end{aligned}
$$

- Let $F_\nu(x)$ be cdf of χ^2_ν distribution with ν df. Then $F_1(W) \approx F_\nu(X^2)$.

RPART (left) and GUIDE (right) trees for peptide data using all 112 vars



GUIDE variable importance scores

- Let $W(X_i, t)$ be the Wilson-Hilferty chi-squared value of X_i at node t
- If X_i is constant at t, define $W(X_i, t) = 1$
- Let $n(t)$ be sample size at t
- The unscaled importance score of X_i is

$$\sum_t n(t) W(X_i, t)$$

  where the sum is over the intermediate nodes

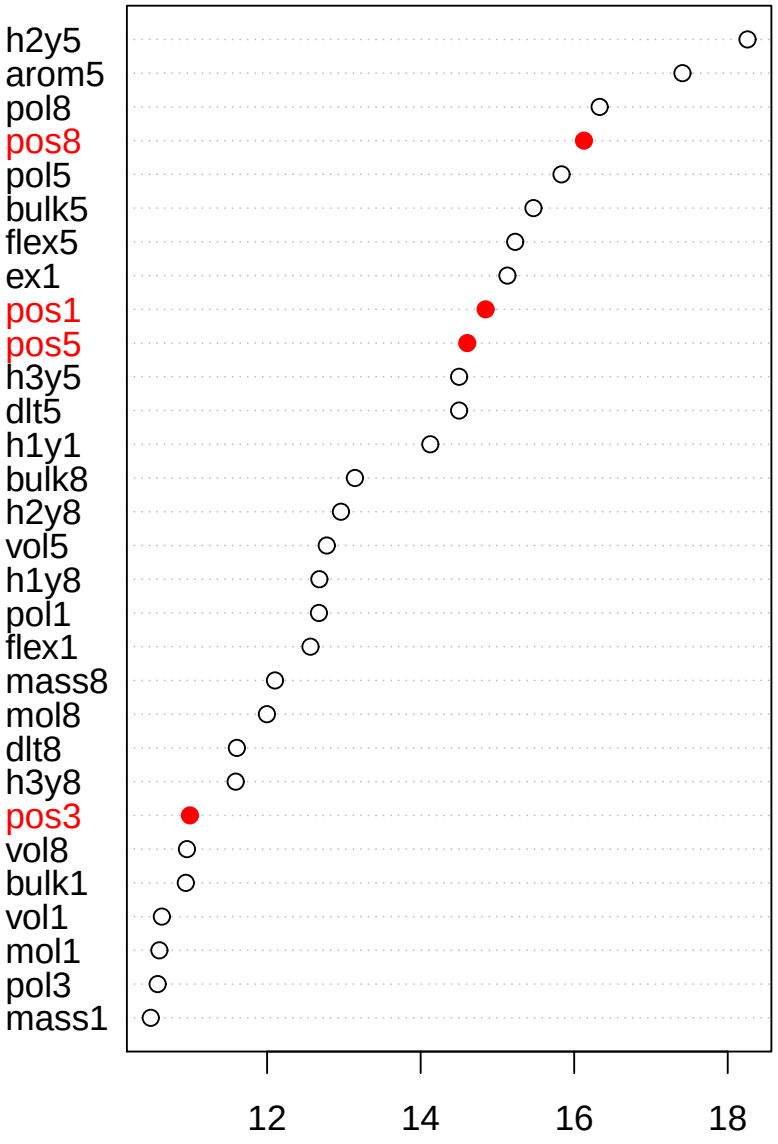
Null distribution of unscaled importance scores

- If X_i is independent of Y, then the scores
  - are linear combinations of approx. independent chi-squared variables
  - can be approximated by a scaled chi-squared distribution (Satterthwaite, 1946)
- Threshold τ for separating important from unimportant variables is the upper-α quantile of the corresponding chi-squared distribution, where

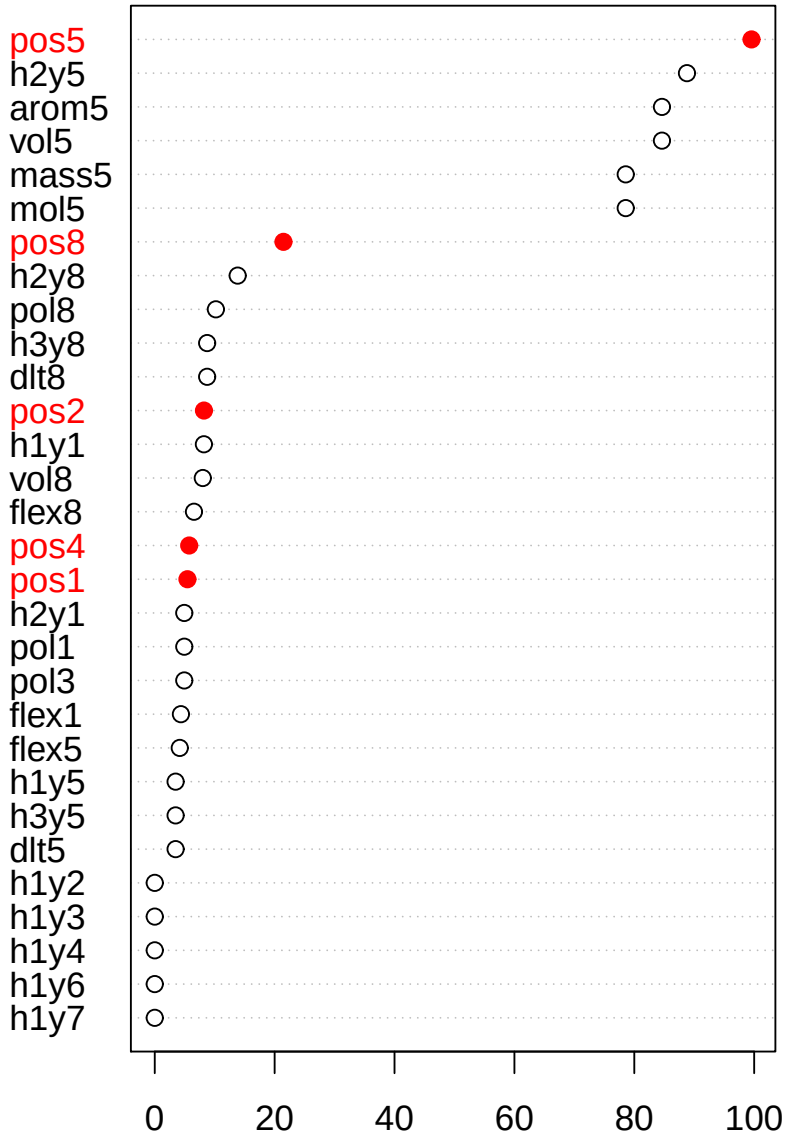
$$\alpha = k_0/K$$

  and k_0 is the expected number of unimportant variables under the null distribution (k_0 = 2 for classification, 1 for regression)
- Scaled importance scores are IMP$(X_i) = \tau^{-1} \sum_t n(t) W(X_i, t)$
- Hence those X_i with IMP$(X_i) < 1$ are considered unimportant

Top 30 variables by GUIDE and RPART



GUIDE top 30 scores



RPART top 30 scores

Importance scoring of GUIDE vs Random forest

- GUIDE scores are weighted sums of chi-squared statistics with node sizes as weights (Loh, 2012)
- GUIDE produces a threshold score for identifying noise variables
- RF scores are differences in accuracy between permuting and not permuting variable one at a time
- RF does not have a threshold score for identifying noise variables
- RF (as implemented in R) requires prior imputation of missing values
- GUIDE does not require missing value imputation

Low birth weight data: Missing values and highly unbalanced classes

- Data from 2006 CDC Natality Public Use File
- Birth weight and 33 other variables for 4.3 million births in U.S. in 2006
- 8.26% have low birth weight
- 61% of subjects have missing values
- Goal: what factors and how are they predictive of low birth weight?

Variables for birth data (#missing in bold)

apncu	Adequacy of prenatal care utilization: 1=inadequate, 2=intermediate, 3=adequate, 4=adequate+, 5=unknown [**2,143,286**]
attend	Attendant: 1=MD, 2=Doctor of Osteopathy, 3=Certified nurse midwife, 4=other midwife, 5=other, 9=unknown/not stated [**2,232**]
bfacil	Birth place: 1=In hospital, 2=not in hosp., 3=unknown [**355**]
cigs	Cigarettes per day: 0-97, 98=98 or more, 99=unknown [**2,104,492**]
dmeth	Delivery method: 1=vaginal (excluding vaginal after previous C-section), 2=vaginal after previous C-section, 3=primary C-section, 4=repeat C-section, 9=not stated [**14,700**]
drinks	Drinks: 0=non drinker, 1=1, 2=2, 3=3-4, 4=5 or more drinks/week, 5=unknown/not stated [**2,086,856**]
fagecomb	Father's combined age: 10-98 age in years, 99=unknown/not stated [**611,546**]

frace	Father's race: 1=White, 2=Black, 3=Amer Indian/Alaskan native, 4=Asian/Pacific Islander, 9=unknown/not stated [**775,450**]
fracehisp	Father's race/Hispanic origin: 1=Mexican, 2=Puerto Rican, 3=Cuban, 4=Central or S. American, 5=Other and unknown Hispanic, 6=non-Hispanic white, 7=non-Hispanic black, 8=non-Hispanic other races, 9=unknown/not stated [**637,910**]
gest	Gestation: 1=under 20 wks, 2=20-27, 3=28-31, 4=32-33, 5=34-36, 6=37-39, 7=40, 8=41, 9=42 and over, 10=unknown [**24,628**]
lowbwt	Low birth weight indicator (bwt $<$ 2500)
mage	Mother's age: 12=10-12, 13=13,. . . , 32=32 years
mar	Mother's marital status: 1=yes, 2=no, 9=unknown/not stated
meduc	Mother's education: 1=0-8, 2=9-11, 3=12, 4=13-15, 5=16 years and over, 6=not stated [**2,103,502**]
mpcb	Month prenatal care began: 1=1st, 2=2nd, 3=3rd trimester, 4=no prenatal care, 5=unknown/not stated [**1,562,097**]

mrace	Mother's race: 1=White, 2=Black, 3=Amer Indian/Alaskan native, 4=Asian/Pacific Islander
mracehisp	Mother's race/Hispanic origin: 1=Mexican, 2=Puerto Rican, 3=Cuban, 4=Central or S. American, 5=Other and unknown Hispanic, 6=non-Hispanic white, 7=non-Hispanic black, 8=non-Hispanic other races, 9=unknown/not stated [**30,147**]
plural	Plurality: 1=single, 2=twin, 3=triplet, 4=quadruplet, 5$\geq$quintuplet
previs	Number prenatal visits: 0-49, 99=unknown/not stated [**134,567**]
restatus	Residence status: 1=resident, 2=intrastate nonres., 3=interstate nonres., 4=foreign
sex	Sex of infant: M or F
tbo	Total birth order: 1-7, 8=total birth order of 8 or more, 9=unknown/not stated [**30,508**]
wtgain	Weight gain: 0-97, 98=98+lbs, 99=unknown/not stated [**752,049**]

Logistic regression

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	3.7275166	0.0424679	87.772	< 2e-16	***
bfacil2	-0.1868176	0.0762437	-2.450	0.01427	*
mage	-0.0040537	0.0010267	-3.948	7.87e-05	***
restatus2	0.1148237	0.0080626	14.241	< 2e-16	***
restatus3	0.2299749	0.0210306	10.935	< 2e-16	***
restatus4	0.1785475	0.3702205	0.482	0.62961	
mrace2	0.1497132	0.0526596	2.843	0.00447	**
mrace3	-0.1899855	0.0914809	-2.077	0.03782	*
mrace4	0.1541109	0.0877779	1.756	0.07914	.
mracehisp2	0.3841715	0.0367728	10.447	< 2e-16	***
mracehisp3	0.1953111	0.1054410	1.852	0.06398	.
mracehisp4	-0.0085673	0.0361126	-0.237	0.81247	
mracehisp5	0.2423411	0.0444570	5.451	5.00e-08	***
mracehisp6	0.1480889	0.0226697	6.532	6.47e-11	***
mracehisp7	0.4128809	0.0563987	7.321	2.47e-13	***
mracehisp8	0.2197294	0.0876985	2.506	0.01223	*

```
mar2          0.1845052  0.0095344   19.351  < 2e-16 ***
meduc        -0.0666695  0.0041055  -16.239  < 2e-16 ***
fagecomb     -0.0024821  0.0008000   -3.102  0.00192 **
frace2        0.0496725  0.0530778    0.936  0.34935
frace3       -0.2160319  0.0967624   -2.233  0.02558 *
frace4        0.0574721  0.0941127    0.611  0.54142
fracehisp2    0.1689273  0.0370073    4.565 5.00e-06 ***
fracehisp3    0.0659272  0.0985604    0.669  0.50356
fracehisp4    0.0318342  0.0361529    0.881  0.37857
fracehisp5    0.2080938  0.0458732    4.536 5.73e-06 ***
fracehisp6   -0.0245387  0.0220995   -1.110  0.26684
fracehisp7    0.0793392  0.0565907    1.402  0.16092
fracehisp8    0.2031302  0.0935563    2.171  0.02992 *
tbo          -0.1034129  0.0028329  -36.504  < 2e-16 ***
mpcb2        -0.0399053  0.0126069   -3.165  0.00155 **
mpcb3        -0.0876767  0.0290933   -3.014  0.00258 **
mpcb4         0.0743985  0.0441908    1.684  0.09226 .
previs       -0.0510736  0.0012691  -40.243  < 2e-16 ***
wtgain       -0.0244573  0.0002887  -84.730  < 2e-16 ***
```

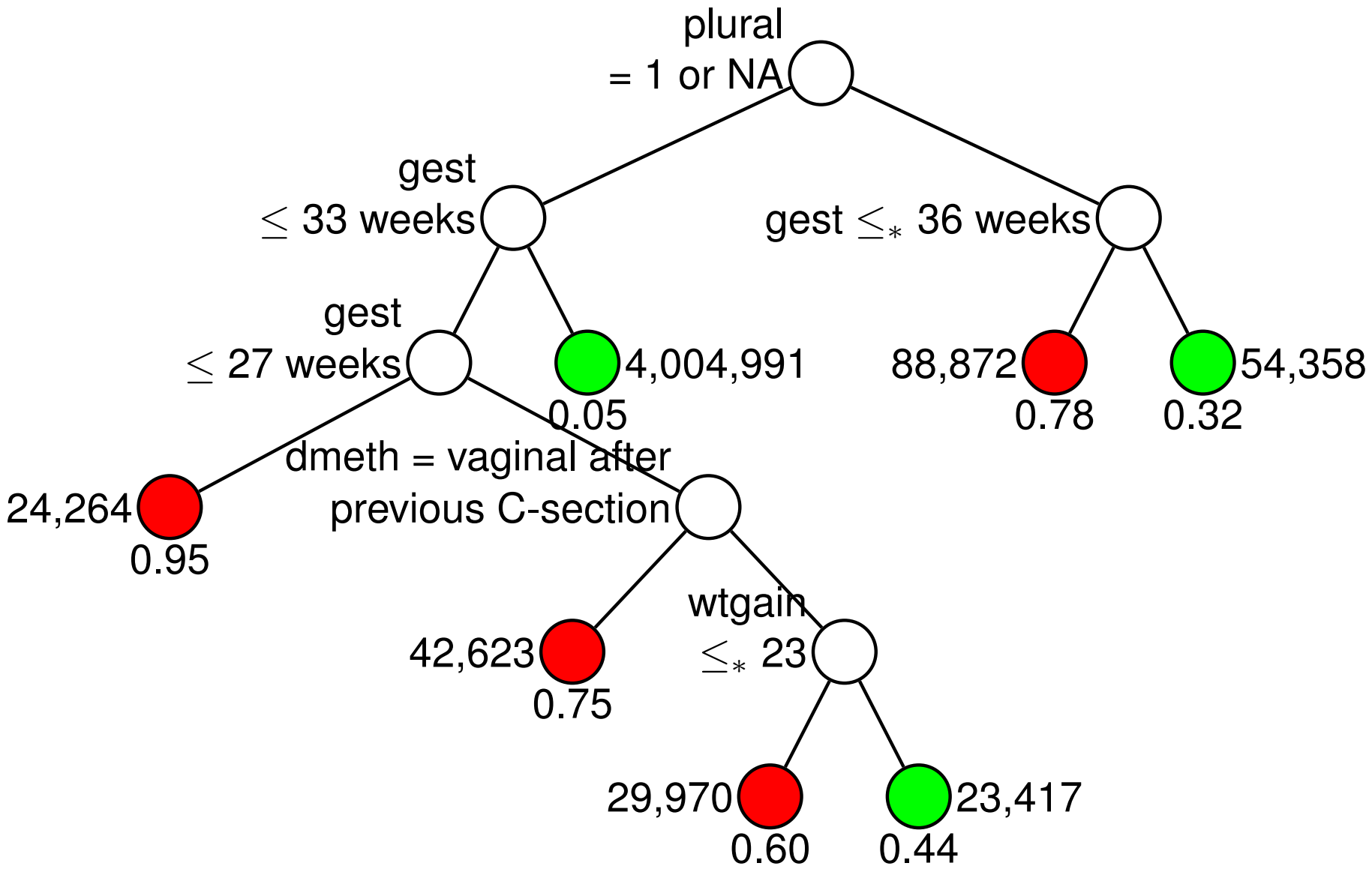
```
apncu        0.1474178  0.0058539   25.183  < 2e-16 ***  <===
cigs         0.0524019  0.0010045   52.169  < 2e-16 ***
drinks       0.1899323  0.0239897    7.917 2.43e-15 ***
dmeth2       0.3609829  0.0078543   45.960  < 2e-16 ***
attend2     -0.1522606  0.0175571   -8.672  < 2e-16 ***
attend3     -0.2852335  0.0180681  -15.787  < 2e-16 ***
attend4     -0.6124455  0.1095952   -5.588 2.29e-08 ***
attend5      0.2981869  0.0690817    4.316 1.59e-05 ***
plural       2.5658474  0.0120097  213.648  < 2e-16 ***
sexM        -0.3283656  0.0073779  -44.507  < 2e-16 ***
gest        -1.3641028  0.0046431 -293.794  < 2e-16 ***
```

Why logistic model cannot be safely interpreted

Model is based on complete subset of **39% of data**

1. Therefore the results require these assumptions:
   (a) Data are **randomly missing**
   (b) **No interactions** on log-odds scale
2. Even if these assumptions are valid:
   (a) Model coefficients **depend** on presence/absence of other variables
   (b) Significance of coefficients are similarly dependent
   (c) Signs of coefficients may be counter-intuitive; see apncu

GUIDE tree with estimated priors



Pr(low birth wt) below nodes, sample sizes beside nodes

Chi-squared tests

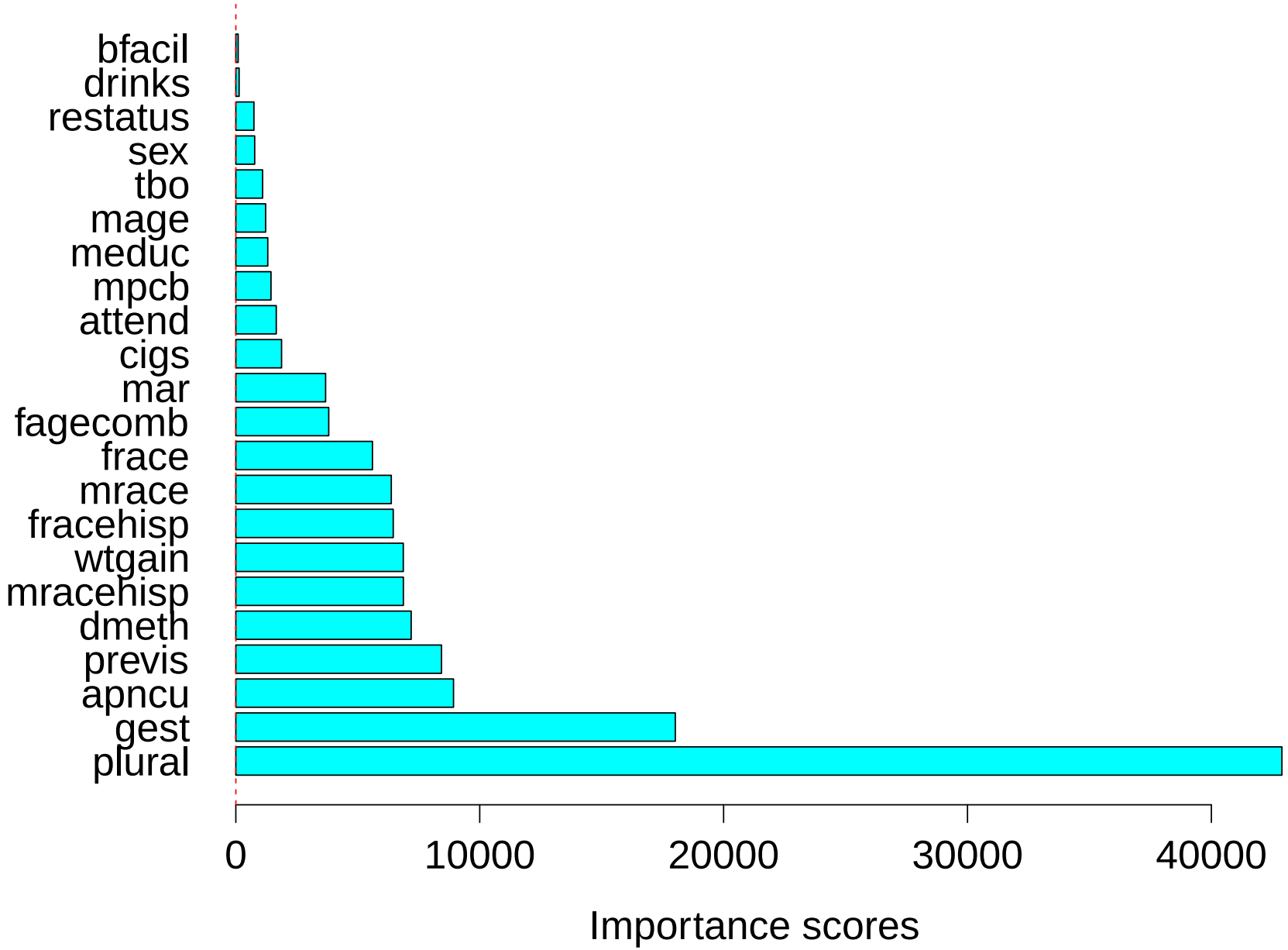
gest (X^2 = 118090, df = 3, p-value $<$ 2.2E-16)

	≤ 6	(6, 7]	> 7	NA
low	325911	11970	11651	2881
not low	2522022	789198	583115	21747

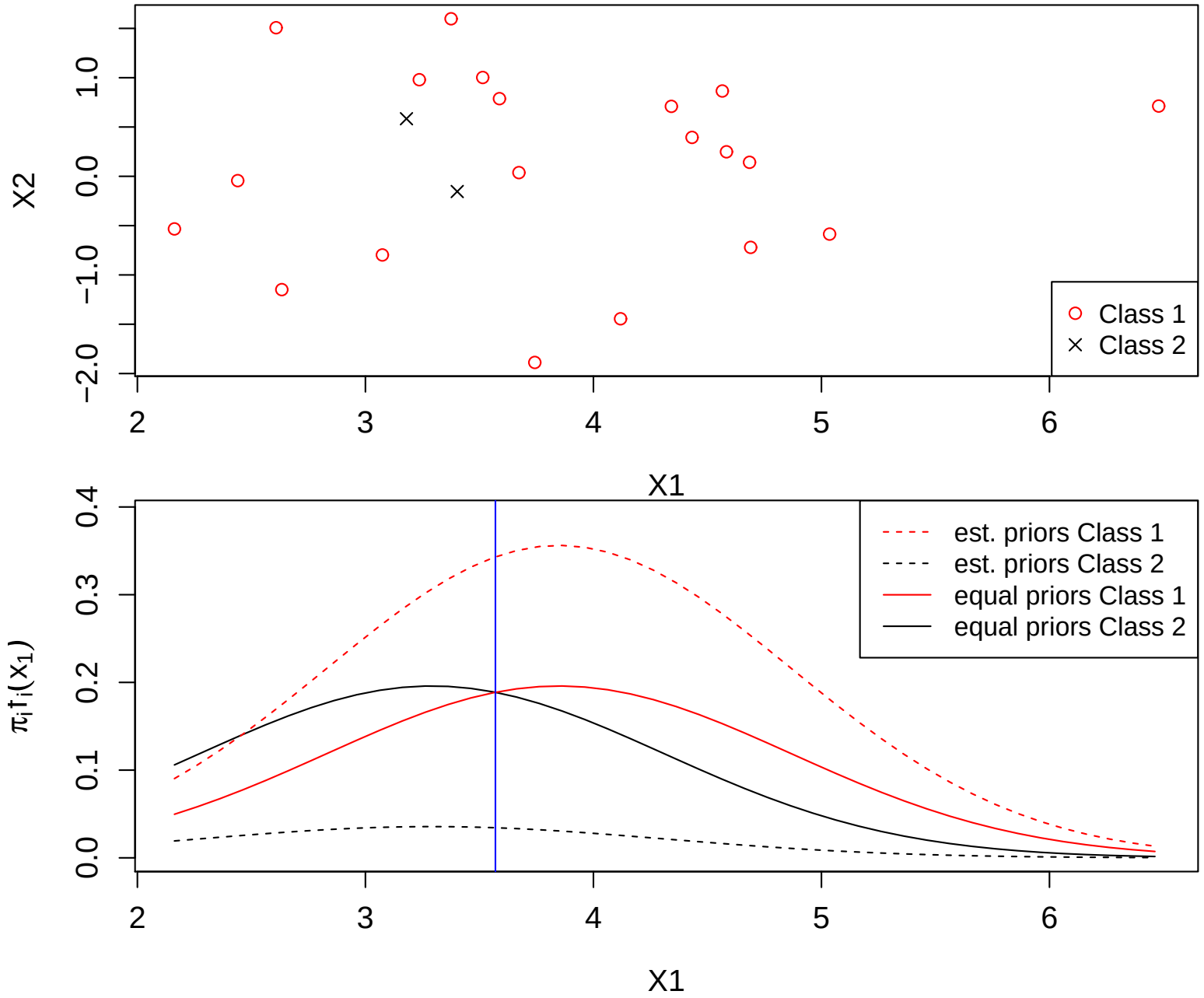
plural (X^2 = 508590, df = 1, p-value $<$ 2.2E-16)

