Classification and Regression Trees and Forests

Wei-Yin Loh
Department of Statistics
University of Wisconsin–Madison
loh@stat.wisc.edu
http://www.stat.wisc.edu/∼loh/

Interface Conference on Applied Statistics, Washington, D.C., 9–10 December, 2013

Course outline

1. **Motivating examples**
   (a) Least squares regression: impact of air pollution on house prices
   (b) Poisson regression: defects in soldering circuit boards
   (c) Multiresponse data: interactions of variables in production of concrete
   (d) Longitudinal data: hourly wages of high-school dropouts
   (e) Censored data and differential treatment effects: breast cancer survival
   (f) Simple classification: Fisher's iris data
   (g) Classification with unequal costs: attitudes towards mammography
   (h) Unbalanced classes: characterizing dissatisfied credit card holders

2. **Classification tree algorithms**
   (a) THAID (Messenger and Mandell, 1972), CART (Breiman et al., 1984), RPART (Therneau and Atkinson, 2013, 2012)
   (b) FACT (Loh and Vanichsetakul, 1988), QUEST (Loh and Shih, 1997), CRUISE (Kim and Loh, 2001, 2003), GUIDE (Loh, 2009)

(c) C4.5 (Quinlan, 1993), CHAID (Kass, 1980), CTREE (Hothorn et al., 2006)

(d) More examples: peptide binding; fish identification; car prediction

(e) Missing values, selection bias, accuracy, speed, and tree complexity

3. **Regression tree algorithms**

(a) Piecewise constant least squares models: AID (Morgan and Sonquist, 1963), CART, RPART, GUIDE (Loh, 2002)

(b) Piecewise linear least squares, quantile regression, subgroup identification of differential treatment effects, and longitudinal data effects: GUIDE (Loh and Zheng, 2013)

(c) Others: M5 (Quinlan, 1992), MOB (Zeileis et al., 2008)

(d) More examples: college tuition; primary biliary cirrhosis of the liver; progression of CD4 counts in AIDS

(e) Missing values, selection bias, accuracy, speed, and tree complexity

4. **Conclusion**

Learning objectives

1. Recognize the fundamental difference between
   (a) inference-based approach of traditional statistical methods and
   (b) data description and prediction objectives of decision tree methods
2. Discover the ways tree methods enrich the statistician's toolbox
3. Know the key ideas that differentiate decision tree algorithms
4. Observe their impact on performance (e.g., computational speed, selection bias) and extensibility (e.g., multiresponse data, missing values)
5. Compare the strengths, weaknesses, and limitations of each algorithm

Classification of tree algorithms by purpose

1. Binary classification trees—CART, RPART, CTREE, QUEST, GUIDE
2. Non-binary classification trees—CHAID, C4.5, CRUISE
3. Piecewise-constant least-squares trees—CART, RPART, CTREE, GUIDE
4. Piecewise-linear least-squares regression trees—M5, GUIDE, CTREE
5. Least-median-of-squares regression trees—GUIDE
6. Quantile regression trees—GUIDE
7. Poisson regression trees—RPART, GUIDE, MOB
8. Logistic regression trees—LOTUS (Chan and Loh, 2004), MOB
9. Censored response variables—RPART, GUIDE, MOB
10. Multivariate and longitudinal response variables—GUIDE
11. Tree ensembles—GUIDE, CTREE, MOB, random forest (Breiman, 2001), random survival forest (Ishwaran et al., 2006)

Free software

- C4.5—`www.rulequest.com/Personal/c4.5r8.tar.gz`; see also `www.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/c4.5/tutorial.html`
- CART, C4.5, M5, etc.—`www.cs.waikato.ac.nz/~ml/weka/`
- CRUISE, GUIDE, LOTUS, QUEST—`www.stat.wisc.edu/~loh/`
- RPART, CTREE, MOB, PARTY, RandomForest —`cran.us.r-project.org/`
- LaTeX (text processing package)—`http://www.ctan.org/`
  CRUISE, GUIDE, LOTUS, and QUEST produce LaTeX tree diagrams