	≤ 1	> 1
low	267563	84850
not low	3857702	58380

Importance scores for low birth weight data



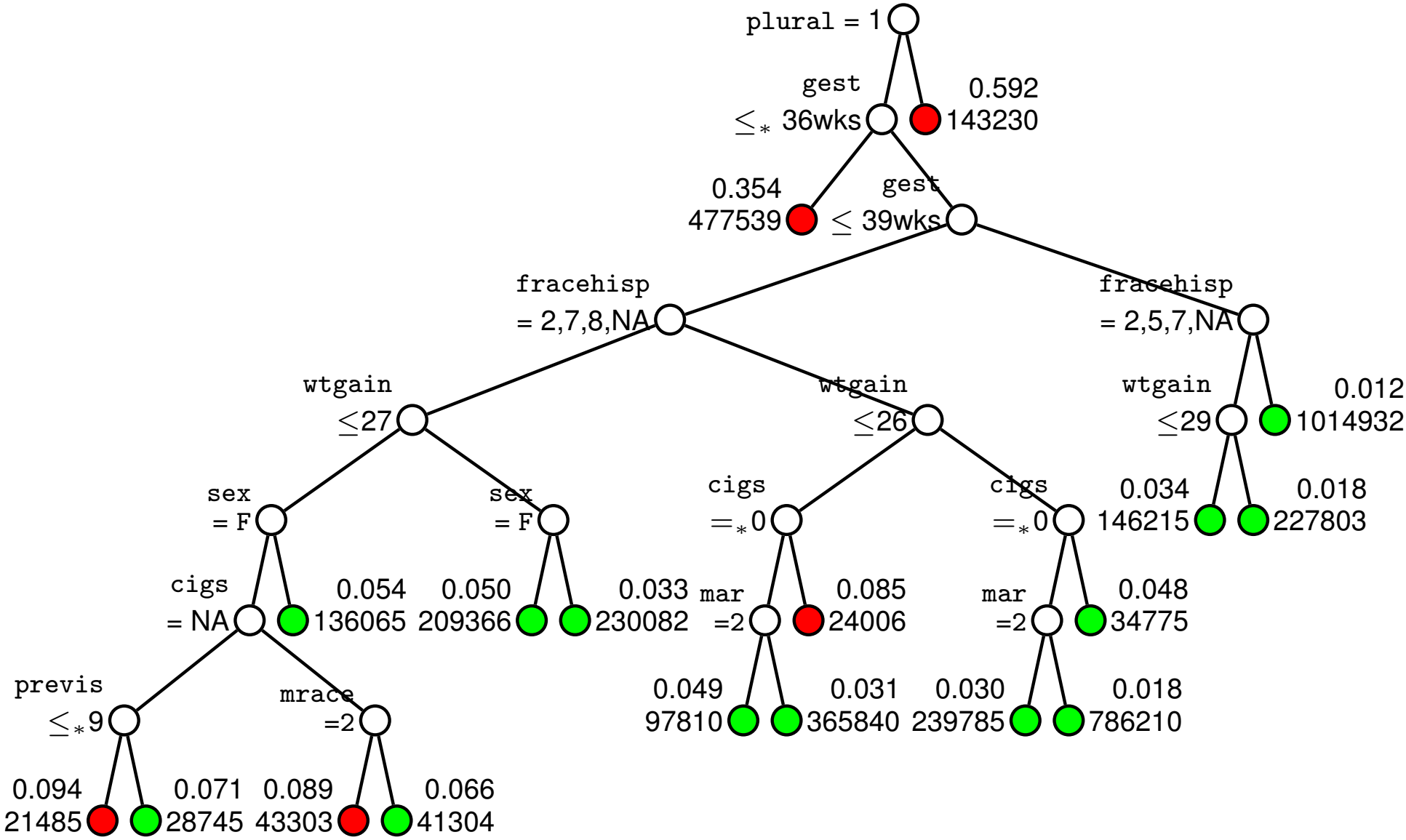
Estimated vs equal priors



Estimated vs equal priors (cont'd.)

- Let $f_j(x)$ be the pop. density and π_j the prior prob. of Class j ($j = 1, 2$)
- If π_j and f_j are given, x is classified in Class 1 if $\pi_1 f_1(x) > \pi_2 f_2(x)$
- If π_j and f_j are estimated, x is classified in Class 1 if $\hat{\pi}_1 \hat{f}_1(x) > \hat{\pi}_2 \hat{f}_2(x)$
- Let N_j = total sample size and $n_j(t)$ = sample size in node t of Class j
- Then $\hat{f}_j(x) = n_j/N_j$ for $x \in t$
- If π_j is estimated from the sample, then $\hat{\pi}_j = N_j/(N_1 + N_2)$ and x is predicted to be in Class 1 if
  $N_1(N_1 + N_2)^{-1} \times n_1/N_1 > N_2(N_1 + N_2)^{-1} \times n_2/N_2$, i.e., $n_1 > n_2$
- If priors are equal, $\pi_j = 1/2$ and x is predicted to be in Class 1 if
  $(1/2)(n_1/N_1) > (1/2)(n_2/N_2)$, i.e., $n_1/n_2 > N_1/N_2$

Equal priors: for finer separation of rare events

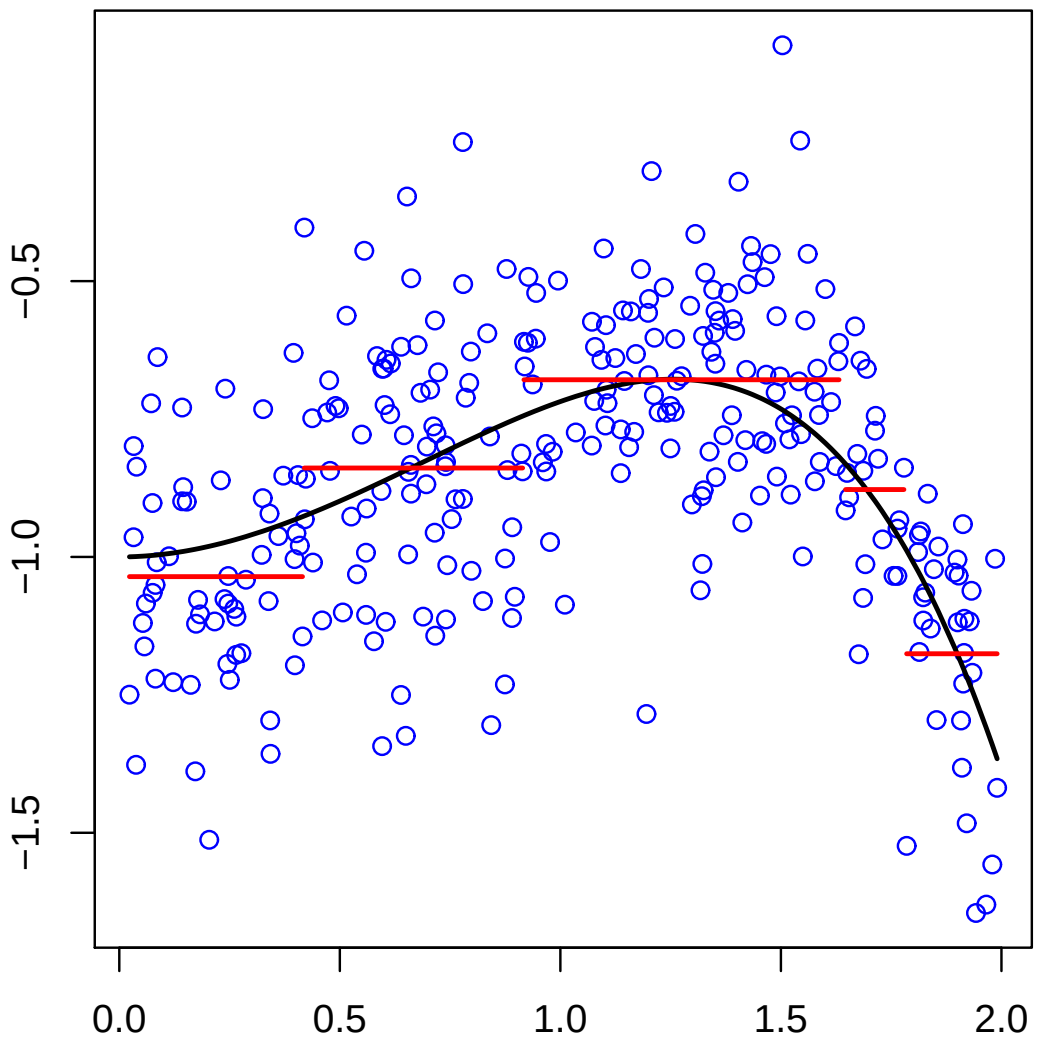


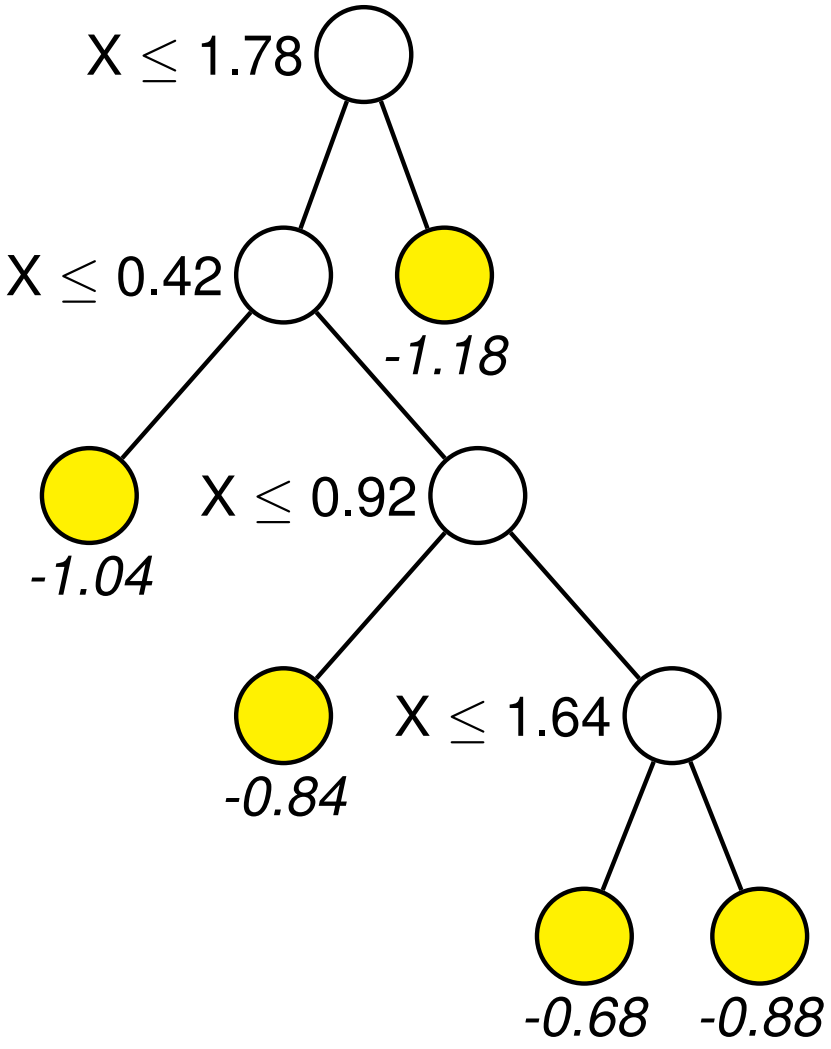
P(low birthwt) and sample size beside nodes

CART regression

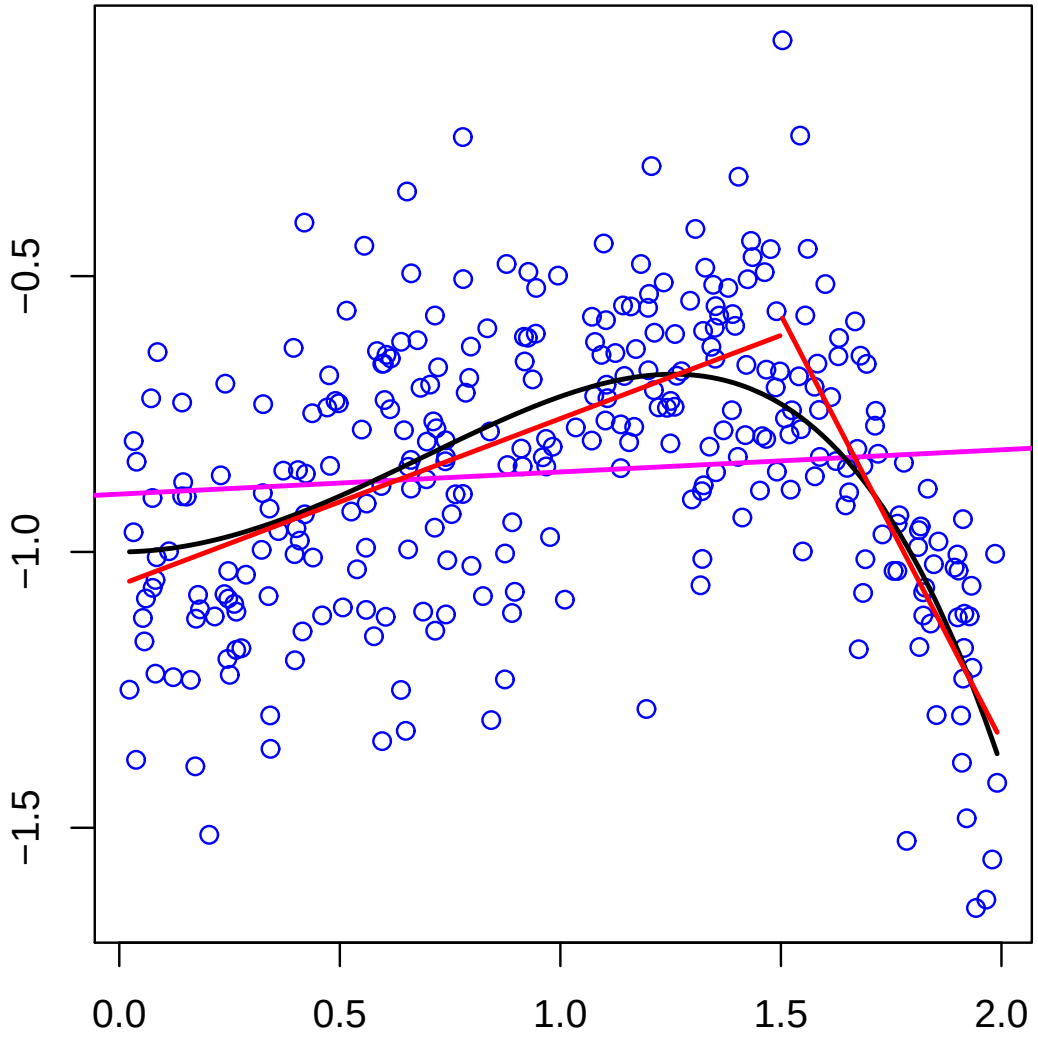
- Fit a constant $\bar{y}$ to each node
- Use residual sum of squares as node impurity and error measure
- Everything else the same as in CART classification

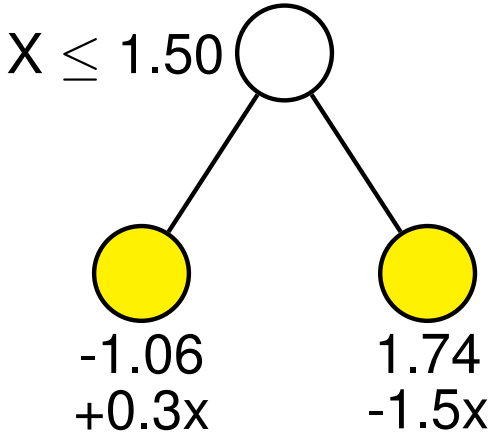
Piecewise-constant regression model





Piecewise-linear regression model





GUIDE regression tree models

- Piecewise constant, multiple linear, stepwise linear, best simple polynomial, and best simple ANCOVA
- Least squares, least median of squares, quantile, Poisson, proportional hazards (with censoring), multi-response, and longitudinal data
- Predictor variables can be used for model fitting only, splitting only, or both
- Unbiased variable selection (bootstrap bias correction for linear models)
- Trees pruned with CART method

Variable roles in GUIDE description files

D: Dependent variable (least-squares, least median of squares, quantile, Poisson, multi-response and longitudinal) or death indicator (proportional hazards)

N: Numerically ordered variable used for fitting and splitting

F: Numerically ordered variable used for fitting only

S: Numerically ordered variable used for splitting only

C: Categorical variable used for splitting only

B: Categorical variable for both for splitting and fitting via dummies

R: Treatment categorical variable for fitting only

W: Weight variable for weighted least squares and case exclusion

T: Survival or observation time (prop. hazards or longitudinal data)

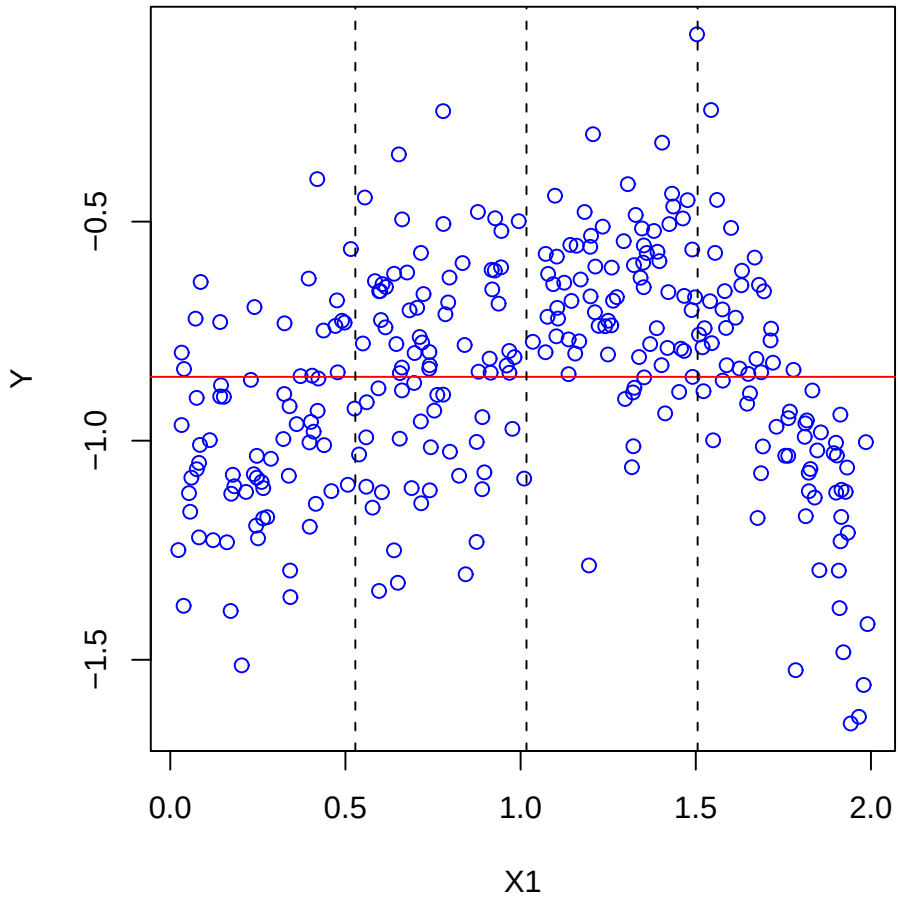
Z: Offset variable (Poisson regression)

X: Excluded variable

GUIDE variable selection for regression

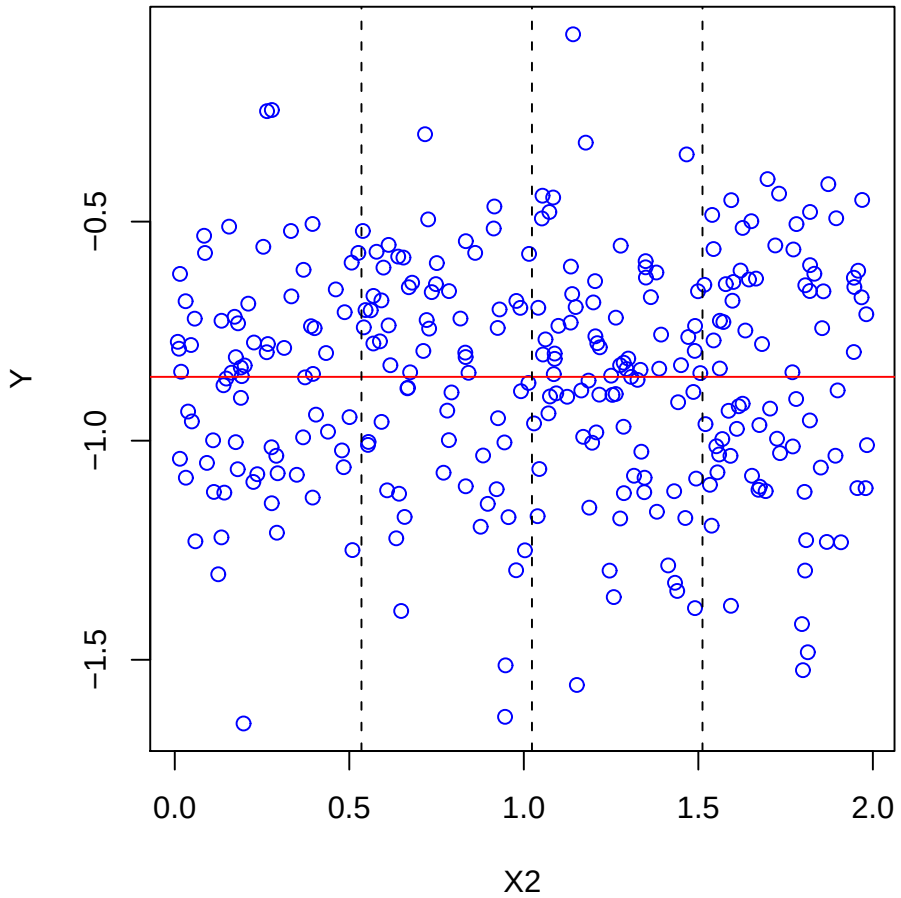
1. Fit a model to the data in the node and obtain the residuals
2. Define a “class” variable that equals +1 if residual is positive, -1 otherwise
3. Follow GUIDE classification procedure to select a variable to split node