Some review papers

1. Lemon et al. (2003), Classification and regression tree analysis in public health: methodological review and comparison with logistic regression, *Annals of Behavioral Medicine*
2. Loh (2008a), Classification and regression tree methods, *Encyclopedia of Statistics in Quality and Reliability*
3. Merkle and Shaffer (2011), Binary recursive partitioning: background, methods, and application to psychology, *British Journal of Mathematical and Statistical Psychology*
4. Loh (2011), Classification and regression trees, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*
5. Loh (2013), Fifty years of classification and regression trees (with discussion), *International Statistical Review*

Linear regression: 1970 Boston housing data (Harrison and Rubinfeld, 1978; Belsley et al., 1980)

Var	Definition	Var	Definition
ID	census tract number	TOWN	township (92 values)
MEDV	median value in $1000	AGE	% built before 1940
CRIM	per capita crime rate	DIS	distance to employment centers
ZN	% zoned for lots $>$ 25K sq.ft.	RAD	accessibility to radial highways
INDUS	% nonretail business	TAX	property tax rate per $10000
CHAS	1 on Charles River, 0 else	PT	pupil/teacher ratio
NOX	nitrogen oxide conc. (p.p.10^9)	B	$($% black - 63$)^2$/10
RM	average number of rooms	LSTAT	% lower-status population

Data: 506 observations (census tracts) in the greater Boston area

Objective: To examine the impact of air pollution on house price

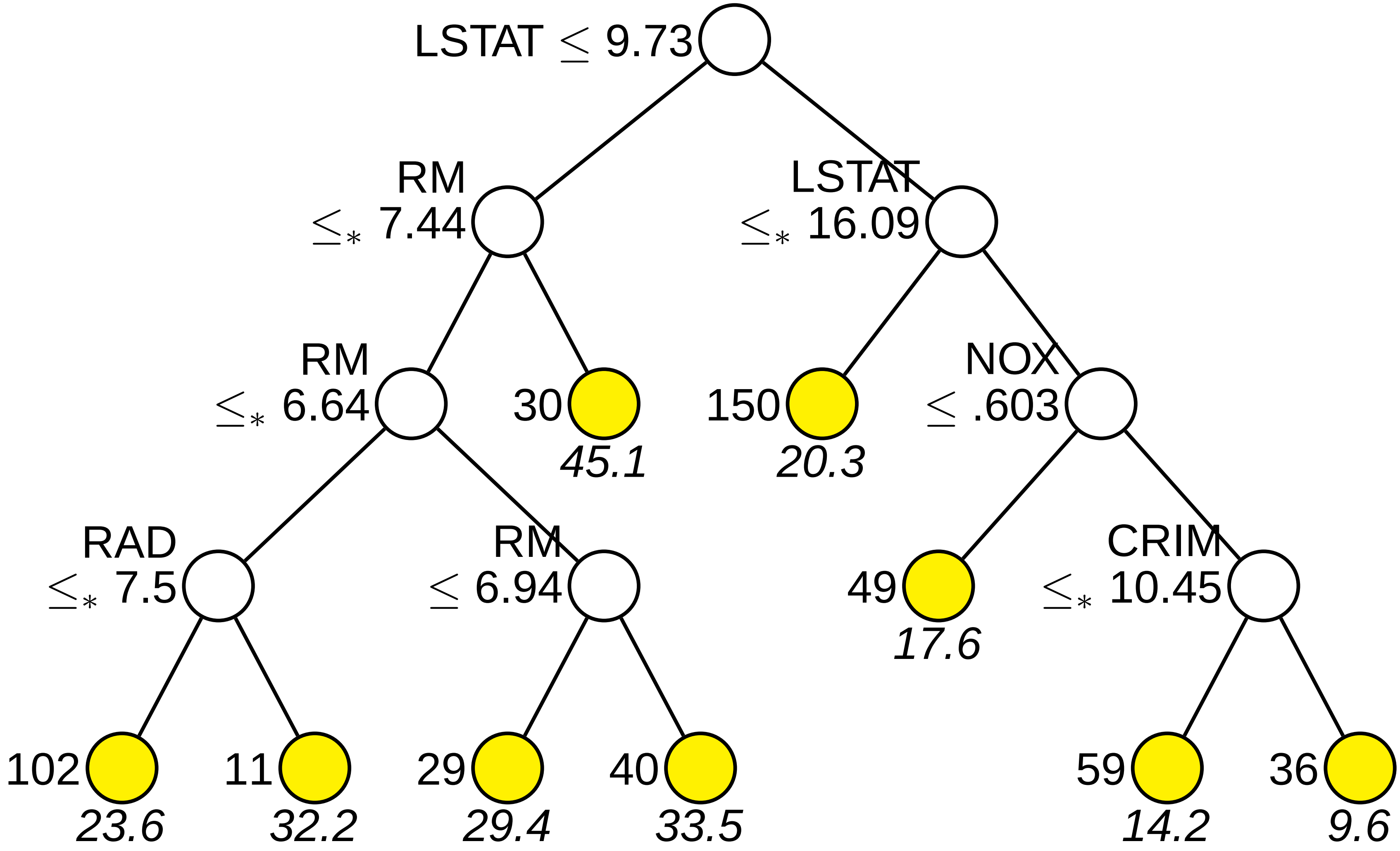
Harrison & Rubinfeld model for log(MEDV)

X	β	t	ρ	X	β	t	ρ
Constant	4.6	30.0		AGE	7.1E-5	0.1	-0.5
CRIM	-1.2E-2	-9.6	-0.5	log(DIS)	-2.0E-1	-6.0	0.4
ZN	9.2E-5	0.2	0.4	log(RAD)	9.0E-2	4.7	-0.4
INDUS	1.8E-4	0.1	-0.5	TAX	-4.2E-4	-3.5	-0.6
CHAS	9.2E-2	2.8	0.2	PT	-3.0E-2	-6.0	-0.5
$\texttt{NOX}^2$	-6.4E-1	-5.7	-0.5	B	3.6E-4	3.6	0.4
$\texttt{RM}^2$	6.3E-3	4.8	0.6	log(LSTAT)	-3.7E-1	-15.2	-0.8

β = coefficient, t = t-statistic, ρ = corr(X, Y)

What can we conclude from this model?

GUIDE piecewise constant model for MEDV

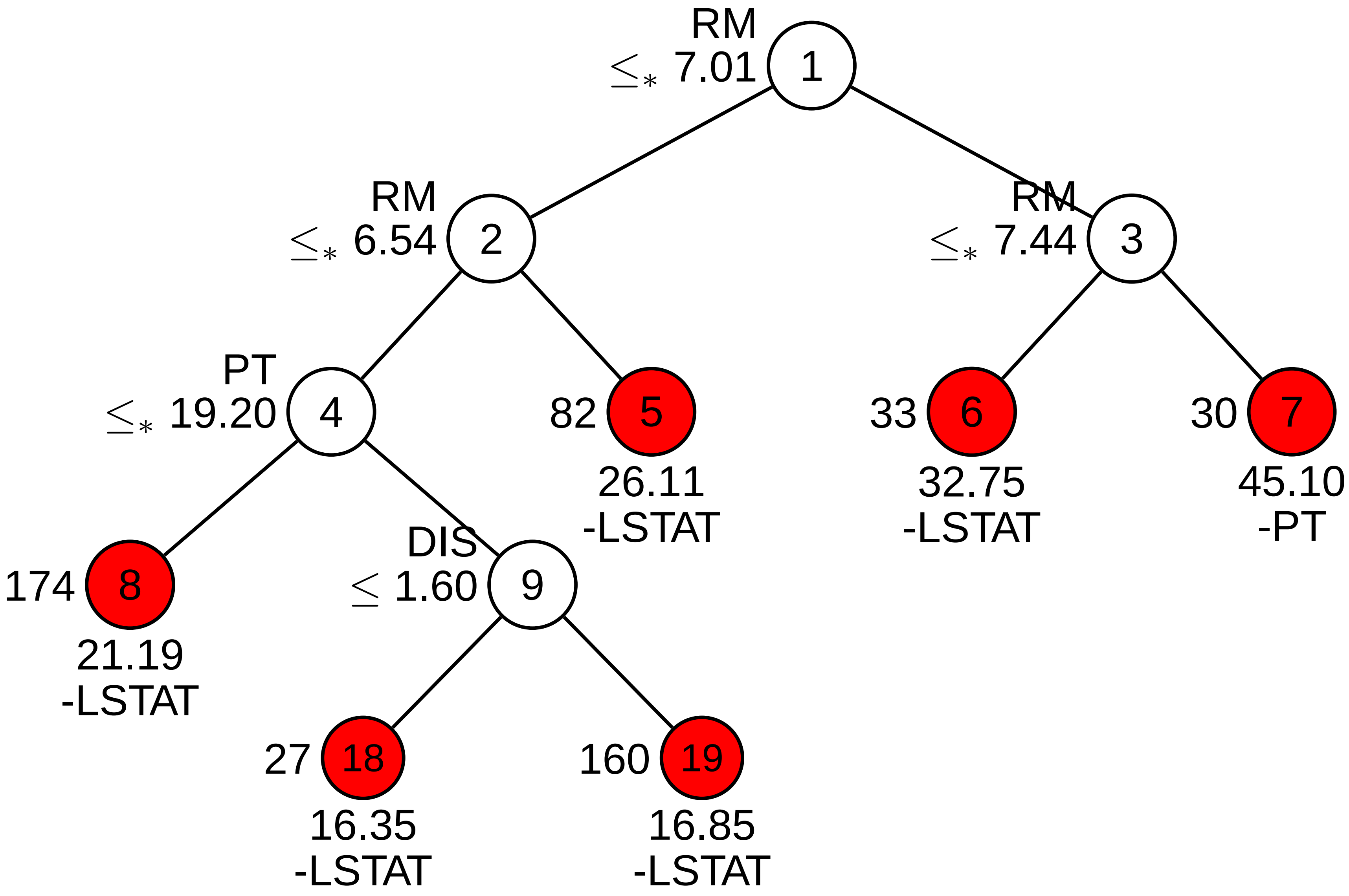


Sample means and sample sizes below and beside nodes.