Split variable selection based on residual patterns



Pos. res.	18	49	68	27
Neg. res.	52	31	10	45

χ^2_3 = 66.7, p = 2×10^{-14}



Pos. res.	37	41	45	39
Neg. res.	34	28	39	37

χ^2_3 = 1.14, p = 0.77

Split selection for piecewise-constant model

1. Fit a constant model to the data in the node and obtain residuals
2. Convert all `n`-variables to `s` type.
3. Do the following curvature tests:
   (a) For each `s`-variable X:
      i. Divide cases into four groups at the X quartiles
      ii. If X has missing values, add a "missing value" group
      iii. Cross-tab with signs of residuals as rows and groups as columns
      iv. Let ν be the df. Convert χ^2_ν to χ^2_1 using the Wilson-Hilferty approximation
   (b) Do the same for each `c`-variable, using its categories to form groups
4. If variable X^* with largest χ^2_1 value is significant at $\alpha_1 = 0.05/K$, find split on X^* that minimizes sum of deviances in subnodes

5. Otherwise do the following interaction tests:

   (a) For each pair (X_i, X_j) of s-variables:

      i. Divide X_i values into 2 groups at sample mean
      ii. Add 3rd group for missing values if necessary
      iii. Do the same for X_j
      iv. Form (X_i, X_j) groups from Cartesian product of X_i and X_j groups
      v. Cross-tab signs of residuals (rows) and (X_i, X_j) groups (columns)
      vi. Compute the Wilson-Hilferty χ^2_1-value

   (b) Do the same for each pair (X_i, X_j) of c-variables, using the (X_i, X_j) category-pairs to form the columns

(c) For each pair (X_i, X_j), where X_i is s and X_j is c:

i. Let the mean of X_i in the node be m_i and the values taken by X_j be $c_1, \ldots, c_j$

ii. Divide the (X_i, X_j) space into sets
$\{X_i = \text{NA}, X_j = c_1\}, \ldots, \{X_i = \text{NA}, X_j = c_j\}$,
$\{X_i \leq m_i, X_j = c_1\}, \ldots, \{X_i \leq m_i, X_j = c_j\}$,
$\{X_i > m_i, X_j = c_1\}, \ldots, \{X_i > m_i, X_j = c_j\}$

iii. Cross-tab signs of residuals (rows) and (X_i, X_j) sets (columns)

iv. Compute the Wilson-Hilferty χ^2_1-value

6. If no interaction test is significant at $\alpha_2 = \min(\alpha_1, 0.2/[K(K-1)])$, split on the variable with the most significant curvature test

7. Otherwise, let (X_i, X_j) be the pair with most significant interaction
   (a) Split node into 4 subnodes in two steps by splitting first on one variable and then on the other
   (b) Using residual sign as class variable, find the best two-step split that minimizes total Gini index in the 4 subnodes
   (c) Split node with the first-step of the best two-step split

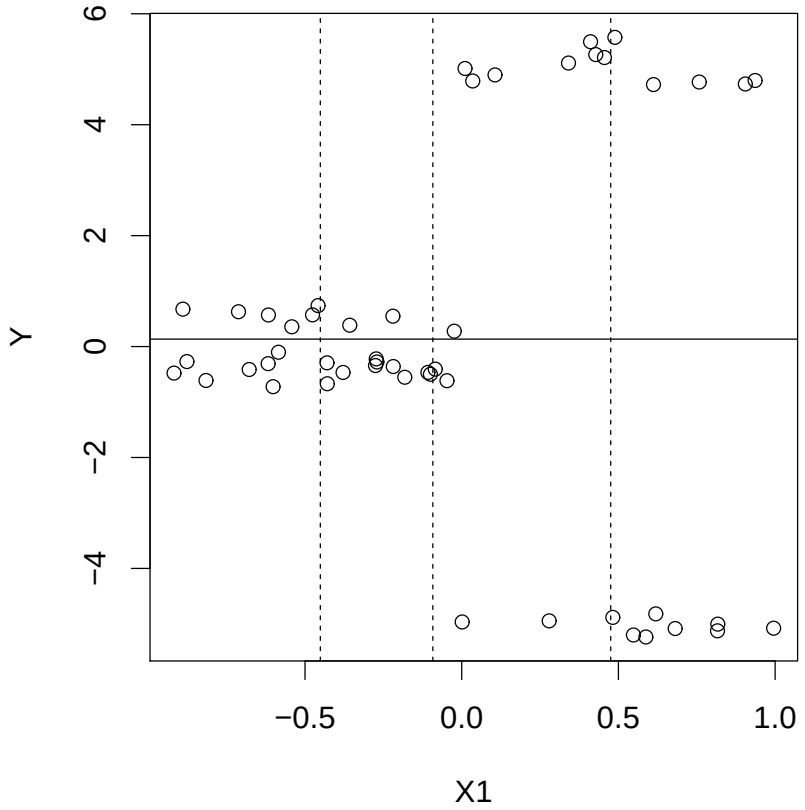
Motivation for interaction test

- Consider the interaction model $Y = f(X_1, X_2) + \epsilon$ where

$$f(X_1, X_2) = \begin{cases} -0.5, & X_1 \leq 0, X_2 \leq 0 \\ 0.5, & X_1 \leq 0, X_2 > 0 \\ 5.0, & X_1 > 0, X_2 \leq 0 \\ -5.0, & X_1 > 0, X_2 > 0 \end{cases}$$

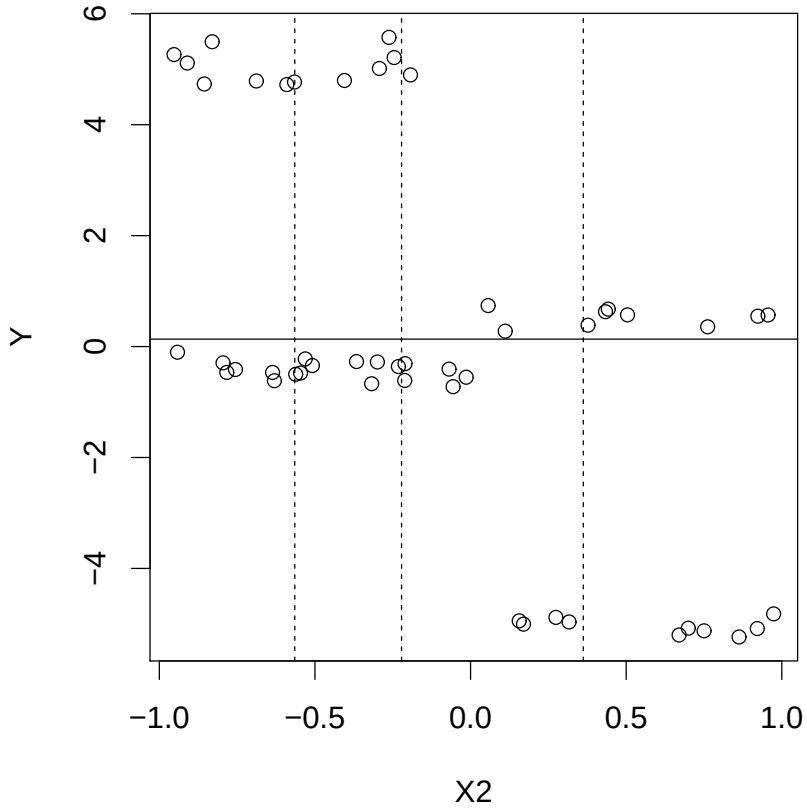
- $X_1, X_2, \ldots \sim U(-1, 1)$ and $\epsilon \sim N(0, 0.04)$
- X_1, $X_2, \ldots, \epsilon$ mutually independent
- Curvature tests are not significant but interaction test is significant

Y versus X_1



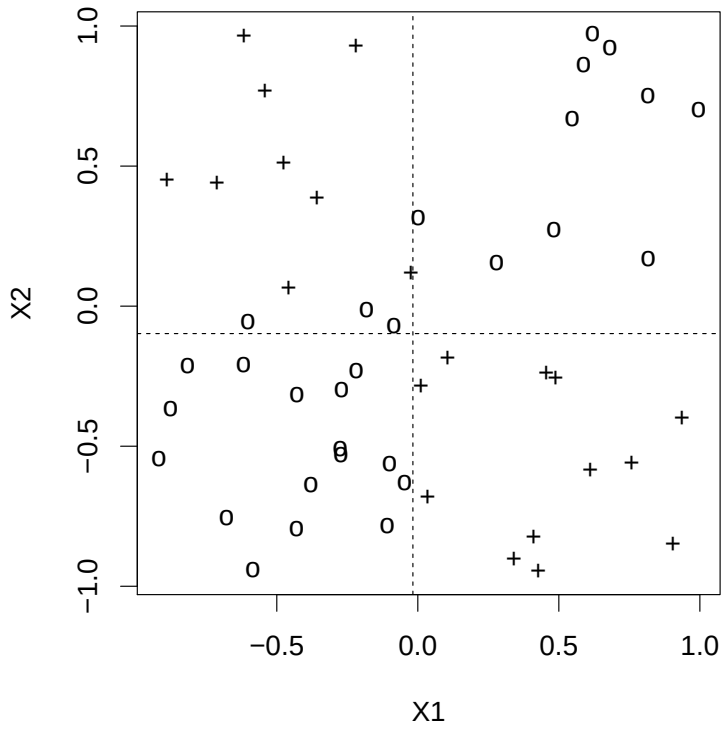
Resids	$X_1 \leq Q_1$	$Q_1 < X_1 \leq Q_2$	$Q_2 < X_1 \leq Q_3$	$X_1 > Q_3$
Positive	6	2	8	5
Negative	7	10	4	8
	$\chi^2_3 = 6.32$, $p = 0.10$, Wilson-Hilferty $\chi^2_1 = 2.71$			

Y versus X_2



Resids	$X_2 \leq Q_1$	$Q_1 < X_2 \leq Q_2$	$Q_2 < X_2 \leq Q_3$	$X_2 > Q_3$
Positive	7	4	3	7
Negative	6	8	9	6
	$\chi^2_3 = 3.29$, $p = 0.35$, Wilson-Hilferty $\chi^2_1 = 0.87$			

Residuals versus X_1 and X_2



	$X_2 \leq \bar{x}_2$		$X_2 > \bar{x}_2$	
Resids	$X_1 \leq \bar{x}_1$	$X_1 > \bar{x}_1$	$X_1 \leq \bar{x}_1$	$X_1 > \bar{x}_1$
Positive (+)	0	12	9	0
Negative (o)	16	0	3	10
	$\chi^2_3 = 40.76$, $p = 7 \times 10^{-9}$, Wilson-Hilferty $\chi^2_1 = 33.4$			

Naive variable selection for piecewise-linear model

1. Fit a linear model to the **n** and **f**-variables in the node and obtain residuals
2. For each **s** and **n**-variable X:
   (a) Divide cases into three or four groups
   (b) Cross-tab data with signs of residuals as rows and groups as columns
   (c) Compute a Wilson-Hilferty χ^2_1-value
3. Do the same for each **c**-variable, using categories to form columns of table
4. Select the variable with the largest χ^2_1 value

Selection bias in linear fit

- Residuals have zero sample correlation with **n**-predictors, but not with **c** and **s**-variables
- χ^2 tests for **n**-variables are less significant than those for **c** and **s**-variables

Simulation experiment

Predictors	Independent	Weakly dependent	Strongly dependent
X_1	T	T	T
X_2	W	W	W
X_3	Z	$T + W + Z$	$W + 0.1Z$
X_4	C_5	$\lfloor UC_{10}/2 \rfloor + 1$	$\lfloor UC_{10}/2 \rfloor + 1$
X_5	C_{10}	C_{10}	C_{10}

- C_k is k-category taking values $\{1, 2, \ldots, k\}$ with equal probabilities
- T is non-categorical uniformly distributed variable on $\{\pm 1, \pm 3\}$
- U is uniform $U(0, 1)$; W is exponential with mean 1; Z is $N(0, 1)$
- C_k, U, T, W, and Z are mutually independent
- $\lfloor . \rfloor$ is the greatest integer function

Selection probabilities for piecewise linear model when Y is independent of predictors (Case 1)

		Independent X_i		Weakly depend. X_i		Strongly depend. X_i	
X_i	Type	Uncorr.	Corr.	Uncorr.	Corr.	Uncorr.	Corr.
X_1	**n**	0	.178	0	.191	0	.178
X_2	**n**	0	.232	0	.206	0	.215
X_3	**n**	0	.200	0	.194	0	.197
X_4	**c**	.469	.181	.519	.200	.532	.214
X_5	**c**	.531	.209	.481	.209	.468	.196

Selection probabilities for piecewise linear model when Y is independent of predictors (Case 2)

		Independent X_i		Weakly depend. X_i		Strongly depend. X_i	
X_i	Type	Uncorr.	Corr.	Uncorr.	Corr.	Uncorr.	Corr.
X_1	**n**	0	.202	0	.181	0	.197
X_2	**n**	0	.217	0	.228	0	.214
X_3	**s**	.352	.203	.288	.134	.313	.121
X_4	**c**	.307	.178	.360	.238	.360	.256
X_5	**c**	.341	.200	.352	.219	.327	.212

Bootstrap estimation of bias correction factor

1. Let Y^* be an N-vector comprised of i.i.d. samples from the elements in Y
2. Fit a linear model to $(Y^*, X_1, \ldots, X_K)$, using only the **n** and **f**-variables
3. Let x^* denote bootstrap Wilson-Hilferty χ^2_1-values and define

   x^*_n = largest x^*-value among the **n**-variables

   x^*_s = largest x^*-value among the **s**-variables

   x^*_c = largest x^*-value among the **c**-variables

   x^*_{nn} = largest x^*-value from the interaction tests among pairs of **n**-variables

   x^*_{ss} = largest x^*-value from the interaction tests among pairs of **s**-variables

   x^*_{cc} = largest x^*-value from the interaction tests among pairs of **c**-variables

   x^*_{ns} = largest x^*-value from the interaction tests among **n**-**s** variable pairs

   x^*_{nc} = largest x^*-value from the interaction tests among **n**-**c** variable pairs

   x^*_{sc} = largest x^*-value from the interaction tests among **s**-**c** variable pairs
4. For each $r > 1$ over a finite set, apply the original variable selection

algorithm to the values $\{rx_n^*, x_s^*, x_c^*, rx_{nn}^*, x_{ss}^*, x_{cc}^*, x_{sc}^*, x_{nc}^*, x_{ns}^*\}$ and let $q(r)$ be the probability that an **n**-variable is selected

5. Plot $q(r)$ and linearly interpolate to find r_0 such that $q(r_0)$ equals the proportion of **n**-variables in the learning sample

6. Fit a model to the real data $(Y, X_1, \ldots, X_K)$, using only the **n** and **f**-variables and let $x_n, x_s, x_c, x_{nn}, x_{ns}, x_{nc}, x_{ss}, x_{sc}, x_{cc}$ be the corresponding Wilson-Hilferty χ_1^2-values

7. Use the values $\{r_0 x_n, x_s, x_c, r_0 x_{nn}, x_{ss}, x_{cc}, x_{sc}, x_{nc}, x_{ns}\}$ to select the split variable

GUIDE split set selection

X **is n or s:** Search over a systematic sample for a split of the form $X \leq c$ to minimize sum of deviances (or squared residuals)

X **is b or c:** Use following heuristic to save computational time —

1. Define $Z = 1$ for a positive residual and $Z = 0$ otherwise
2. For each split of the form $X \in A$, let t_L and t_R be the left and right nodes
3. Compute the (binomial) variances of Z in t_L and t_R
4. Use shortcut method in Breiman et al. p. 101 to find A that minimizes sum of Z variances in t_L and t_R

Default GUIDE method for missing values

1. For piecewise constant models, only cases complete in the **d**, **w**, **t**, and **z** variables are used for split selection and model fitting
2. For other models, missing predictor values are imputed with node means prior to model fitting
3. For split selection, a special category is created for missing categorical variables
4. For each split on a **n** or **s** variable, missing values are sent to the left or right node, depending on which one reduces node impurity more. The split that sends missing values to one node and nonmissing to the other is included.
5. Bootstrap bias-correction is performed for multiple linear models only

Advantages of GUIDE

- GUIDE model uses the variables directly—no transformations needed
- GUIDE fitted function can be displayed with fewer graphs
- GUIDE automatically clusters categories that have similar effects

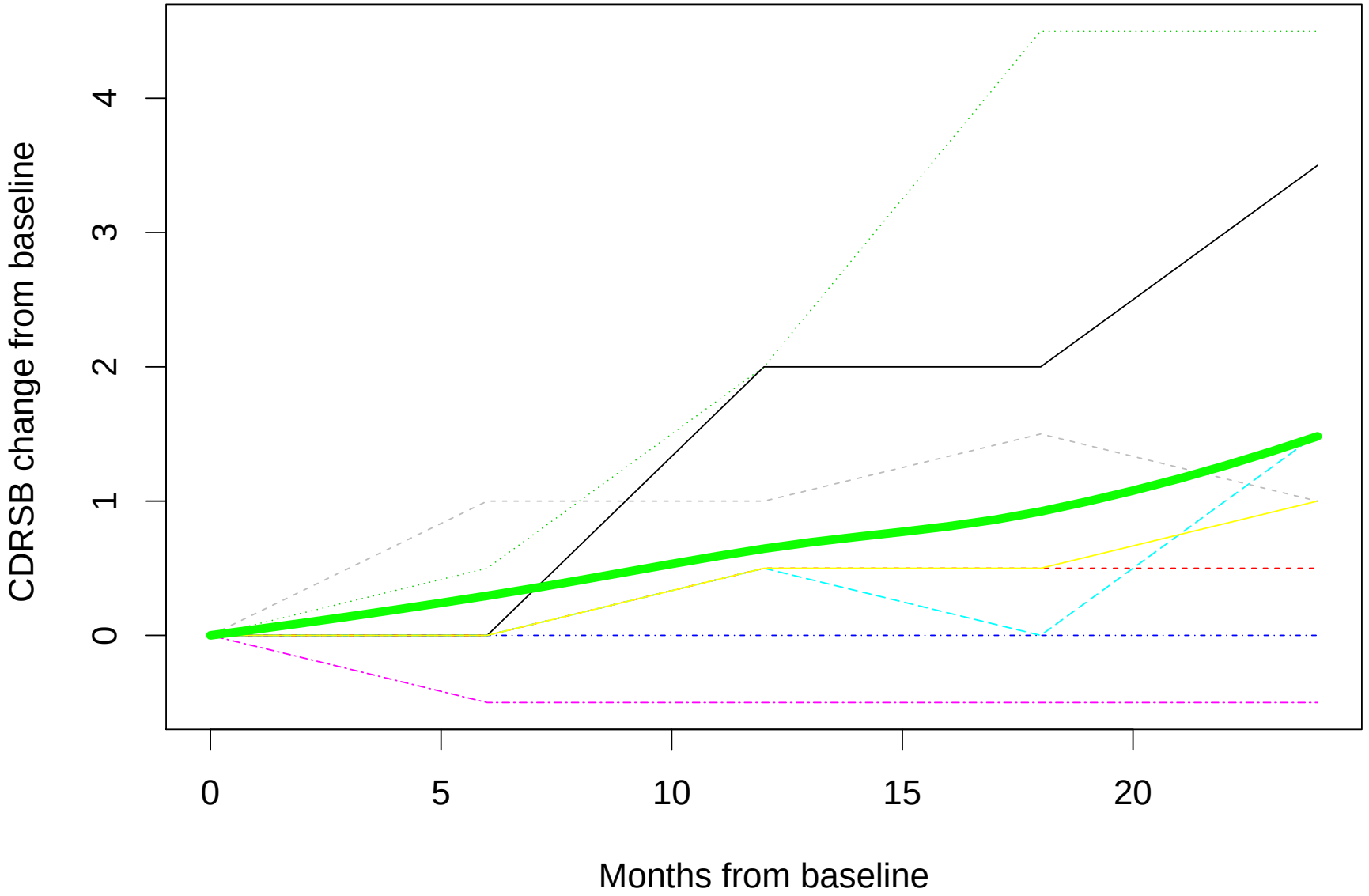
Summary of GUIDE models

1. Classification
   (a) Node models: simple, kernel, and nearest-neighbor
   (b) Splits: univariate and bivariate
2. Regression
   (a) Least squares: weighted, stepwise, multiple linear, best polynomial, and simple linear with ANCOVA
   (b) Least median of squares: multiple linear and best polynomial
   (c) Quantile: multiple linear and best polynomial
   (d) Poisson: multiple linear and best polynomial
   (e) Proportional hazards: multiple linear and best polynomial
   (f) Multi-response and longitudinal data, with and without time variables
3. All models accept missing values and 0-1 weights
4. Ensemble models (bagging and forests) also available

Clustering longitudinal data: Alzheimer's disease

- 1638 subjects observed at baseline, 6, 12, 18, and 24 months
- Only 285 subjects have responses in CDRSB at all time points
- CDRSB = Clinical Dementia Rating Sum of Boxes (lower is better)
- 26 baseline predictor variables