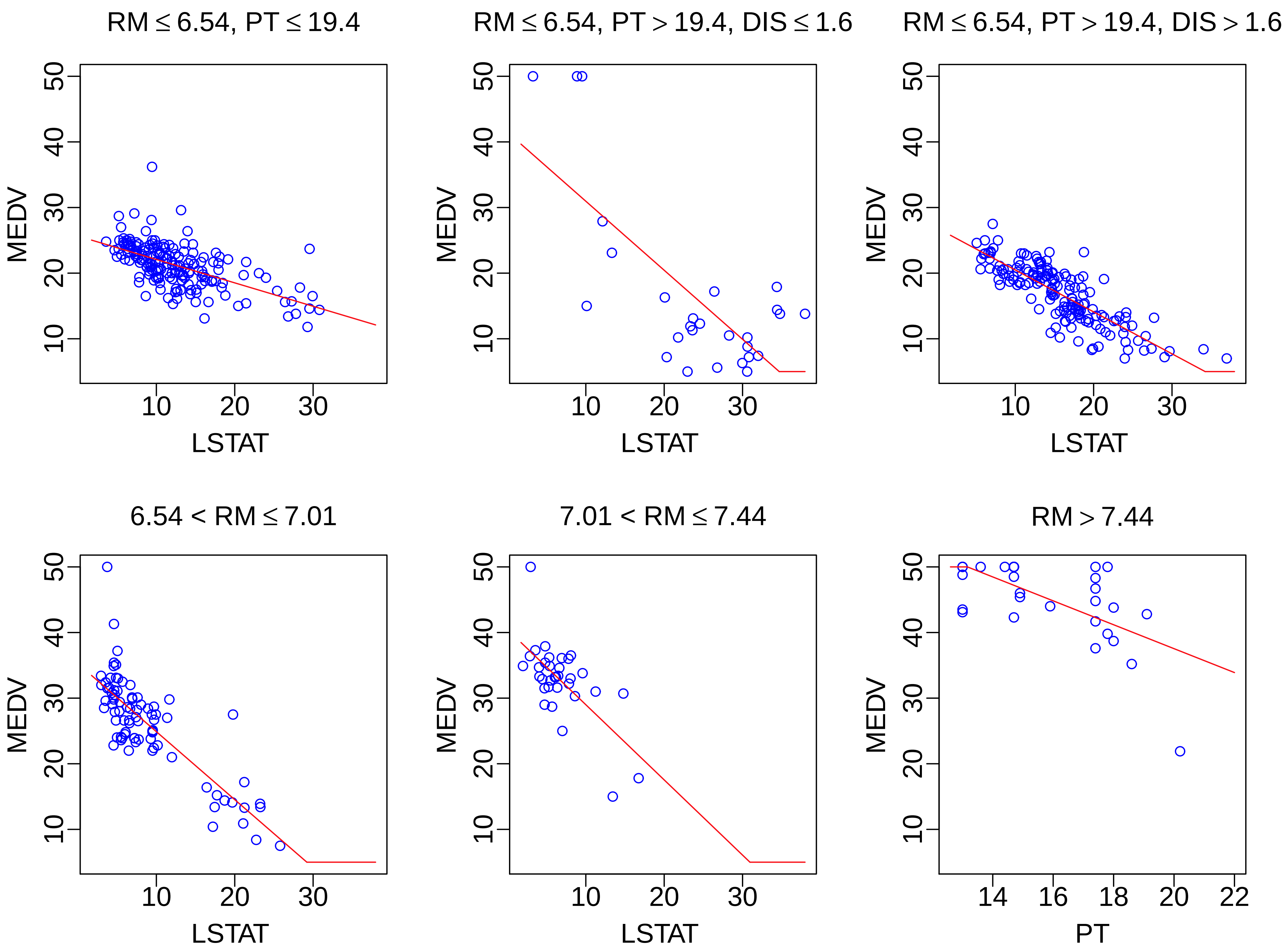
At each intermediate node, a case goes left if and only if the condition is true.

Symbol “$\leq_*$” means “$\leq$ or missing.”

GUIDE piecewise simple linear model for MEDV



Mean MEDV and signed linear predictor beneath each node



RM ≤ 6.54, PT ≤ 19.4

RM ≤ 6.54, PT > 19.4, DIS ≤ 1.6

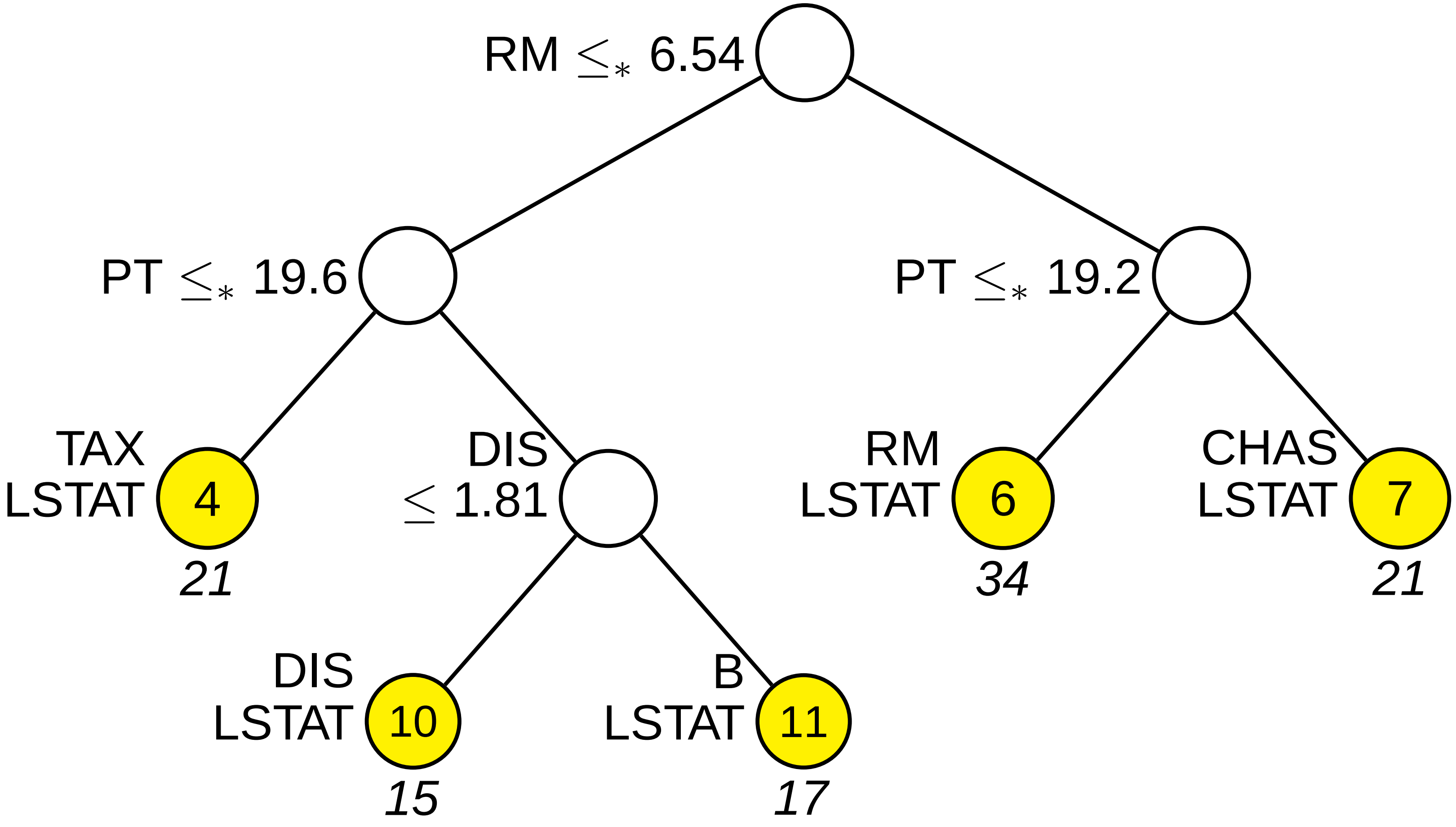