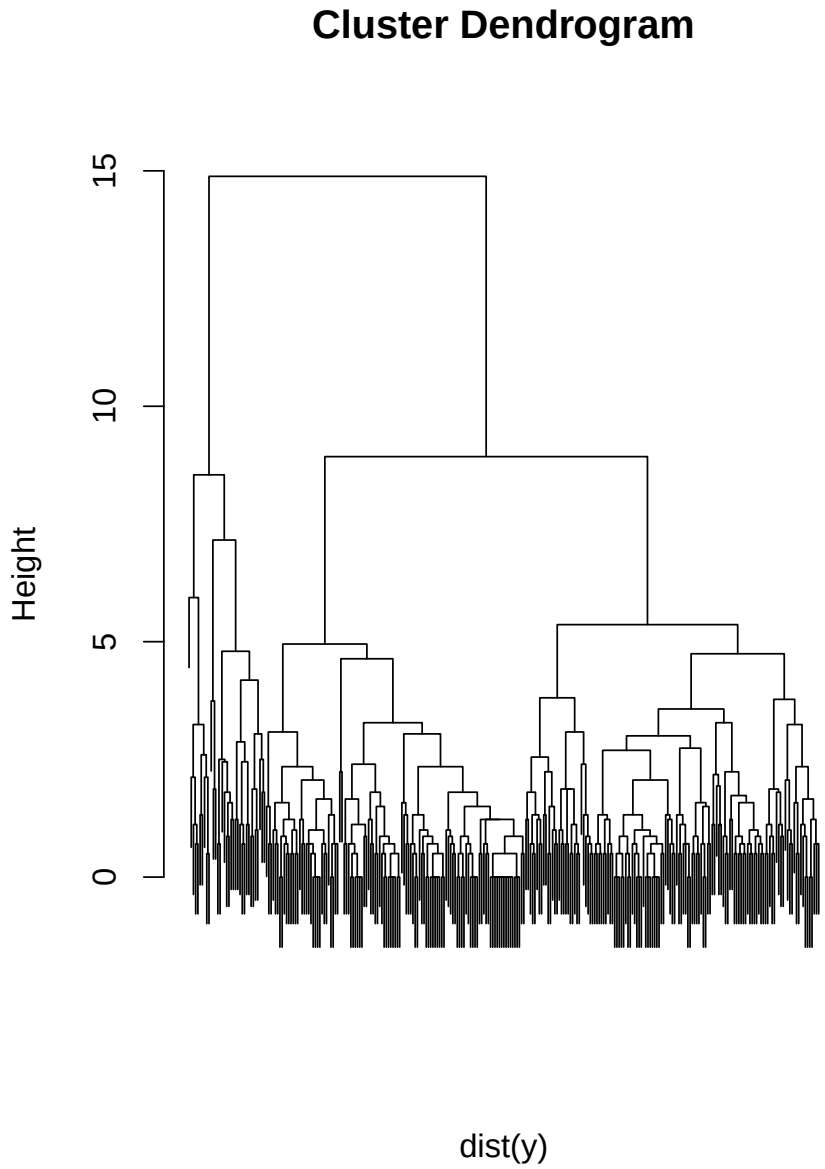
Sample paths with smoothed mean

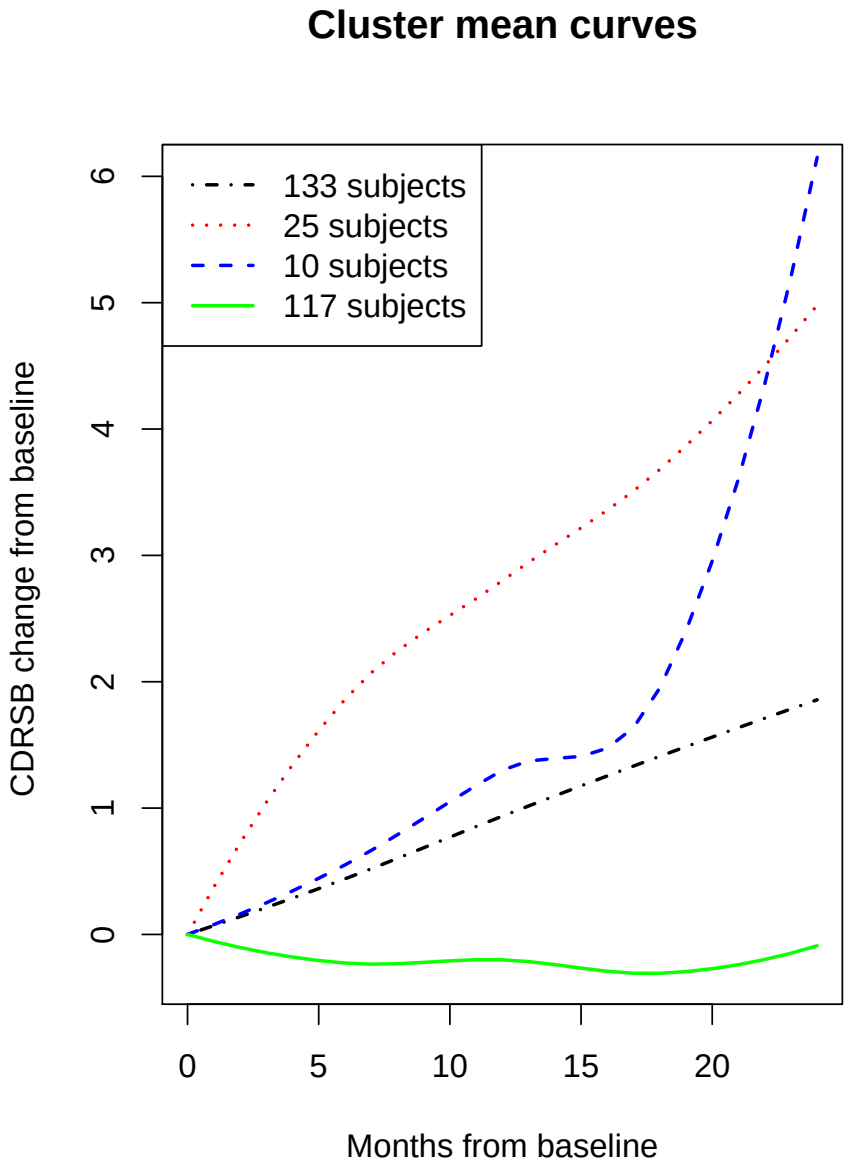


Two methods to group subjects into clusters

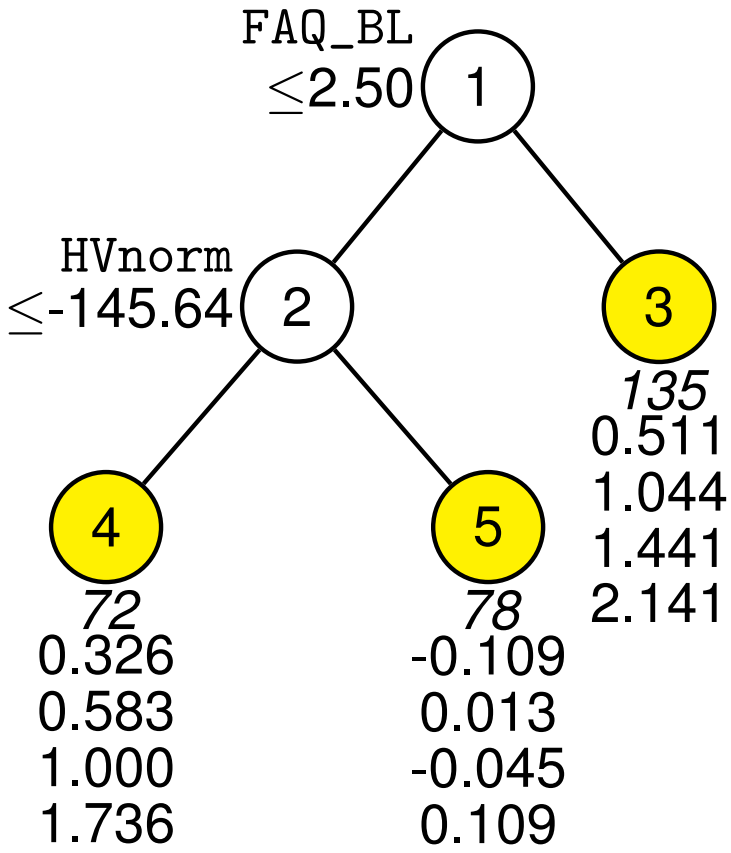
1. Standard clustering methods —
   (a) uses only responses but not covariates
   (b) requires work to understand cluster relationships of covariates
   (c) applicable only to 285 subjects with responses at all time points
2. GUIDE regression tree —
   (a) uses responses and covariates
   (b) automatically determines cluster-covariate relationships
   (c) applicable to completers and noncompleters

Hierarchical clustering for 285 completers



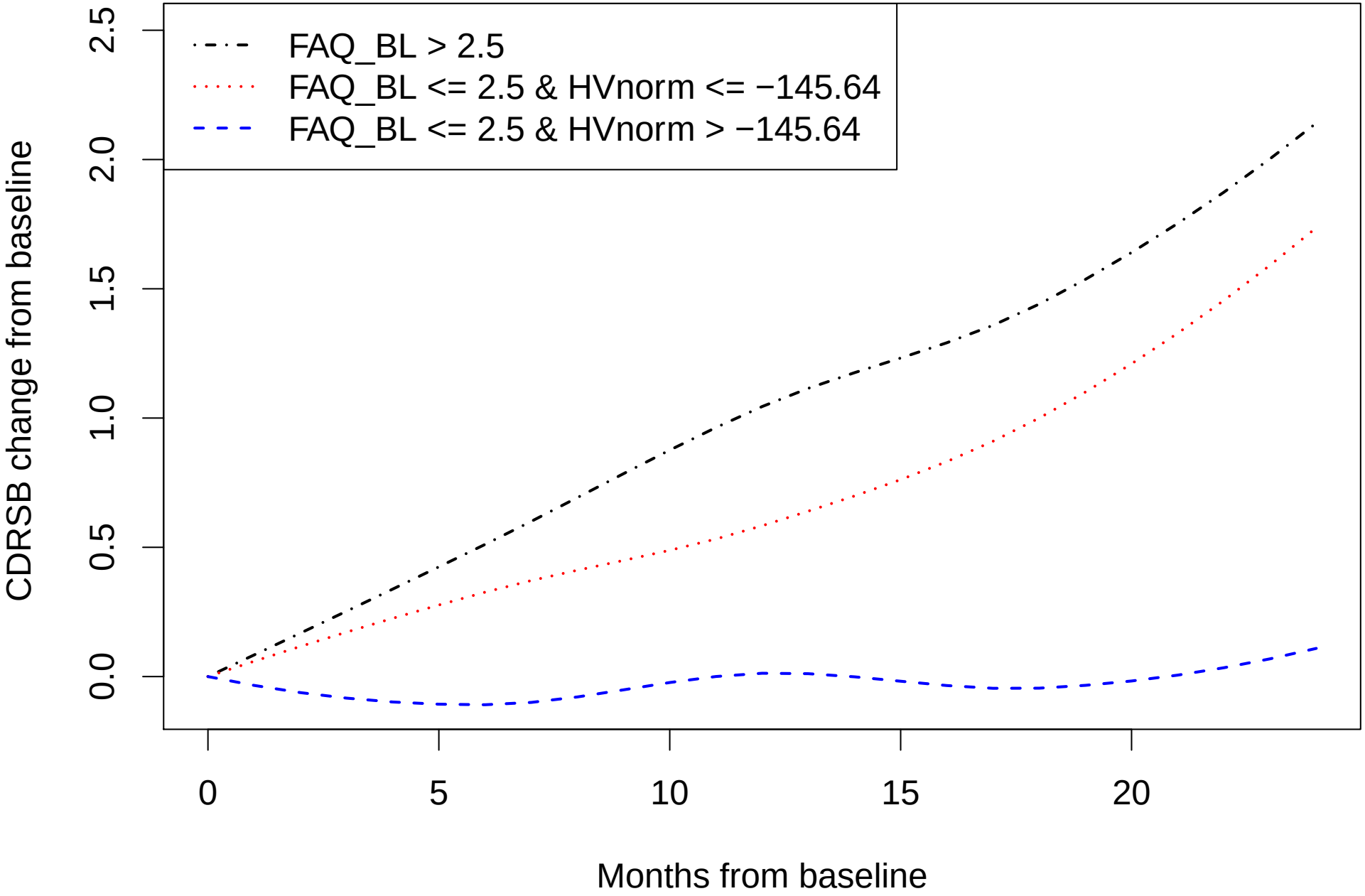


Subgroups for change in CDRSB from baseline for 285 completers (Loh and Zheng, 2013)

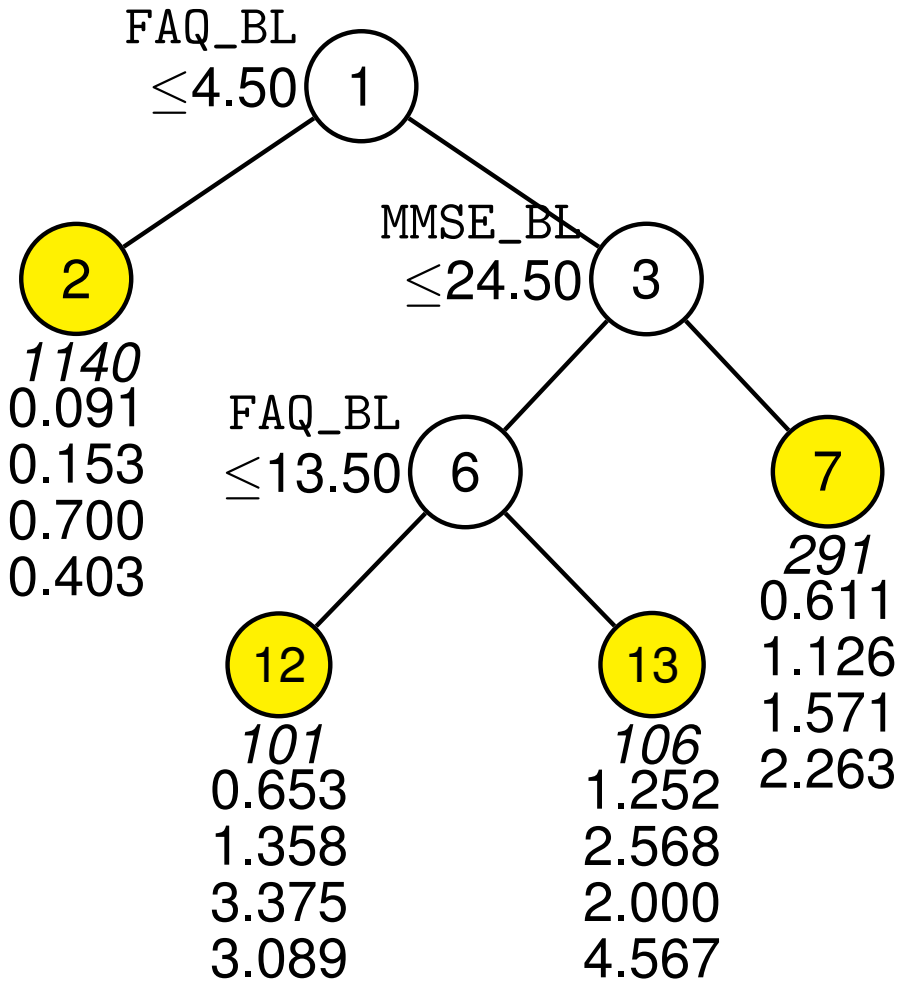


Numbers below nodes are sample size (in *italics*)
and mean CDRSB change at 6, 12, 18 and 24 months
HVnorm = normalized Hippocampal volume (higher is better)
FAQ_BL = Functional Activities Questionnaire at baseline (lower is better)

Mean subgroup paths for 285 completers



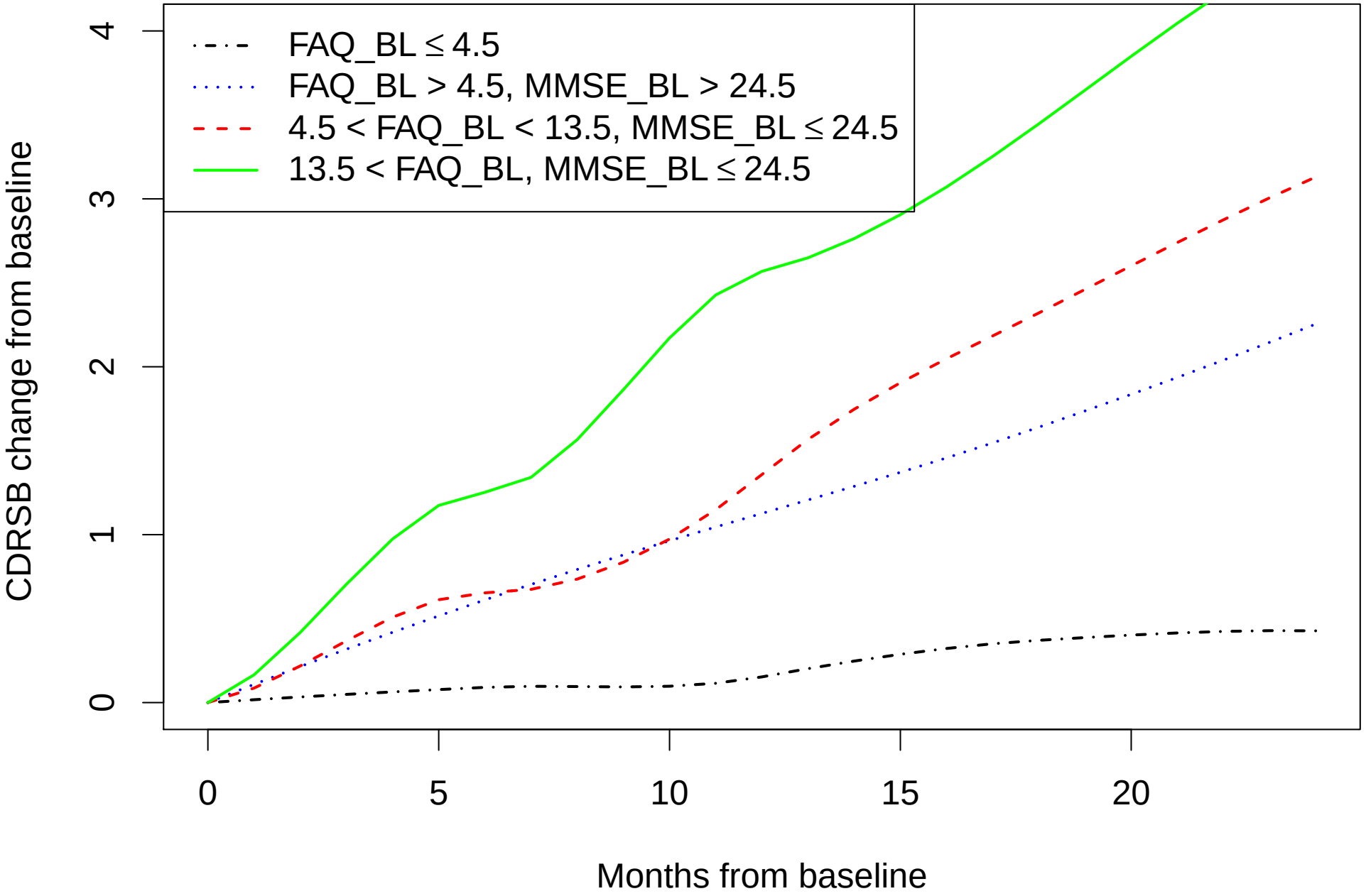
Subgroups for change in CDRSB from baseline for all 1638 subjects



FAQ_BL = Functional Activities Questionnaire at baseline (lower is better)

MMSE_BL = Mini-mental state exam at baseline (higher is better)

Mean subgroup paths for all 1638 subjects



Subgroup identification: five desirable properties

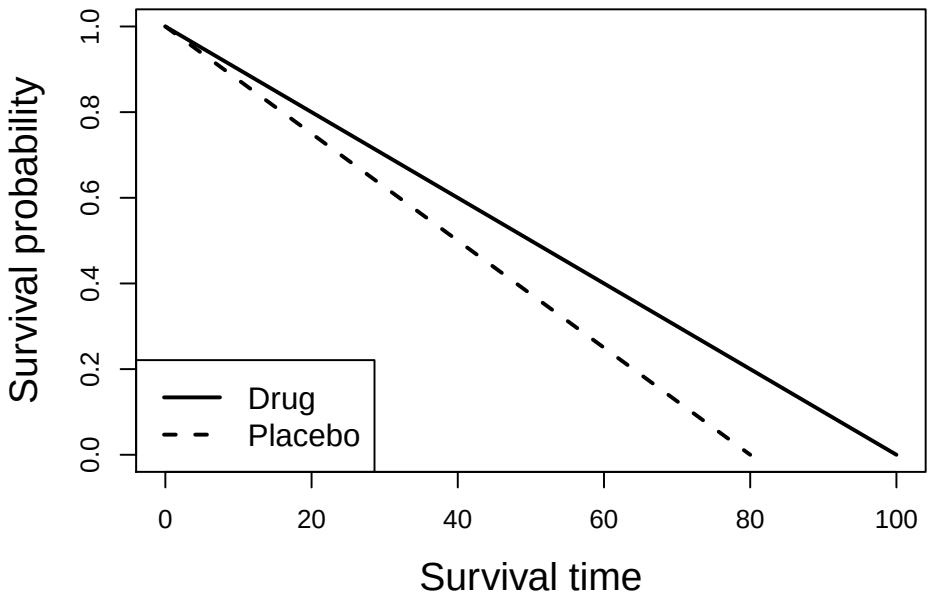
1. **Good accuracy** in identification of subgroups
2. **Unbiased variable selection**
3. **Unbiased estimates** of treatment effects within subgroups
4. **Local-linear control** of prognostic variables
5. **Confidence intervals** for estimated treatment effects

Predictive vs. prognostic variables

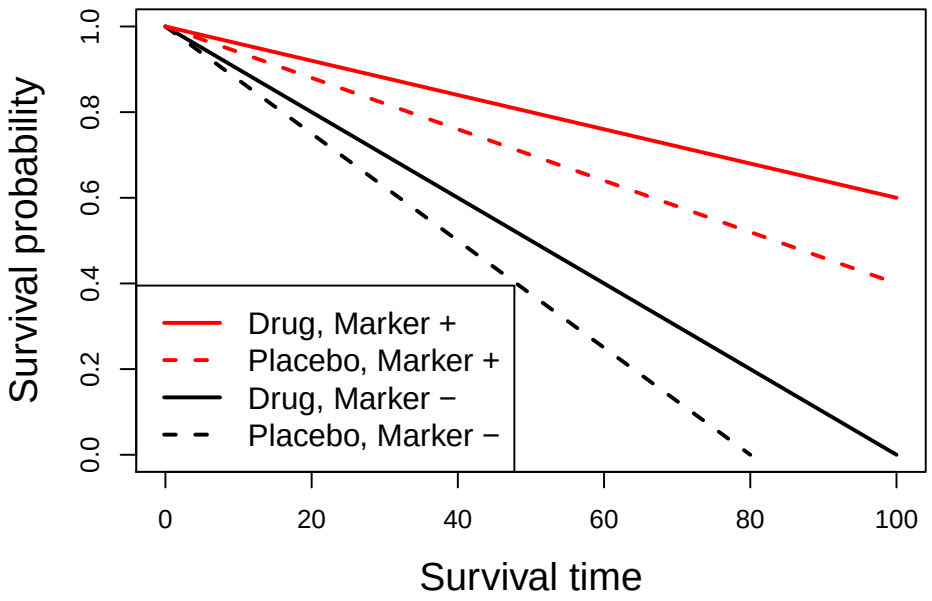
1. **Prognostic** variable is a clinical or biologic characteristic that provides information on the likely outcome of the disease in an untreated individual (e.g., patient age, family history, disease stage, and prior therapy)
2. **Predictive** variable is a characteristic that provides information on the likely benefit from treatment. Such variables can be used to identify subgroups of patients who are most likely to benefit from a given therapy.
3. **Prognostic variables** define the effects of patient or tumor characteristics on the patient outcome, whereas **predictive variables** define the effect of treatment on the tumor.

— Italiano (2011)