RM ≤ 6.54, PT > 19.4, DIS > 1.6

6.54 < RM ≤ 7.01

7.01 < RM ≤ 7.44

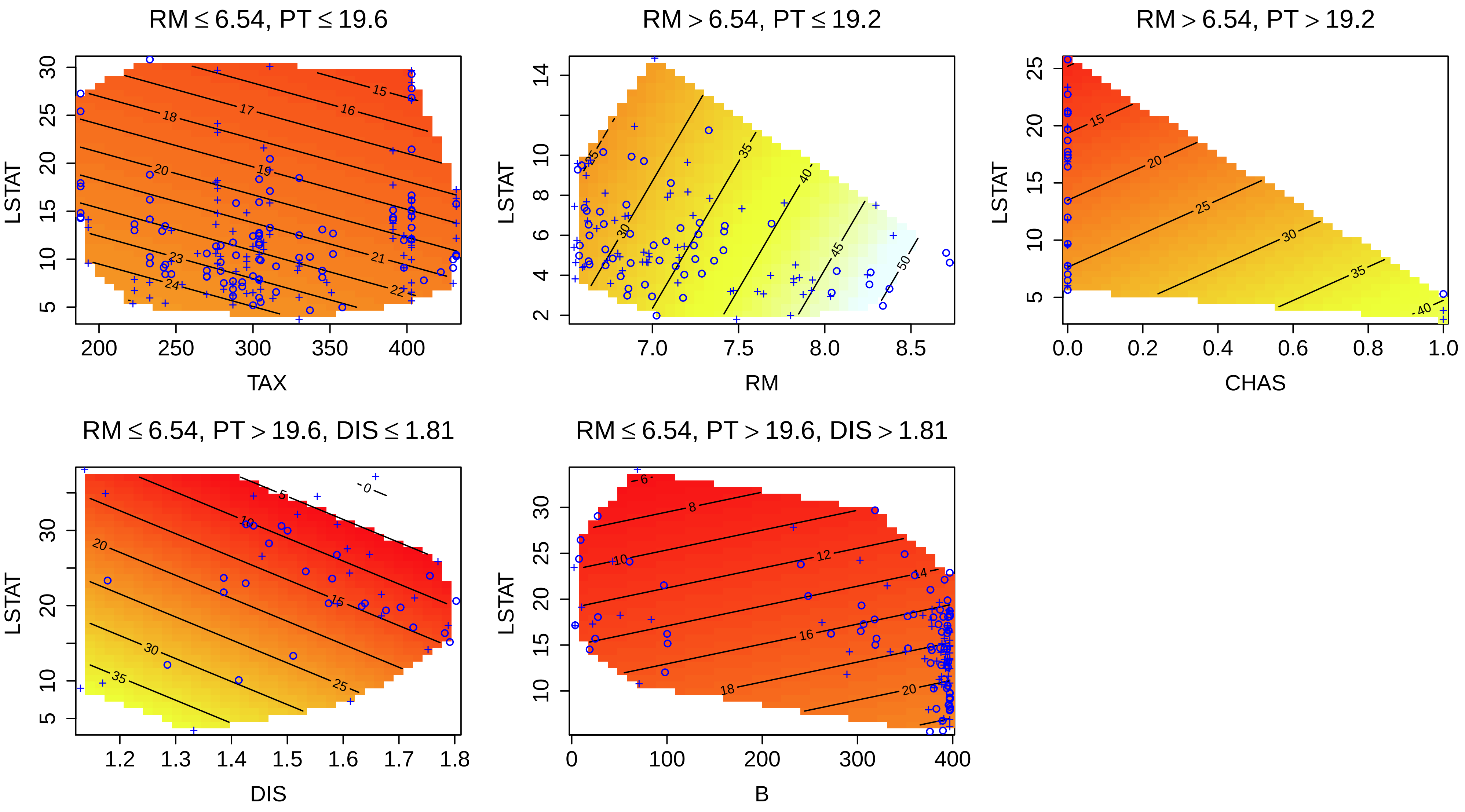
RM > 7.44

GUIDE piecewise two-variable model for MEDV



Mean MEDV beneath each node

Data and fits in GUIDE two-variable model