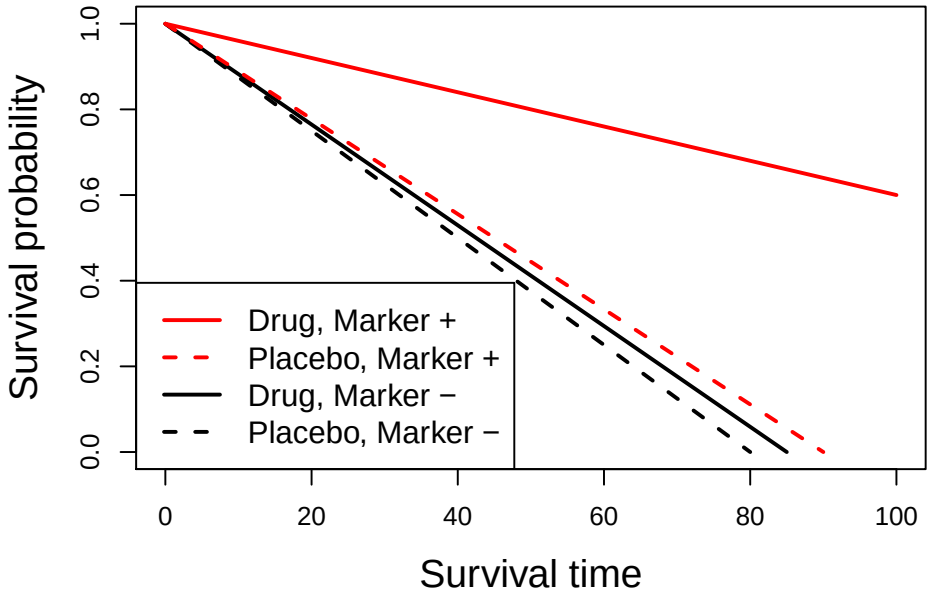
Whole population



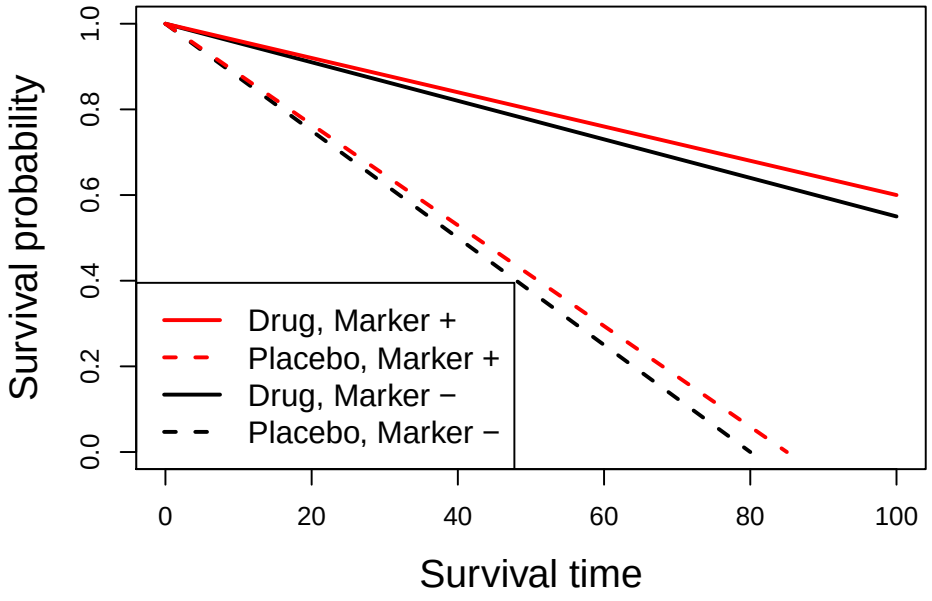
Prognostic biomarker



Predictive biomarker

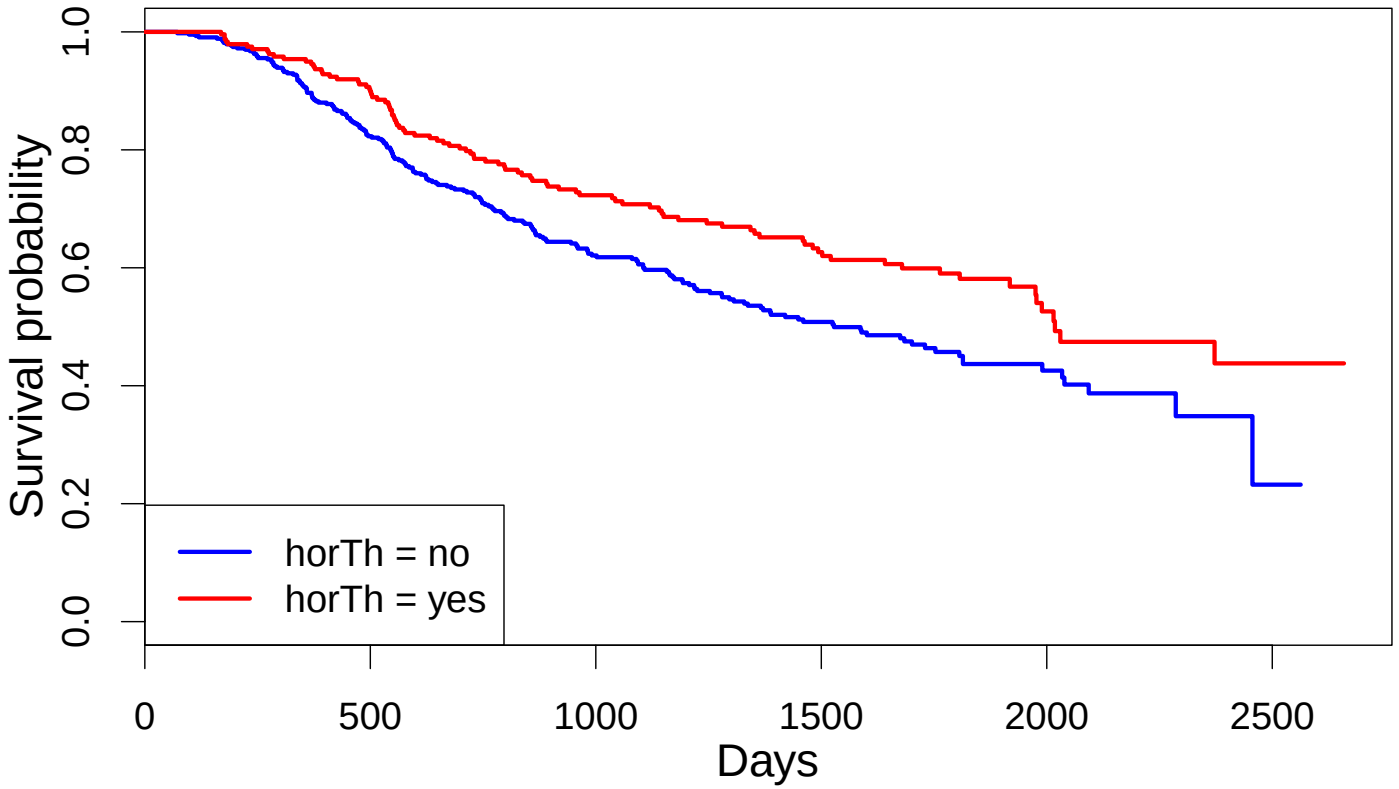


No biomarker effect



Subgroup identification: breast cancer trial

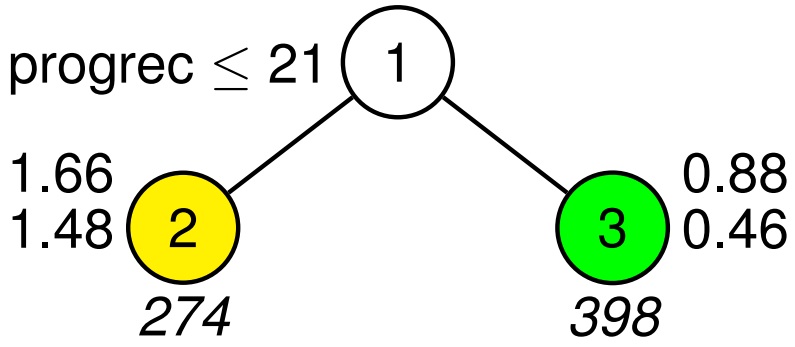
- Randomized clinical trial of 672 subjects with primary node positive breast cancer (Schumacher et al., 1994; data from **TH.data** R package; 14 subjects with censored times less than smallest uncensored time excluded)
- Response is recurrence-free survival time (8–2659 days, 299 uncensored, 387 censored)
- Eight predictor variables with no missing values:
  1. **horTh** (hormone therapy, yes/no)
  2. **age** (21–80 years)
  3. **tsize**(tumor size, 3–120 mm)
  4. **pnodes**(number of positive lymph nodes, 1–51)
  5. **progrec** (progesterone receptor status, 0–2380 fmol)
  6. **estrec** (estrogen receptor status, 0–1144 fmol)
  7. **menostat** (menopausal status, pre/post)
  8. **tgrade** (tumor grade, 1, 2, 3)

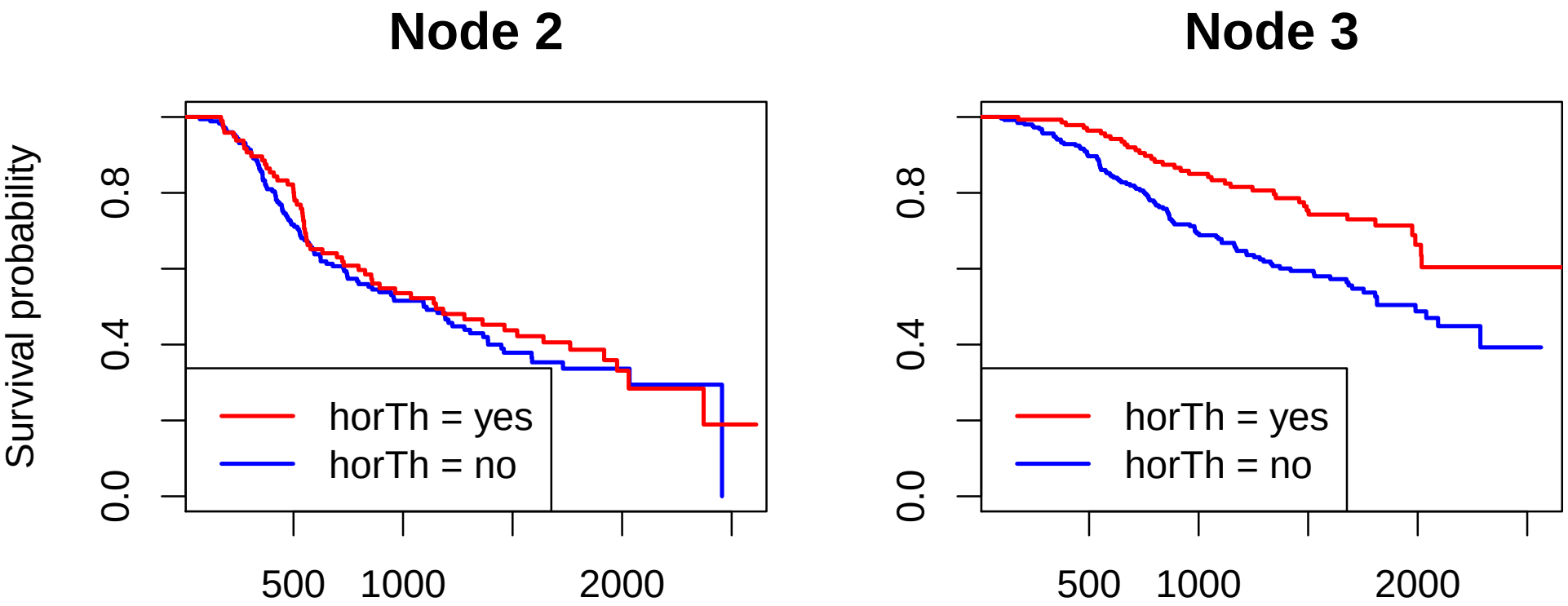


Variable	Coef	p-value	Variable	Coef	p-value
horTh=yes	-0.3463	7.3e-03	tsize	0.0078	4.8e-02
age	-0.0095	3.1e-01	pnodes	0.0488	5.7e-11
meno=Post	0.2585	1.6e-01	progrec	-0.0022	1.1e-04
tgrade.L	0.5513	3.7e-03	estrec	0.0002	6.6e-01
tgrade.Q	-0.2011	9.9e-02			

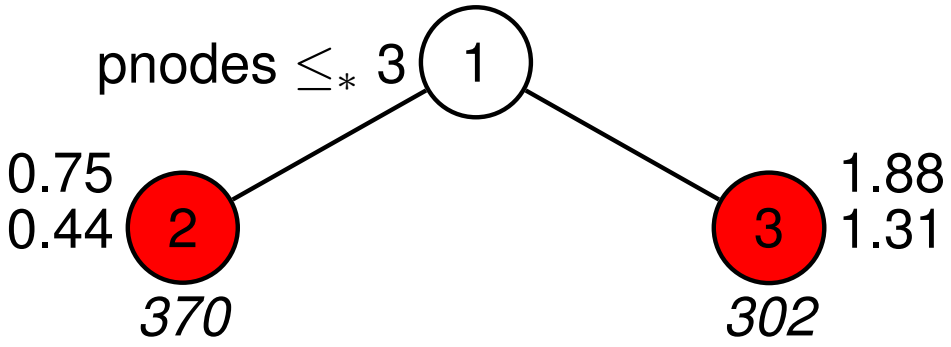
Is there a subgroup where hormone therapy is ineffective?

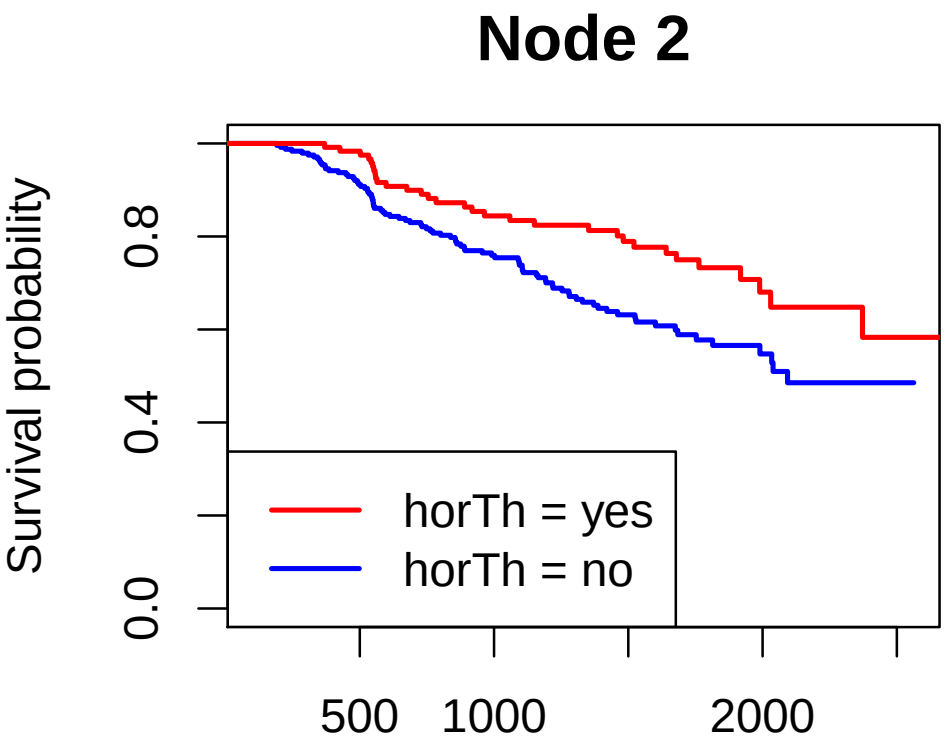
Gi without linear prognostic control

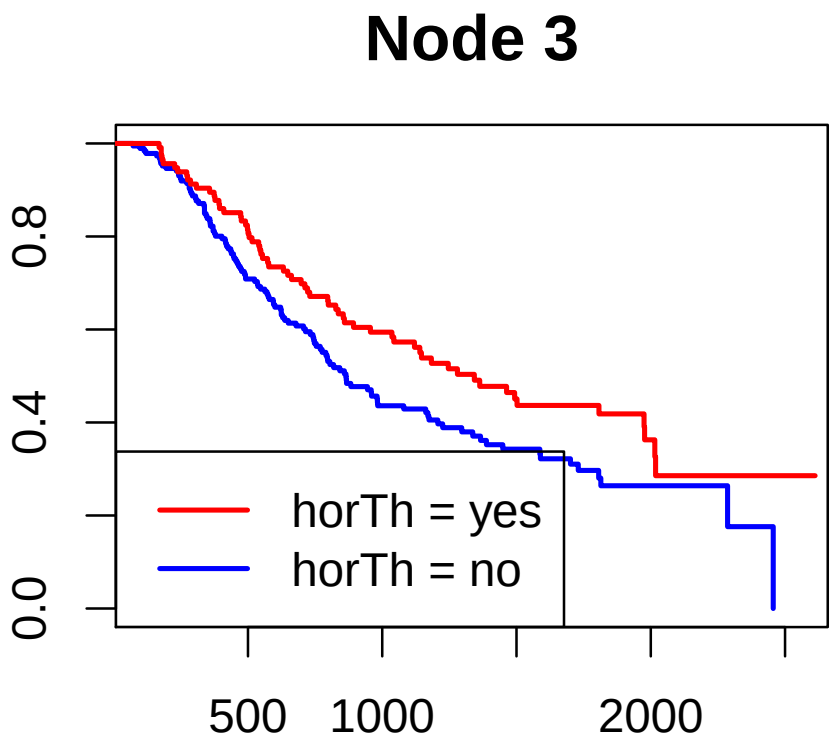




Gs without linear prognostic control







Key idea #1:
use piecewise-*linear* models

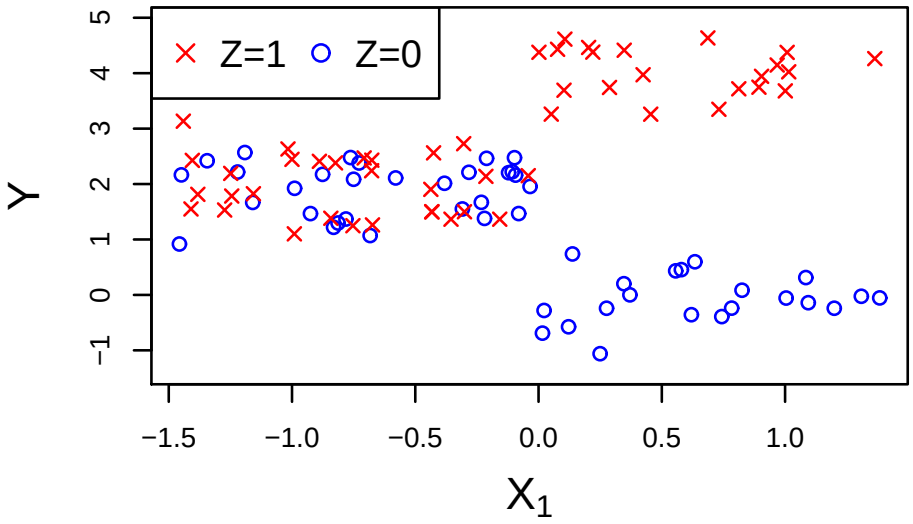
- Suppose treatment variable Z takes values 0, 1, . . .
- Fit the model $EY = \eta + \sum_k \beta_k I(Z = k)$ in each node (so that treatment effects can be estimated)
- CART, RPART, and other piecewise-constant trees inapplicable

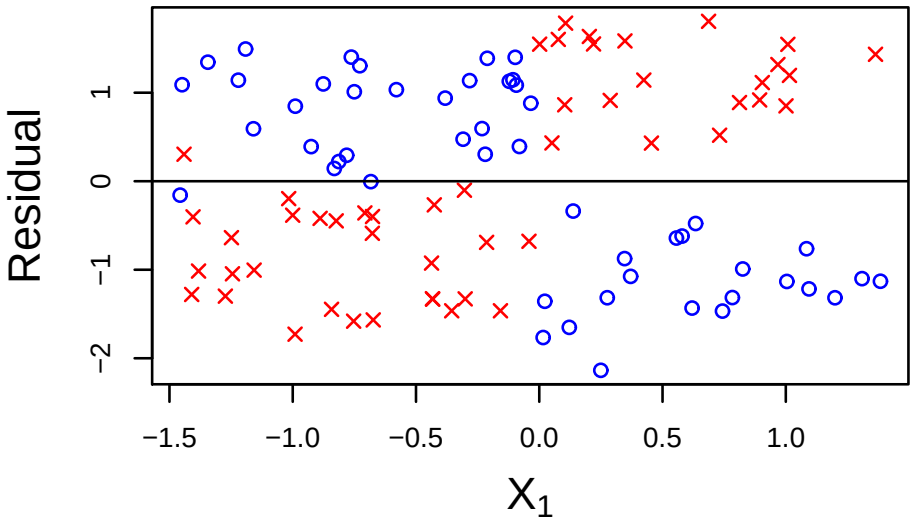
GUIDE (Loh, 2002, 2009) and MOB (Zeileis et al., 2008)

- These algorithms use significance tests to select variables for splitting data
- GUIDE uses chi-squared tests of residual signs vs. each predictor variable
  - missing values are *included*
- CTREE and MOB use permutation tests on score functions
  - missing values are *excluded* (implies missing completely at random)

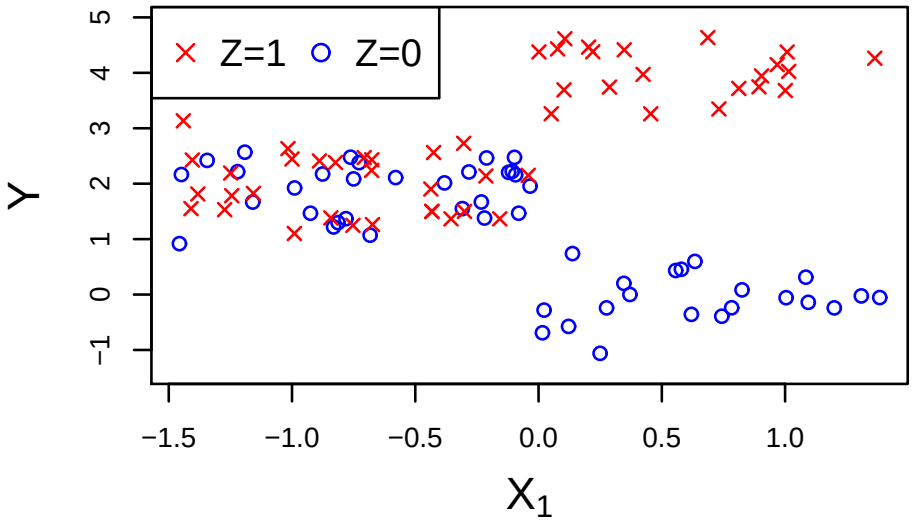
Example with treatment Z = 0, 1

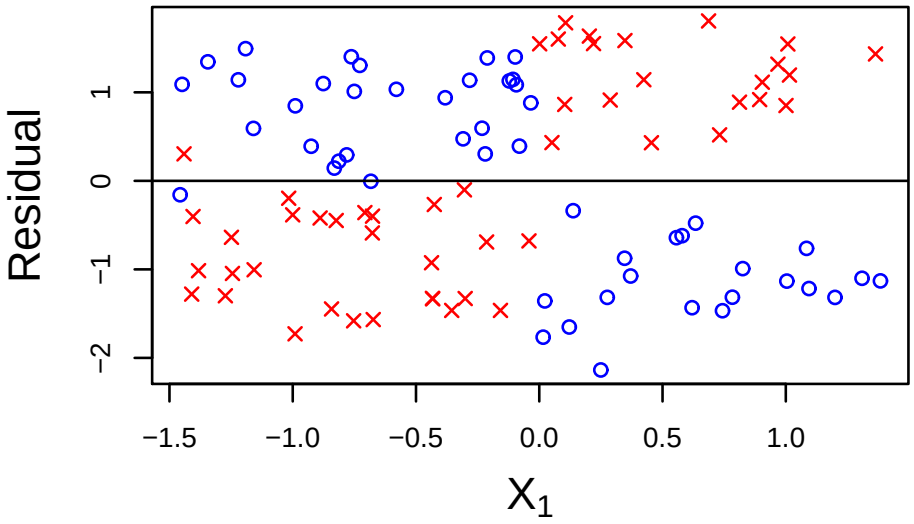
- True model:
  $Y = 1.9 + 0.2I(Z = 1) - 1.8I(X_1 > 0) + 3.6I(X_1 > 0, Z = 1) + \varepsilon$
- $X_2, X_3, \ldots$ are noise
- Fit $EY = \beta_0 + \beta_1 Z$ to data in each node

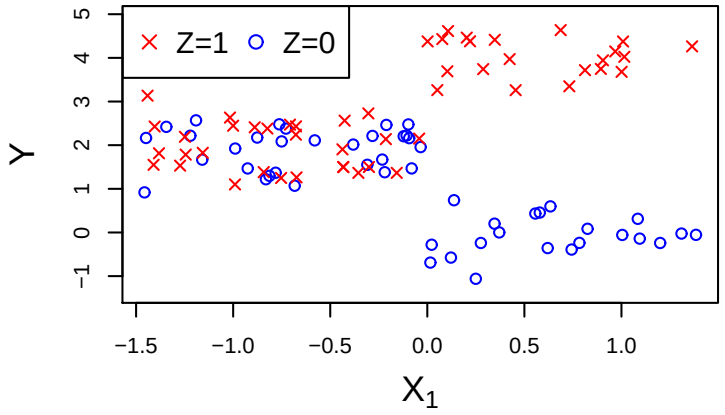


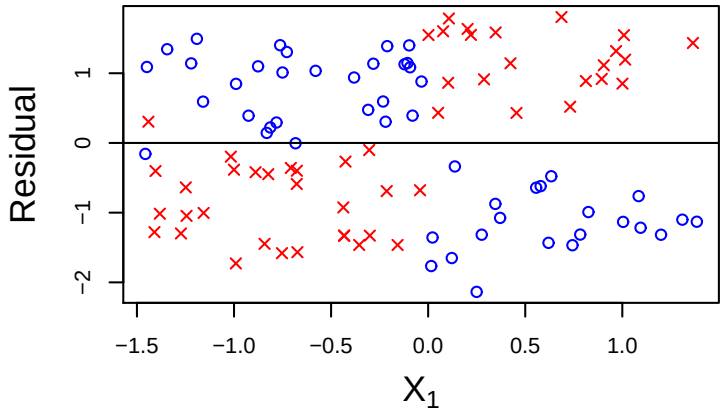


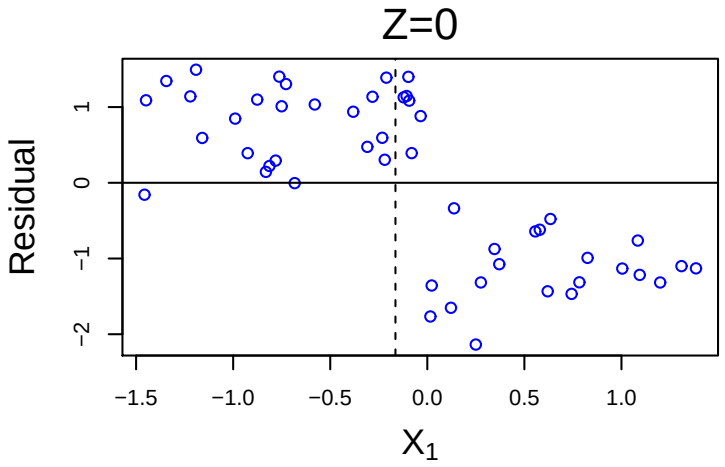
Key idea #2:
examine residual patterns
for each treatment level

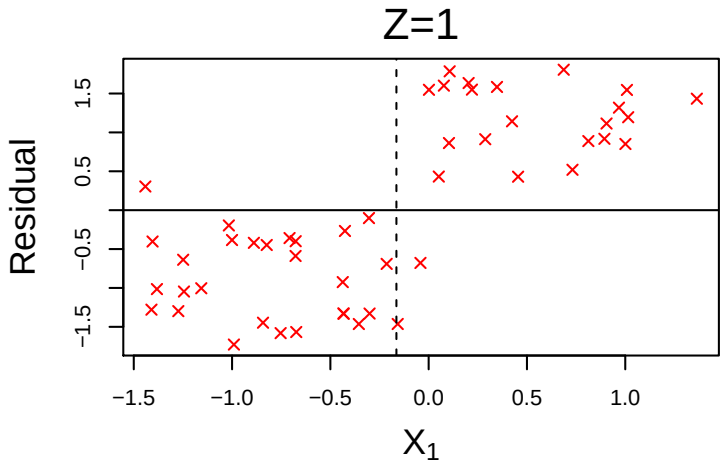












$Z=0$	$X_1 \le \bar{x}_1$	$X_1 > \bar{x}_1$
resid > 0	21	6
resid ≤ 0	2	21

$\chi^2 = 21.2,\ p = 4 \times 10^{-6}$

$Z=1$	$X_1 \le \bar{x}_1$	$X_1 > \bar{x}_1$
resid > 0	1	21
resid ≤ 0	26	2

$\chi^2 = 35.2,\ p = 3 \times 10^{-9}$

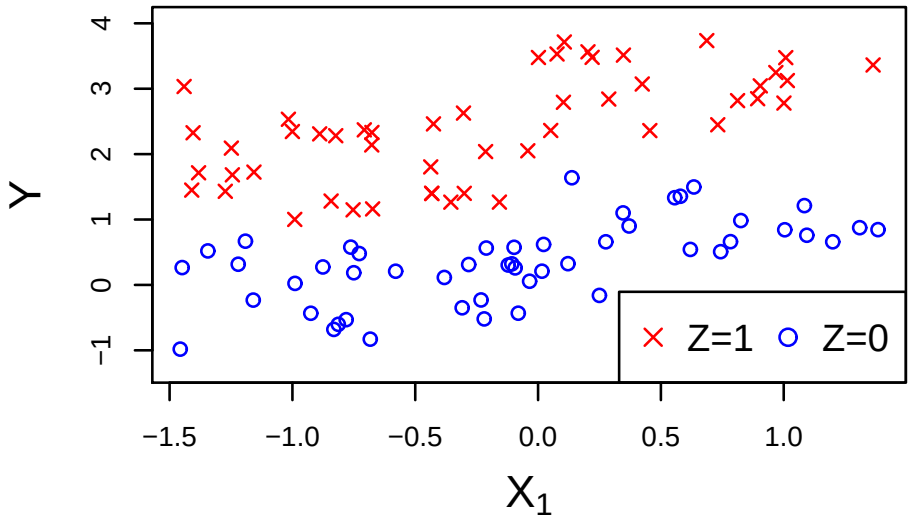
Key idea #3:
why group ordinal variables?

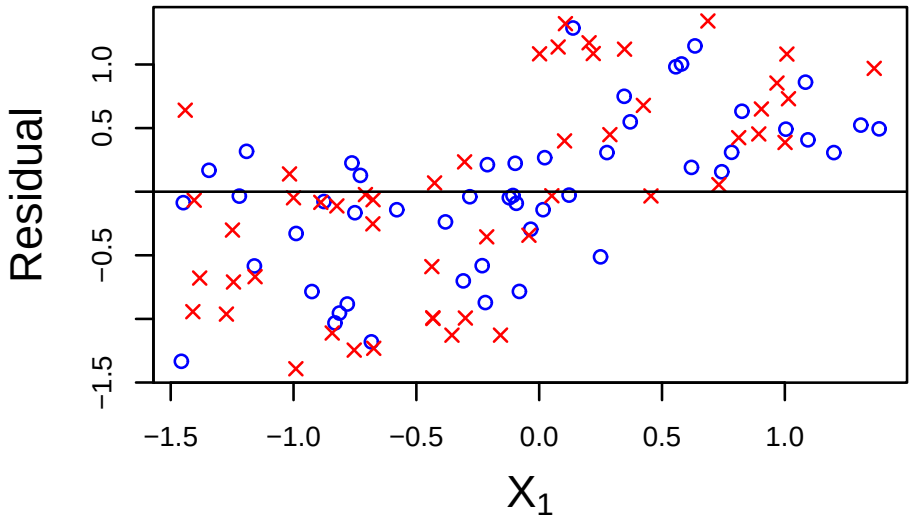
- Grouping values of ordinal X variables may result in power loss
- But grouping allows missing values to be used!

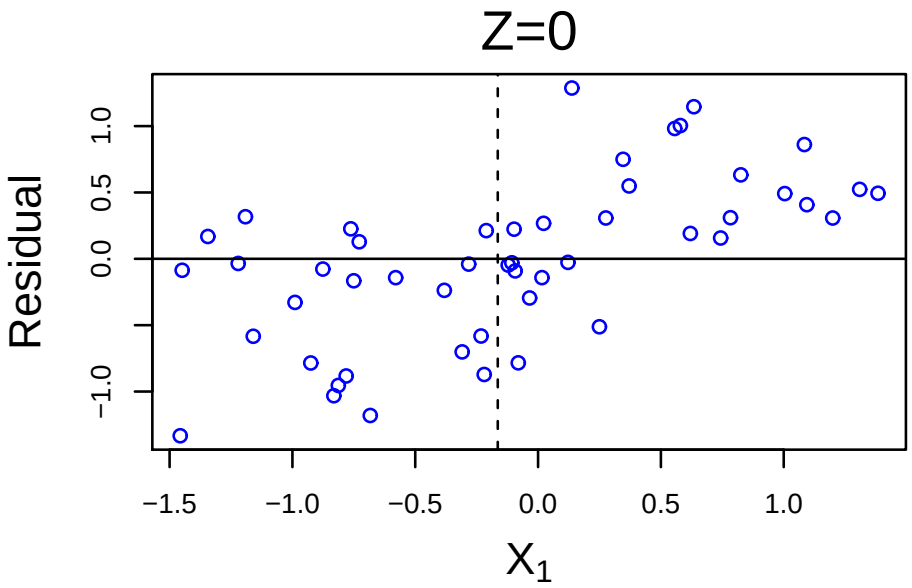
Gs method ("s" for "sum")

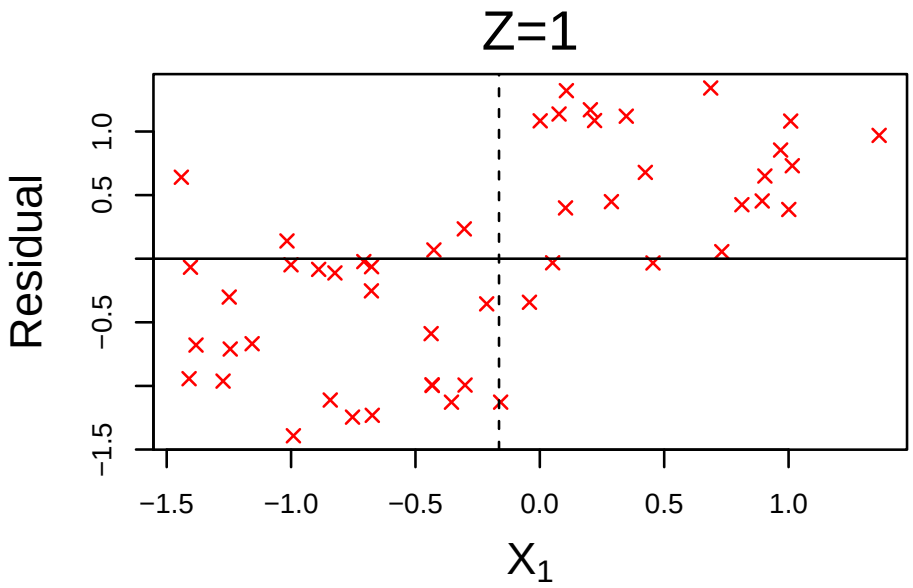
1. Obtain the residuals from the model $EY = \eta + \sum_k \beta_k I(Z = k)$
2. Do for each X variable:
   (a) Do for each value of Z:
      i. Crosstab residual signs vs. grouped values of X
      ii. Add one more group for missing values in X if there are any
      iii. Compute chi-squared statistic of the table
      iv. Convert chi-squared value to one with a single df
   (b) Sum converted chi-squareds over values of Z to get test statistic
3. Let X^* have largest test statistic
4. Find split $X^* \in S$ that minimizes sum of squared residuals in subnodes
5. Partition data and recursively apply procedure to each subnode

Problem: Gs is sensitive to prognostic variables









Key idea #4:
test for treatment interactions

1. Usual approach: add cross-product "interaction" terms if X is ordinal:

$$EY = \eta + \sum_k \beta_k I(Z = k) + \sum_k \gamma_k X I(Z = k)$$

2. Two problems with this:
   (a) Cross-products $XI(Z = k)$ do not represent every kind of interaction
   (b) Cross-products do not allow missing values in X
3. Solution: Use interaction model for categorical variables

$$EY = \eta + \sum_j \alpha_j I(X = j) + \sum_k \beta_k I(Z = k) + \sum_j \sum_k \gamma_{jk} I(X = j, Z = k)$$

with a category for missing values. If X is ordinal, group its values.

GUIDE Gi method for subgroup identification (Loh et al., 2015)

Test lack of fit of *model without interactions*:

1. Do for each X at each node:
   (a) If X is ordinal, categorize it into two groups at its mean
   (b) If X is categorical, let its values form the groups
   (c) Add a group for missing values
   (d) Let H be the factor variable created from the groups
   (e) Test lack of fit of the model $EY = \beta_0 + \sum_j \alpha_j I(H = j) + \sum_k \beta_k I(Z = k)$
2. Let X^* be the variable with the most significant chi-squared
3. Find the split on X^* that minimizes the sum of squared residuals of the model $EY = \eta + \sum_k \beta_k I(Z = k)$ fitted to each of the two subnodes

Extension to censored response data

1. Let U_i and C_i be survival and censoring times of subject i
2. Let $Y_i = \min(U_i, C_i)$ and $\delta_i = I(T_i < C_i)$ be the event indicator
3. Let $\Lambda_0(.)$ be the baseline cumulative hazard function of PH model
4. Estimate coefficients of PH model by iteratively fitting a Poisson regression model with δ_i as response and $\log \Lambda_0(y_i)$ as offset:
   (a) Use the Nelson-Aalen method to get an initial estimate of $\Lambda_0(.)$
   (b) Use GUIDE to construct a Poisson regression tree
   (c) Update $\Lambda_0(.)$ with the tree
   (d) Repeat steps (b) and (c) four more times

Previous methods for binary responses

VT: Virtual twins (Foster et al., 2011) Assume $Y, Z = 0, 1$.

1. Estimate $\tau = P(Y = 1|Z = 1) - P(Y = 1|Z = 0)$ with Random forest (Breiman, 2001) using $Z, X_1, \ldots, X_M$, $ZX_1, \ldots, ZX_M$, and $(1 - Z)X_1, \ldots, (1 - Z)X_M$ as predictors.
2. Use RPART to estimate τ. Subgroups are nodes with large $\hat{\tau}$.

Weaknesses:

1. Selection bias of CART (Breiman et al., 1984) and random forest (RF).
2. No good way to deal with missing values (RF needs prior imputation).
3. Not extensible to three or more treatments and to censored responses.
4. Random subgroups due to RF.

SIDES: (Lipkovich et al., 2011; Lipkovich and Dmitrienko, 2014)

1. Find 5 splits to minimize p-value (e.g., differential treatment effects or difference in efficacy and safety between child nodes).
2. For each split, repeat on most promising child node.

Weaknesses: selection bias; effect bias; inapplicable to missing data

Interaction trees (Su et al., 2008, 2009). Fit proportional hazards (PH) model to node and split it with variable having largest interaction with treatment

Weaknesses: same as SIDES

QUINT: Qualitative interaction tree (Dusseldorp and Van Mechelen, 2014) Split each node to optimize a weighted sum of measures of effect size and subgroup size.

Strength: Allows simultaneous control of effect size and subgroup size

Weaknesses:

1. Selection bias.
2. Needs one treatment to be better in one subgroup and worse in other.
3. Not easily extensible to three or more treatments.
4. Not easily extensible to censored responses.

AIDS clinical trial

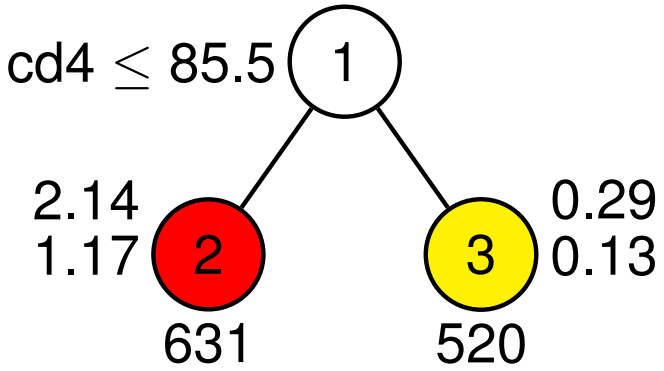
- 1151 subjects in a double-blind, randomized AIDS clinical trial (Hosmer et al., 2008)
- 3-drug regime with indinavir (IDV) vs 2-drug regime without IDV
- Subjects were on HIV-infected and had $\leq$ 200 CD4 cells/mm^3 at baseline and $\geq$ 3 months prior zidovudine (ZDV) therapy
- Randomization stratified by CD4 cell counts at time of screening
- Outcome was time to AIDS defining event or death
- Trial stopped early due to efficacy meeting pre-specified level of significance at interim analysis
- No missing values

time	days to AIDS diagnosis or death
censor	1: AIDS defining event or death, 0: otherwise
tx	1: treatment includes IDV, 0: otherwise
strat2	CD4 stratum at screening (0: CD4 $\leq$ 50, 1: otherwise)
sex	1: male, 2: female
raceth	1: white non-Hispanic, 2: black non-Hispanic, 3: Hispanic, 4: Asian, Pacific Islander, 5: American Indian, Alaskan native, 6: other/unknown
ivdrug	IV drug use history (1: never, 2: currently, 3: previously)
hemophil	Hemophiliac (1: yes, 2: no)
karnof	Karnofsky performance scale (100: normal, no complaints, no evidence of disease; 90: normal activity possible, minor signs/symptoms of disease; 80: normal activity with effort, some signs/symptoms of disease; 70: cares for self, normal activity/active work not possible)
cd4	baseline CD4 count
priorzdv	months of prior ZDV use
age	age in years at enrollment

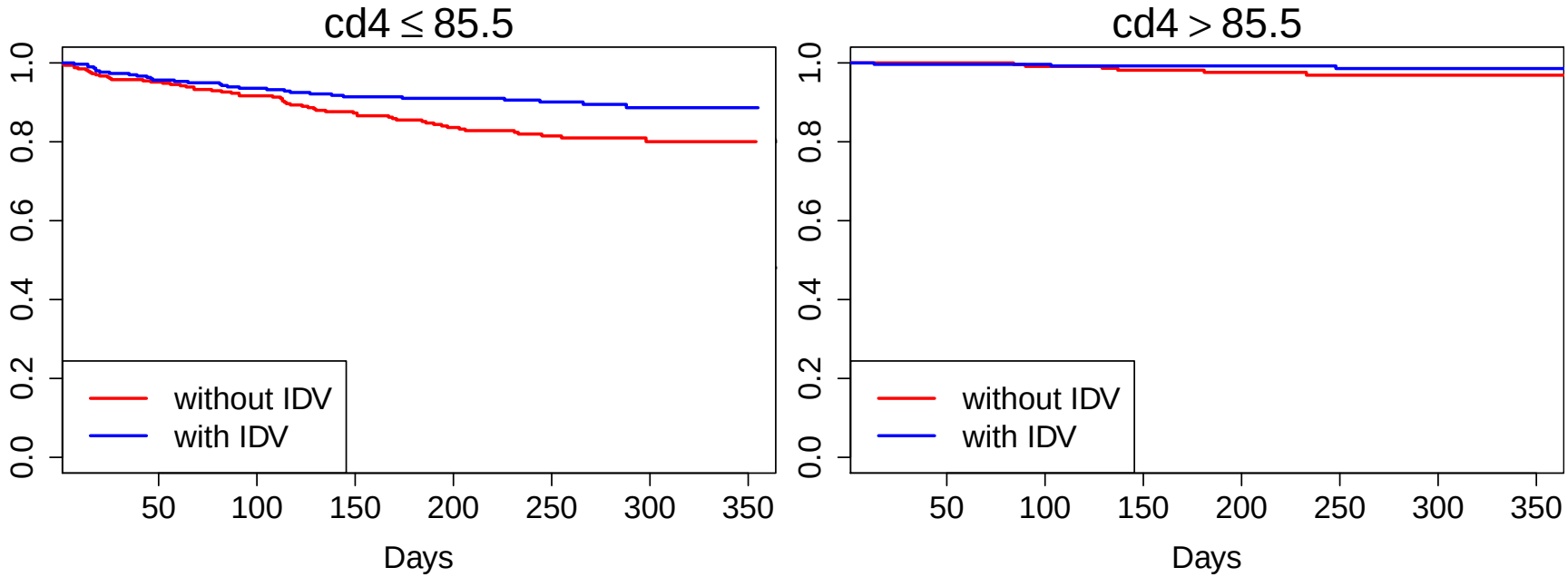
Cox proportional hazards model

	coef	exp(coef)	se(coef)	z	p
tx	-0.675535	0.509	0.21606	-3.127	1.8e-03
strat2	-0.055075	0.946	0.34922	-0.158	8.7e-01
sex	0.119640	1.127	0.28546	0.419	6.8e-01
raceth	0.067347	1.070	0.11802	0.571	5.7e-01
ivdrug2	0.748667	2.114	1.03791	0.721	4.7e-01
ivdrug3	-0.640297	0.527	0.33854	-1.891	5.9e-02
hemophil	0.116408	1.123	0.60060	0.194	8.5e-01
karnof	-0.055668	0.946	0.01215	-4.580	4.6e-06
cd4	-0.014120	0.986	0.00384	-3.681	2.3e-04
priorzdv	-0.000389	1.000	0.00387	-0.100	9.2e-01
age	0.022484	1.023	0.01137	1.978	4.8e-02

Gi model with no linear prognostic control

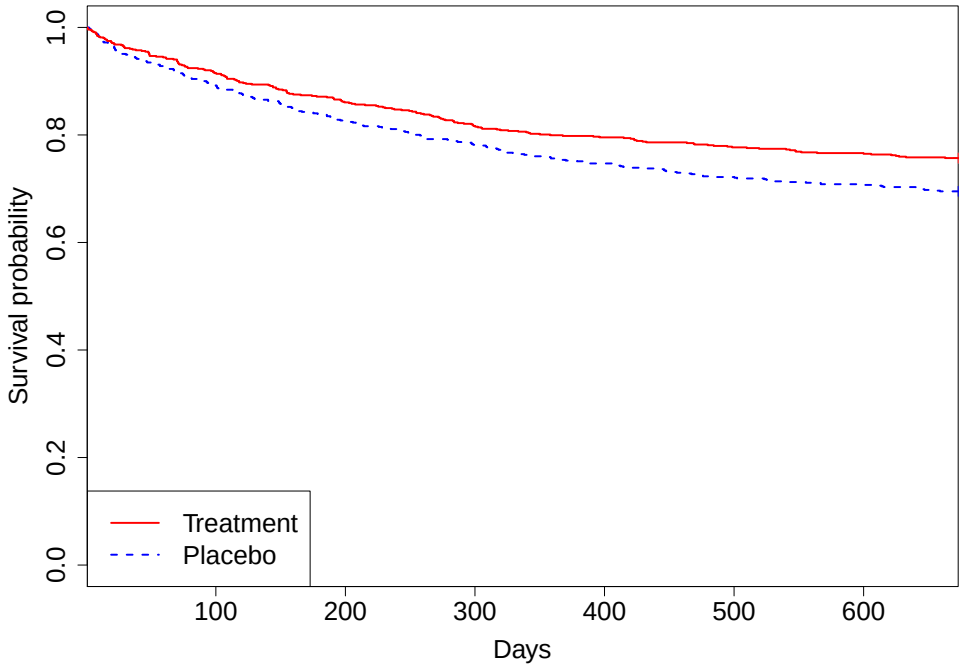


Numbers beside terminal nodes are estimated relative risks (relative to average for sample ignoring covariates) corresponding to treatment levels 0 and 1

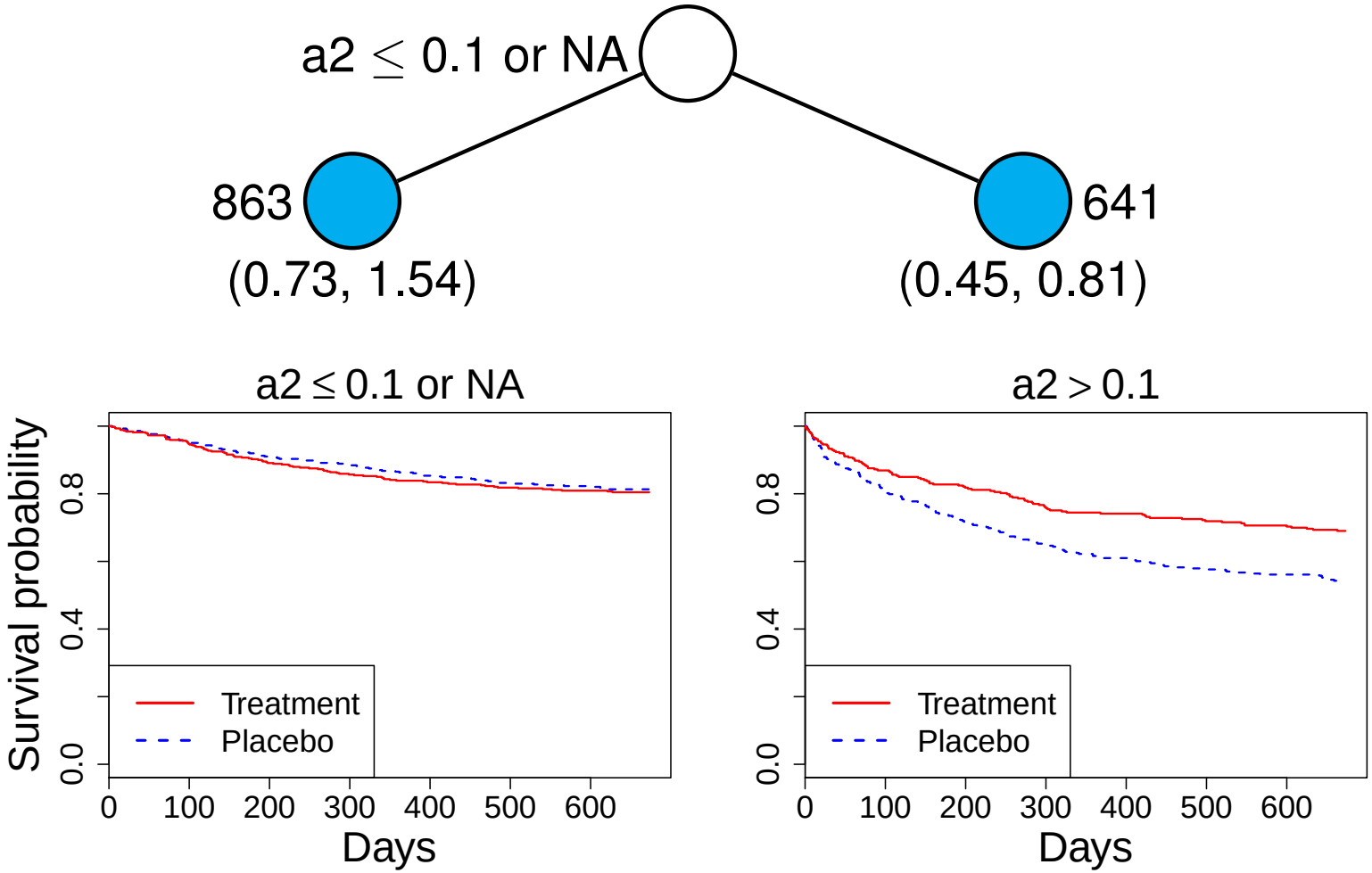


Retrospective candidate gene study (large numbers of missing values)

- 1504 subjects randomized to treatment or placebo
- Response is survival time in days, with 63% censored
- 23 baseline (17 ordered, 6 categorical) and 282 genetic (cat.) variables
- **95% of subjects have missing values**; only 7 variables are complete



GUIDE model with bootstrap intervals of RR



At each node, a case goes to the left child node if stated condition is satisfied.

Sample sizes are beside terminal nodes.

95% bootstrap intervals for relative risk of treatment vs. placebo below nodes.

Extension to multiple responses

Do at each node:

1. For each response variable Y_j, find chi-squared of each X variable
2. Choose the variable X^* with largest sum of chi-squared values over j
3. Choose the split on X^* that yields smallest sum of squared residuals over all response variables

Further extension to correlated response variables

Apply above to principal components (PCA) or linear discriminant coords (LDA) of Y variables computed at each node

How to do inference after subgroup selection?

1. Subgroups are random because they are results of search algorithms
2. Hence, unlike classical theory, true subgroup effects θ are also random
3. Statistical significance of estimates $\hat{\theta}$ must account for the search
4. Some methods (e.g., SIDES) use permutation tests
   (a) Permuting treatment labels has low power because it tests for absence of treatment effects everywhere
   (b) Permuting X values makes no sense because the null hypothesis H_0 cannot be true to begin with
5. Confidence intervals are better because H_0 is not required but how?
6. Bootstrap comes to mind but cannot only apply bootstrap data to one tree
7. Search algorithm must be applied to bootstrap samples to get bootstrap trees — then what?

Motivation: linear regression

- Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be iid from F such that $Y_i = \mu + \beta X_i + \epsilon_i$, where ϵ_i is iid with mean 0 and variance σ^2
- Let $\nu = n - 1$, $\hat{\beta}$ be LS estimate of β, and $t_{\nu,\alpha}$ be upper-α quantile of t_ν dist
- If $\epsilon_i \sim N(0, \sigma^2)$, $P_F(\beta \in \hat{\beta} \pm t_{\nu,\alpha/2}\hat{\sigma}/\sqrt{n}) = 1 - \alpha$
- Suppose we continue to use the same t-interval
- If F is known, use simulation to find $\alpha' = \alpha'(F, n)$ such that

$$P_F(\beta \in \hat{\beta} \pm t_{\nu,\alpha'/2}\hat{\sigma}/\sqrt{n}) = 1 - \alpha \qquad (*)$$

- If F is unknown, replace F with $\hat{F}$ and define $\hat{\alpha} = \alpha'(\hat{F}, n)$
- Bootstrap interval is $\hat{\beta} \pm t_{\nu,\hat{\alpha}/2}\hat{\sigma}/\sqrt{n}$
- As $n \to \infty$, $\hat{F} \to F$ and $\hat{\alpha} \to \alpha'$. Therefore under smoothness conditions

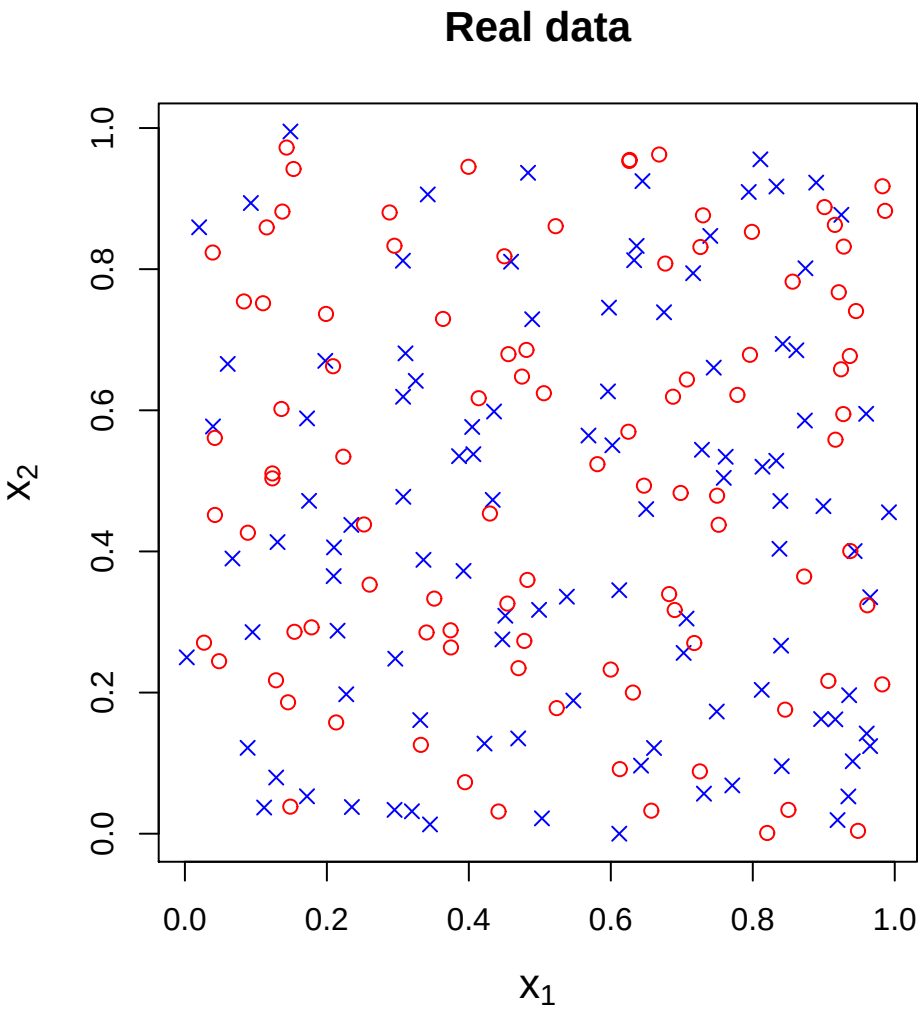
$$P_F(\beta \in \hat{\beta} \pm t_{\nu,\hat{\alpha}/2}\hat{\sigma}/\sqrt{n}) \to 1 - \alpha$$

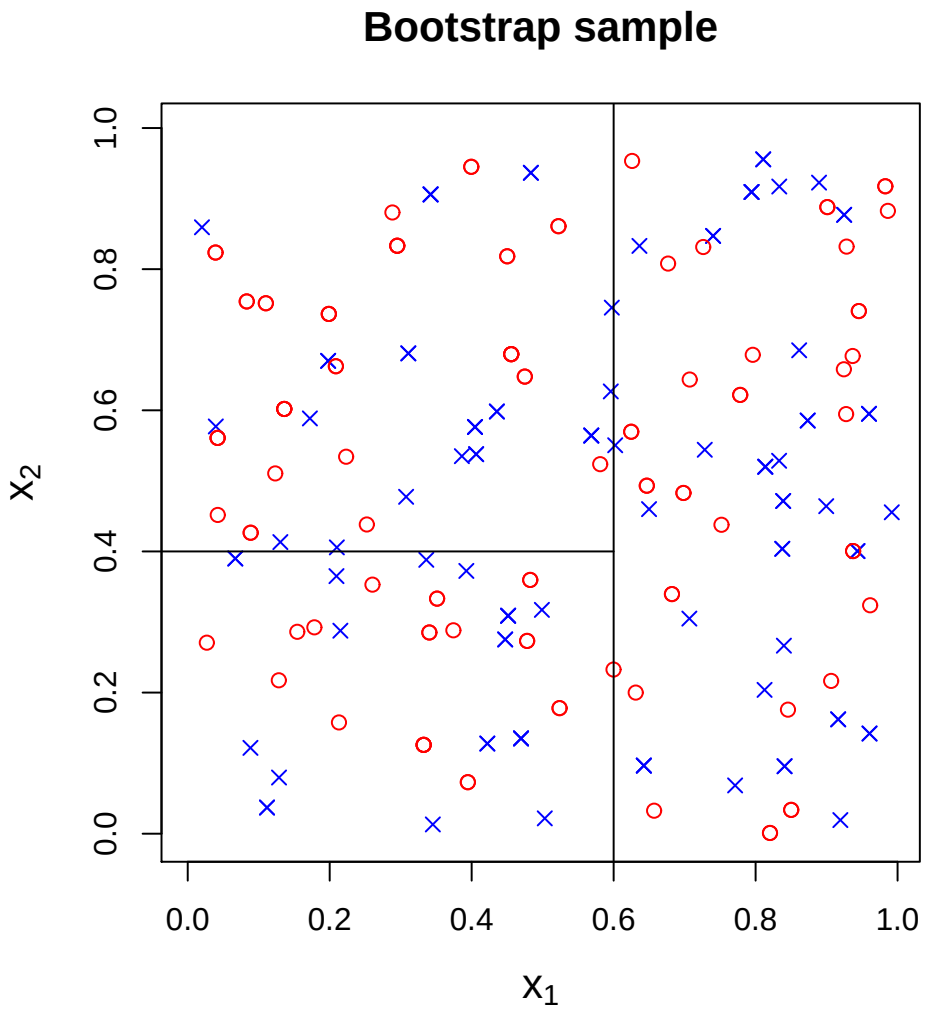
Bootstrap calibration (Loh, 1987, 1991)

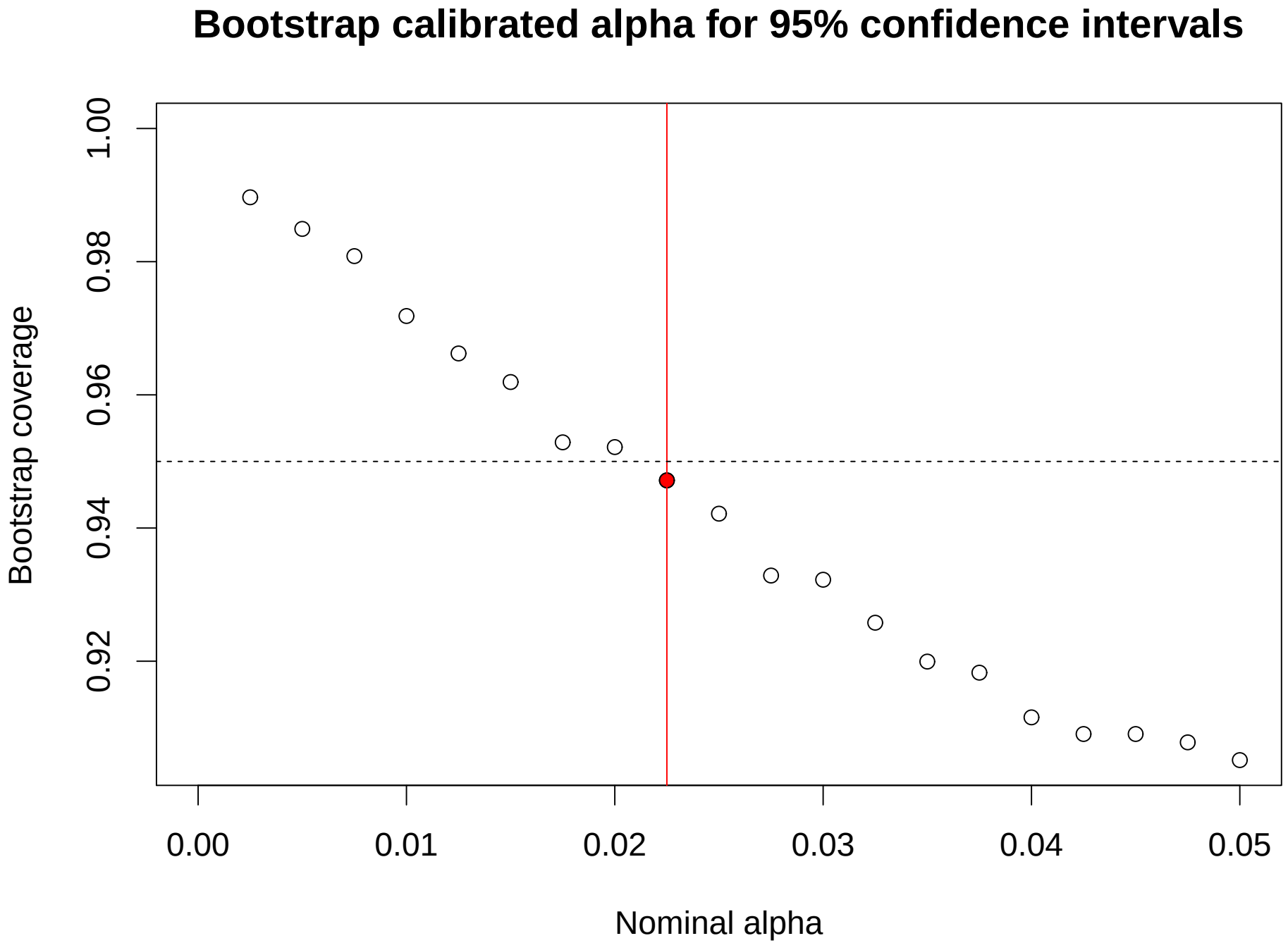
- Naïve intervals too short — do not account for subgroup search
- Need to increase nominal confidence level
- Use bootstrap to estimate true confidence levels
- Increase nominal level of intervals to reach desired level

95% bootstrap calibrated intervals (Loh, 1987)

1. Let F be true (unknown) distribution of data
2. Given sample of data, construct a tree model
3. Given γ, construct a nominal 100γ% interval at each terminal node
4. Let $C(F, \gamma)$ be true average coverage of nominal 100γ% intervals
5. Let γ_F be such that $C(F, \gamma_F) = 0.95$
6. If we know F, construct nominal $100\gamma_F$% intervals and we are finished
7. Because F is unknown, let $\hat{F}$ be its **bootstrap** estimate
8. Use simulation to find **calibrated level** $\gamma_{\hat{F}}$ such that $C(\hat{F}, \gamma_{\hat{F}}) = 0.95$
9. Construct desired intervals at nominal level $\gamma_{\hat{F}}$


Real data
x_2
0.0
0.2
0.4
0.6
0.8
1.0
x_1


Bootstrap sample
x_2
0.0
0.2
0.4
0.6
0.8
1.0
x_1


Bootstrap calibrated alpha for 95% confidence intervals
Bootstrap coverage
1.00
0.98
0.96
0.94
0.92
0.00
0.01
0.02
0.03
0.04
0.05
Nominal alpha

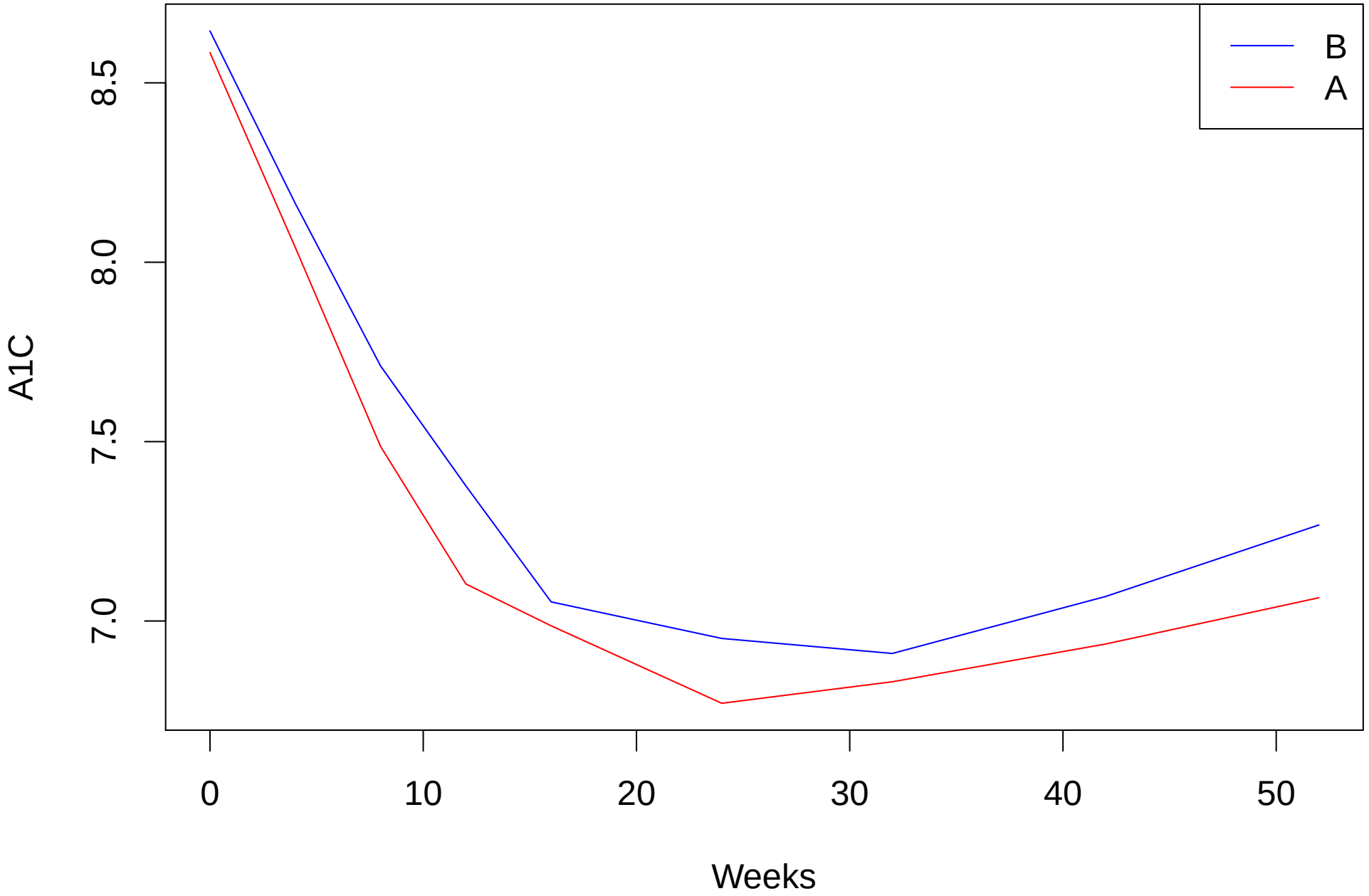
Type 2 diabetes longitudinal study with missing values in responses and covariates

- 1249 subjects from a multi-center, randomized double-blind trial (Charbonnel et al., 2004)
- Subjects randomized to a 52-week treatment period of drug A or drug B
- 24 baseline (time 0) variables measured for each subject as well as their HbA1c at 10 time points (-2, 0, 4, 8, 12, 16, 24, 32, 42, and 52 weeks)
- Analysis based on 747 subjects (364 on A and 383 on B) with HbA1c values at every time point
- Drug A increases amount of insulin produced by the pancreas
- Drug B improves how body uses insulin (“insulin sensitizer”)

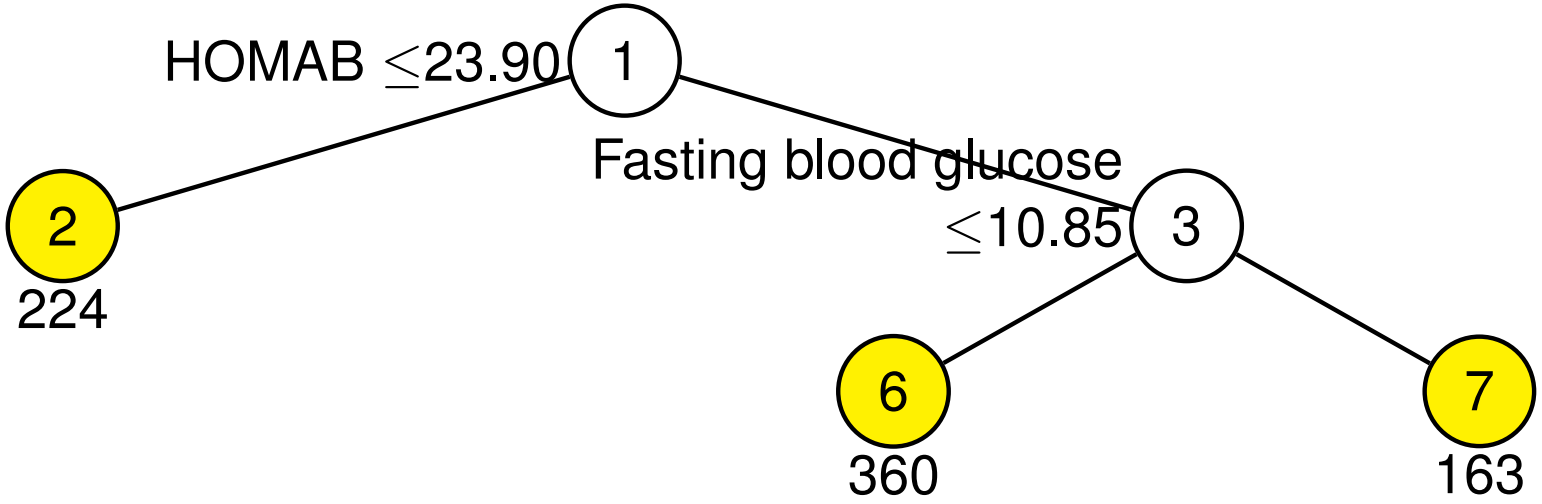
Baseline variables and their missing values

Variable	#Missing	Variable	#Missing
HDL	7	Age	0
LDL	77	Weight	1
Total cholesterol	6	BMI	0
Triglycerides	6	Waist	4
Creatinine	0	A1CBase	0
Fasting insulin	46	HomaS	62
ALT	0	HomaIR	62
AST	0	HomaB	62
GGT	0	Diastolic blood pressure	0
C-peptide	593	Systolic blood pressure	0
Diabetes duration	0	Pulse	0
Fasting blood glucose	0		

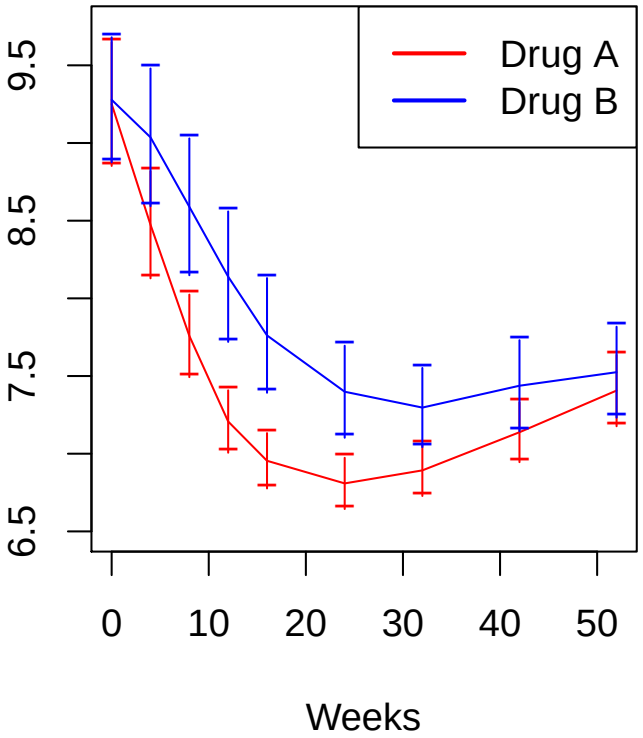
HbA1c means for 747 subjects



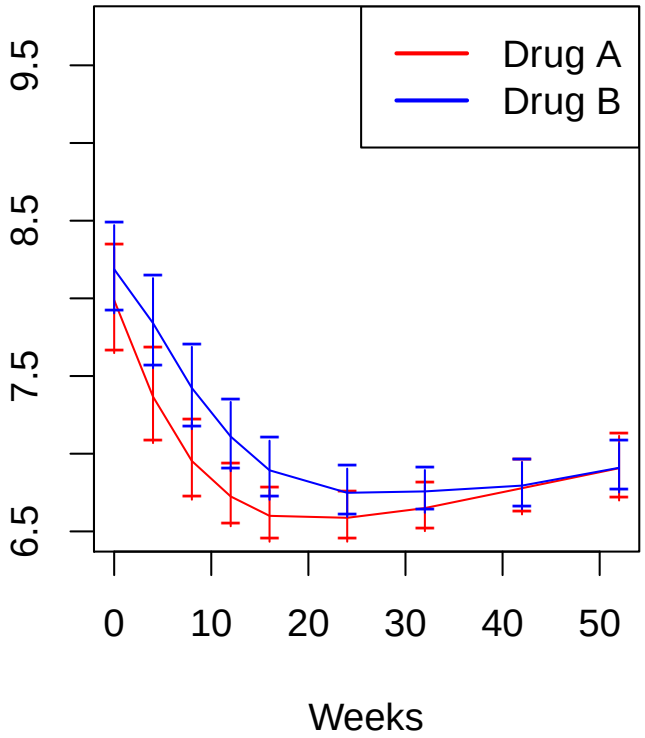
GUIDE tree with 95% bootstrap CIs (Loh et al., 2016)



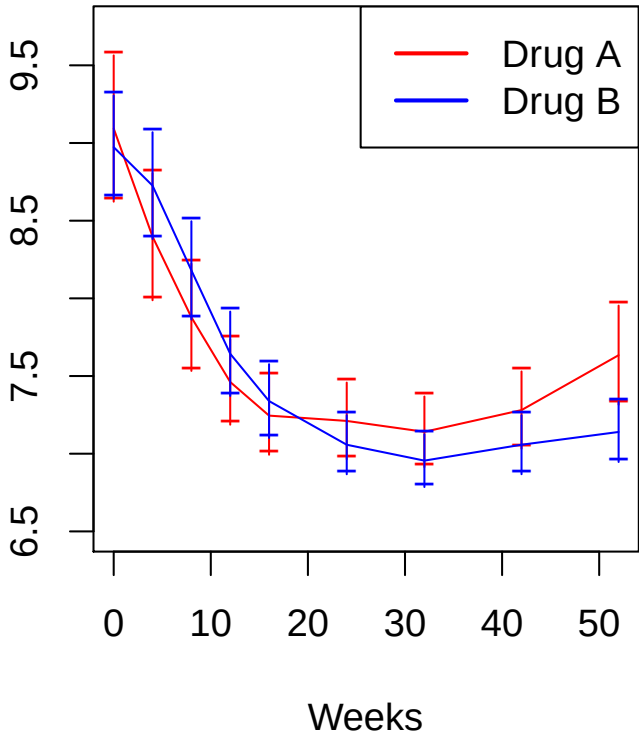
Node 2



Node 6



Node 7



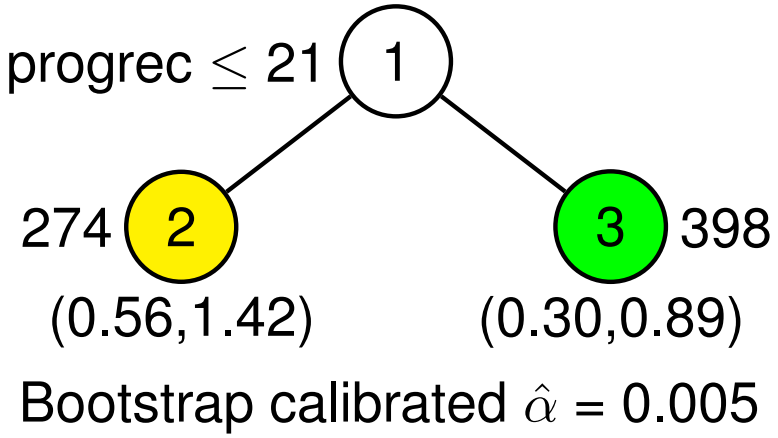
Simulated relative bias of treatment means (simulation SEs in parentheses)

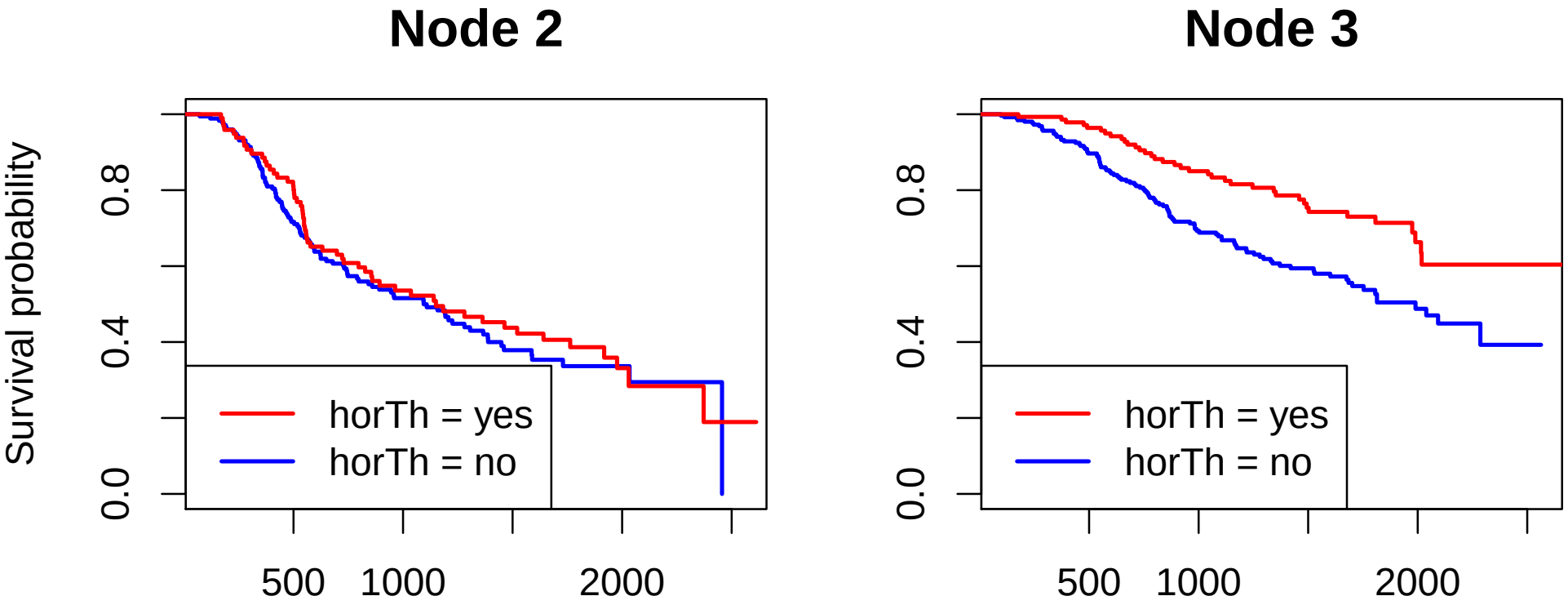
	Average relative bias	
n	Drug A	Drug B
100	2e-04 (4e-04)	6e-04 (4e-04)
500	-6e-04 (2e-04)	6e-04 (2e-04)
1000	-2e-04 (1e-04)	3e-04 (1e-04)

Simulated average coverage of 95% CIs (simulation SEs about 0.002)

	Naïve intervals		Bootstrap calibrated intervals		
n	Drug A	Drug B	Drug A	Drug B	Overall
100	0.917	0.920	0.930	0.947	0.939
500	0.930	0.936	0.938	0.959	0.949
1000	0.932	0.942	0.940	0.962	0.951

95% bootstrap intervals for RR (therapy vs none)





Simulated average coverage of 95% CIs for treatment effect (without linear prognostic control) for breast cancer data ($\pm$ 2 simulation SEs)

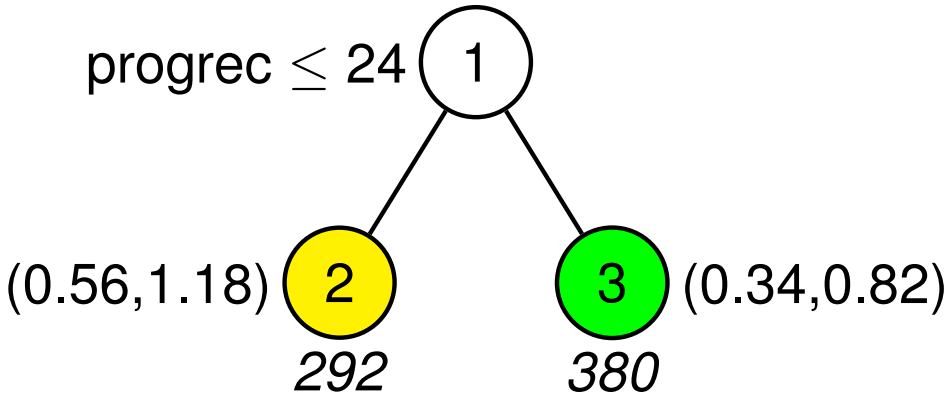
Naïve t interval	0.837 $\pm$ 0.022
Bootstrap calibrated interval	0.935 $\pm$ 0.015

based on 1100 simulation trials each, with 25 bootstraps

Linear control of prognostic variables

- Many studies include prognostic variables (e.g., age, tumor size)
- Treatment randomization balances the overall effects of these variables
- But balance may be upset within subgroups
- Leads to confounding between prognostic and treatment effects
- Solution: control for linear prognostic effects within subgroups

95% bootstrap intervals for RR due to horTh with linear control of prognostic variables



Bootstrap calibrated $\hat{\alpha}$ = 0.0225

	Node 2		Node 3	
	coef	p-value*	coef	p-value*
constant	0.010	0.934	-0.335	0.005
pnodes	0.087	0.000	0.040	0.000
horTh=yes	-0.209	0.206	-0.643	0.001

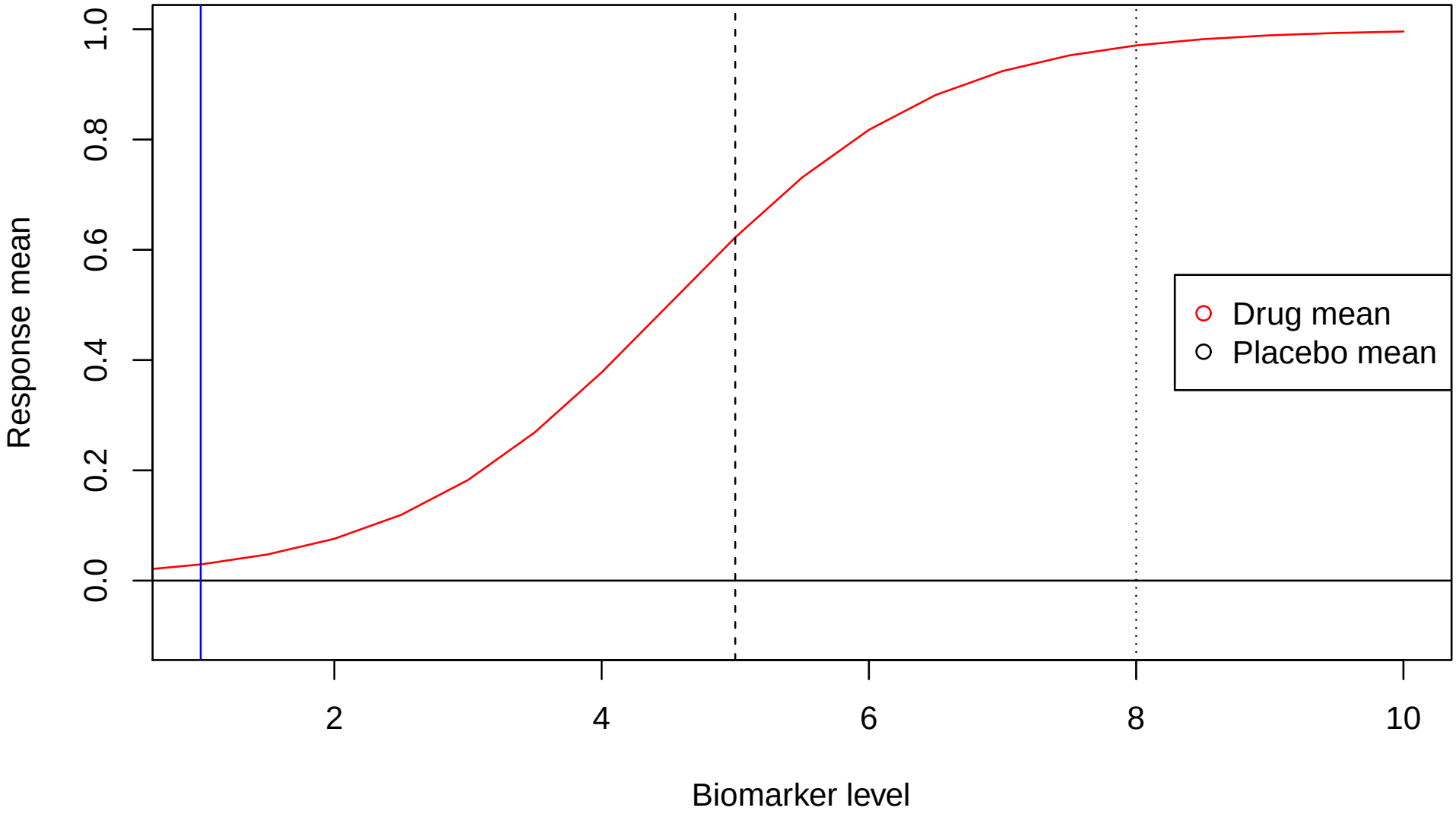
* unadjusted p-values

Coverage ($\pm$ 2 SEs) of 95% CIs for treatment effect with prognostic control for breast cancer data

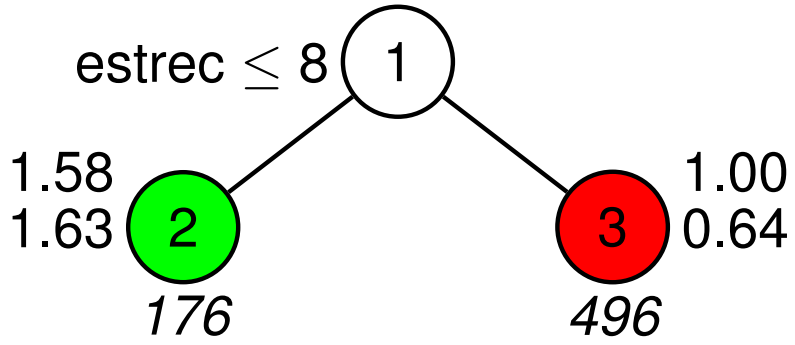
Naïve t interval	0.892 $\pm$ 0.018
Bootstrap calibrated interval	0.960 $\pm$ 0.011

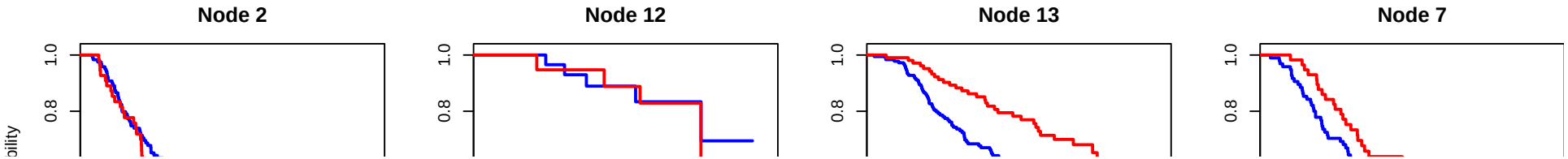
based on 1200 simuation trials with 25 bootstraps per trial

Often overlooked fact: subgroups are not unique

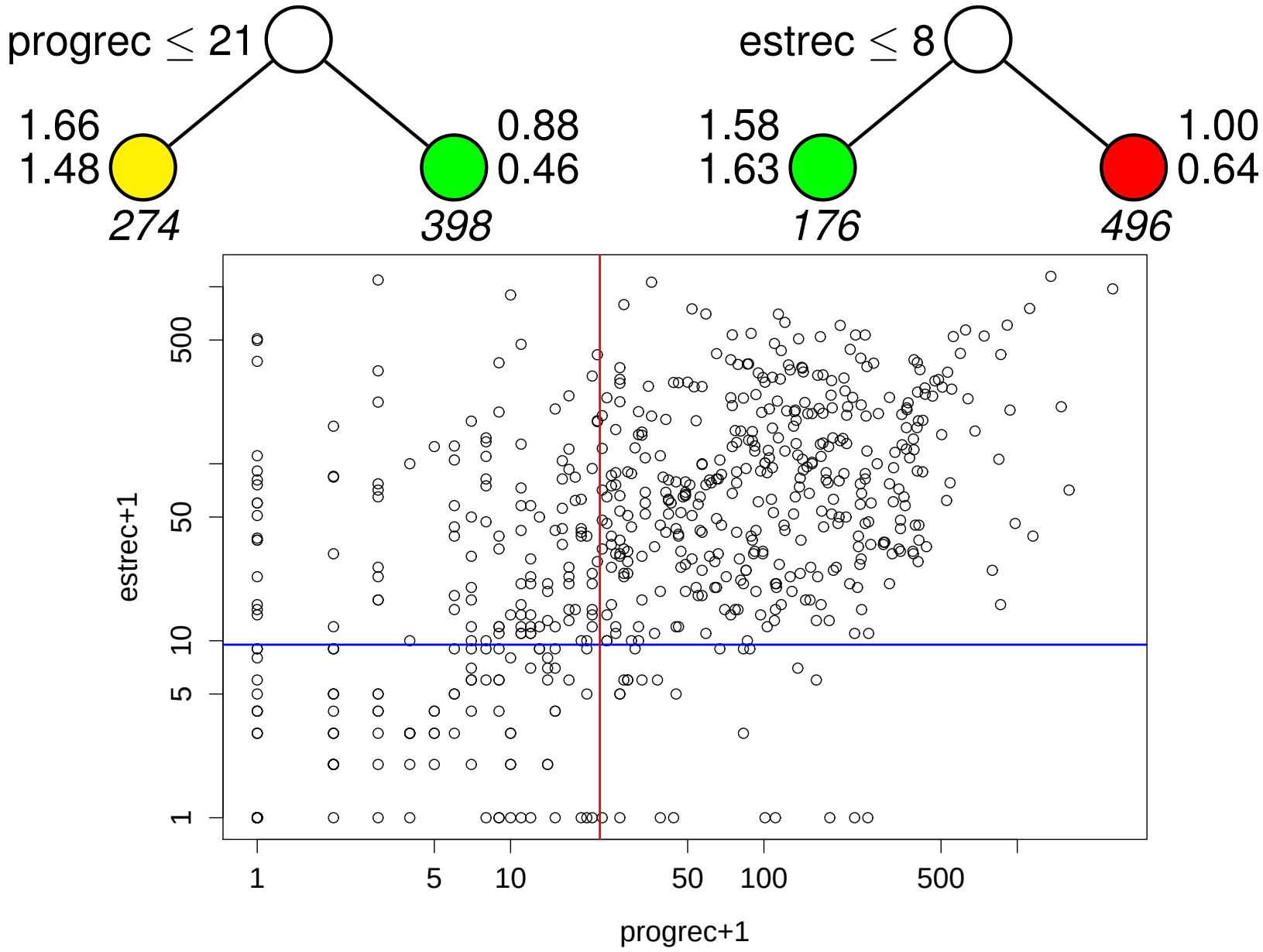


GUIDE model for breast cancer without progrec





Non-unique subgroups

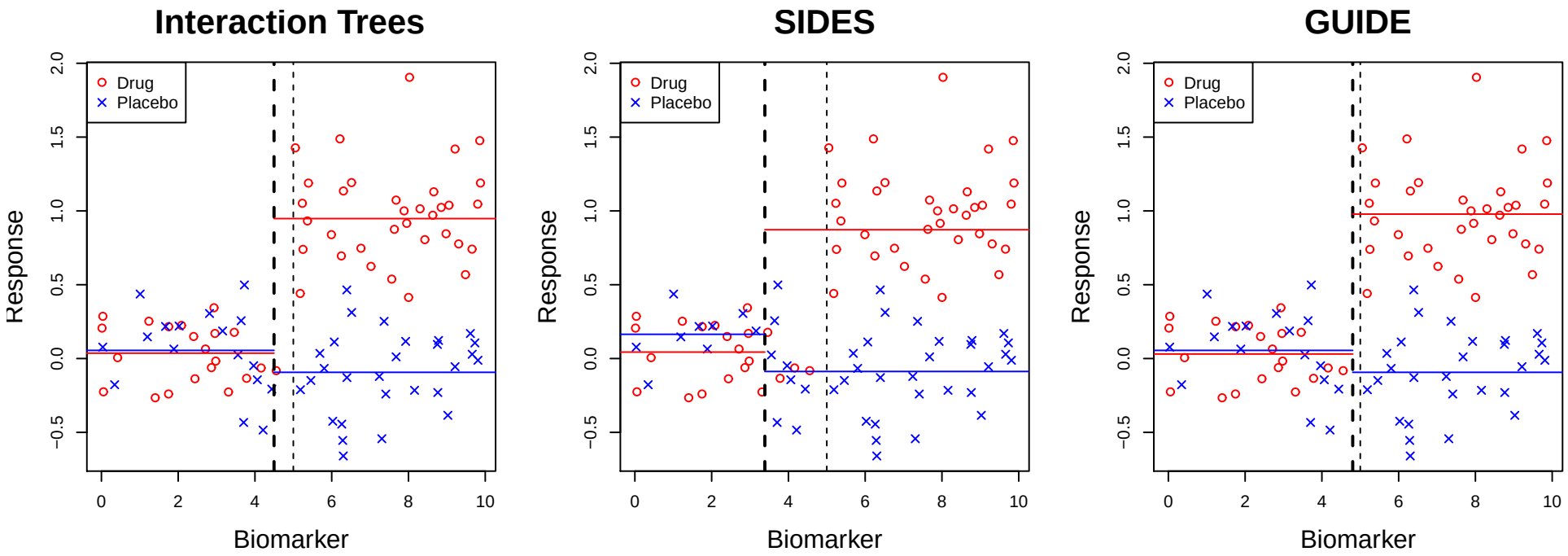


Conclusions

1. Subgroups are not unique: trade-off between subgroup size and effect size
2. Greedy methods such as **Interaction Trees** (Su et al., 2008), **Virtual Twins** (Foster et al., 2011) and **SIDES** (Lipkovich et al., 2011; Lipkovich and Dmitrienko, 2014) yield optimistically biased estimates that require difficult (and typically unsuccessful) bias corrections
3. **GUIDE** is unbiased — does not over-estimate treatment differences
4. **GUIDE** can optionally control for prognostic variables within subgroups
5. **SIDES** uses permutation p-values to assess statistical significance—but such tests of **complete absence** of treatment effect may have low power if treatment has an effect in **some** subgroups but not in others
6. Best way to assess statistical significance is bootstrap intervals

IT and SIDES vs GUIDE

(true subgroup marked by black dashed line)



Interaction Trees maximizes t statistic

SIDES minimizes p-value

GUIDE minimizes sum of squared residuals

References

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Chapman & Hall/CRC.

Charbonnel, B. H.and Matthews, D. R., Schernthaner, G., Hanefeld, M., and Brunetti, P. (2004). A long-term comparison of Pioglitazone and Gliclazide in patients with Type 2 diabetes mellitus: a randomized, double-blind, parallel-group comparison trial. *Diabetic Medicine*, 22:399–405.

Chaudhuri, P., Huang, M.-C., Loh, W.-Y., and Yao, R. (1994). Piecewise-polynomial regression trees. *Statistica Sinica*, 4:143–167.

Chaudhuri, P., Lo, W.-D., Loh, W.-Y., and Yang, C.-C. (1995). Generalized regression trees. *Statistica Sinica*, 5:641–666.

Dusseldorp, E. and Van Mechelen, I. (2014). Qualitative interaction trees: a tool to identify qualitative treatment-subgroup interactions. *Statistics in Medicine*, 33:219–237.

Foster, J. C., Taylor, J. M. G., and Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, 30:2867–2880.

Hosmer, Jr., D. W., Lemeshow, S., and May, S. (2008). *Applied Survival Analysis*. Wiley, New York, 2nd edition.

Italiano, A. (2011). Prognostic or predictive? It's time to get back to definitions! *Journal of Clinical Oncology*, 29:4718.

Kim, H. and Loh, W.-Y. (2001). Classification trees with unbiased multiway splits. *Journal of the American Statistical Association*, 96:589–604.

Lipkovich, I. and Dmitrienko, A. (2014). Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using SIDES. *Journal of Biopharmaceutical Statistics*, 24:130–153.

Lipkovich, I., Dmitrienko, A., Denne, J., and Enas, G. (2011). Subgroup identification based on differential effect search — a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine*, 30:2601–2621.

Loh, W.-Y. (1987). Calibrating confidence coefficients. *Journal of the American Statistical Association*, 82:155–162.

Loh, W.-Y. (1991). Bootstrap calibration for confidence interval construction and selection. *Statistica Sinica*, 1:477–491.

Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, 12:361–386.

Loh, W.-Y. (2009). Improving the precision of classification trees. *Annals of Applied Statistics*, 3:1710–1737.

Loh, W.-Y. (2012). Variable selection for classification and regression in large *p*, small *n* problems. In Barbour, A., Chan, H. P., and Siegmund, D., editors, *Probability Approximations and Beyond*, volume 205 of *Lecture Notes in Statistics—Proceedings*, pages 133–157, New York. Springer.

Loh, W.-Y., Fu, H., Man, M., Champion, V., and Yu, M. (2016). Identification of subgroups with differential treatment effects for longitudinal and multiresponse variables. *Statistics in Medicine*, 35:4837–4855.

Loh, W.-Y., He, X., and Man, M. (2015). A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in Medicine*, 34:1818–1833.

Loh, W.-Y. and Shih, Y.-S. (1997). Split selection methods for classification trees. *Statistica Sinica*, 7:815–840.

Loh, W.-Y. and Vanichsetakul, N. (1988). Tree-structured classification via generalized discriminant analysis (with discussion). *Journal of the American Statistical Association*, 83:715–728.

Loh, W.-Y. and Zheng, W. (2013). Regression trees for longitudinal and multiresponse data. *Annals of Applied Statistics*, 7:495–522.

Milik, M., Sauer, D., Brunmark, A. P., Yuan, L., Vitiello, A., Jackson, M. R., Peterson, P. A., Skolnick, J., and Glass, C. A. (1998). Application of an artificial neural network to predict specific class I MHC binding peptide sequences. *Nature Biotechnology*, 16:753–756.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2:110–114.

Su, X., Tsai, C. L., Wang, H., Nickerson, D. M., and Bogong, L. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10:141–158.

Su, X., Zhou, T., Yan, X., Fan, J., and Yang, S. (2008). Interaction trees with censored survival data. *International Journal of Biostatistics*, 4. Article 2.

Wilson, E. B. and Hilferty, M. M. (1931). The distribution of chi-square. *Proceedings of the National Academy of Sciences of the United States of America*, 17:684–688.

Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, 17:492–514.