RM ≤ 6.54, PT ≤ 19.6

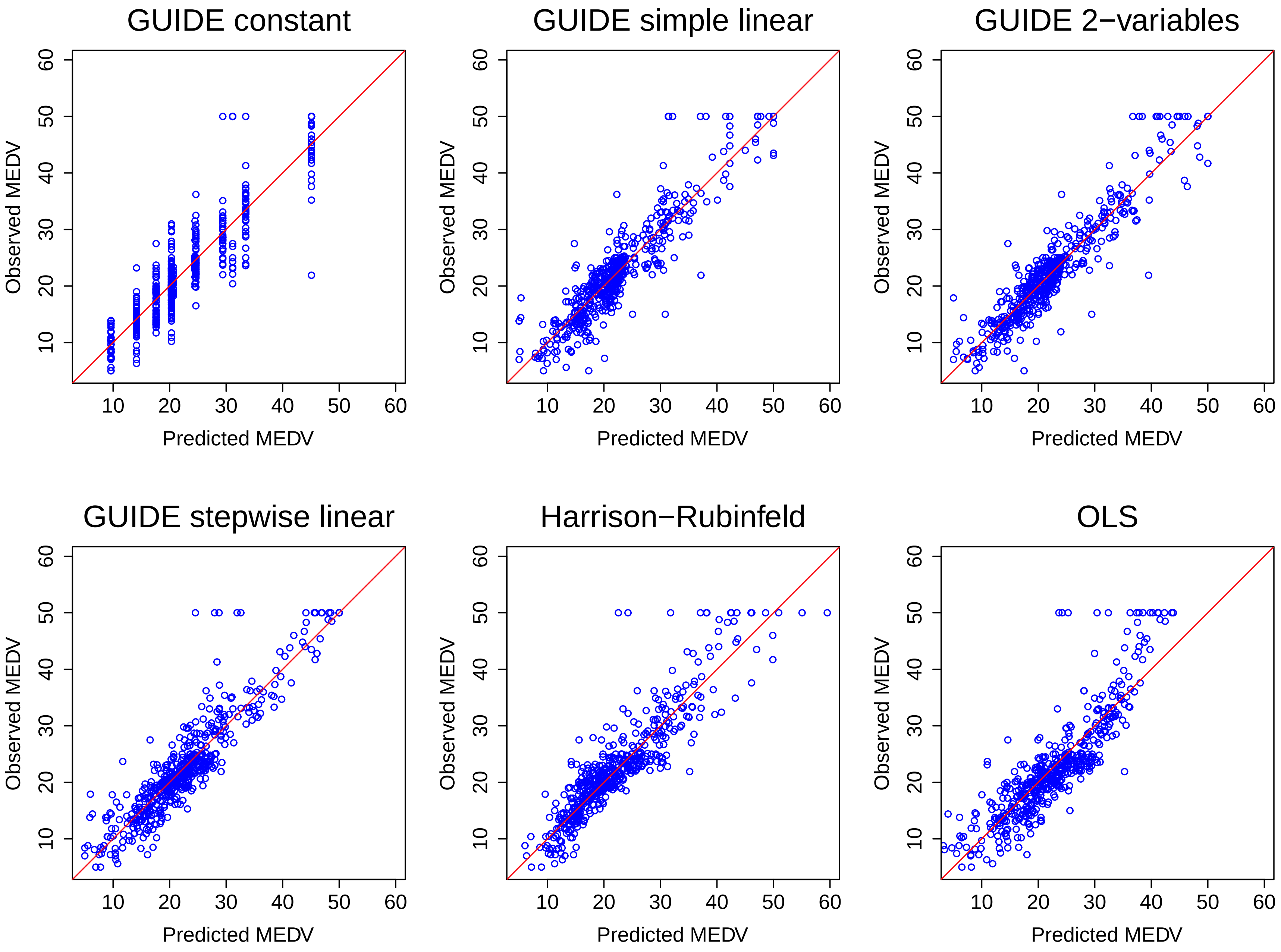
RM > 6.54, PT ≤ 19.2

RM > 6.54, PT > 19.2

RM ≤ 6.54, PT > 19.6, DIS ≤ 1.81

RM ≤ 6.54, PT > 19.6, DIS > 1.81

Comparison of models



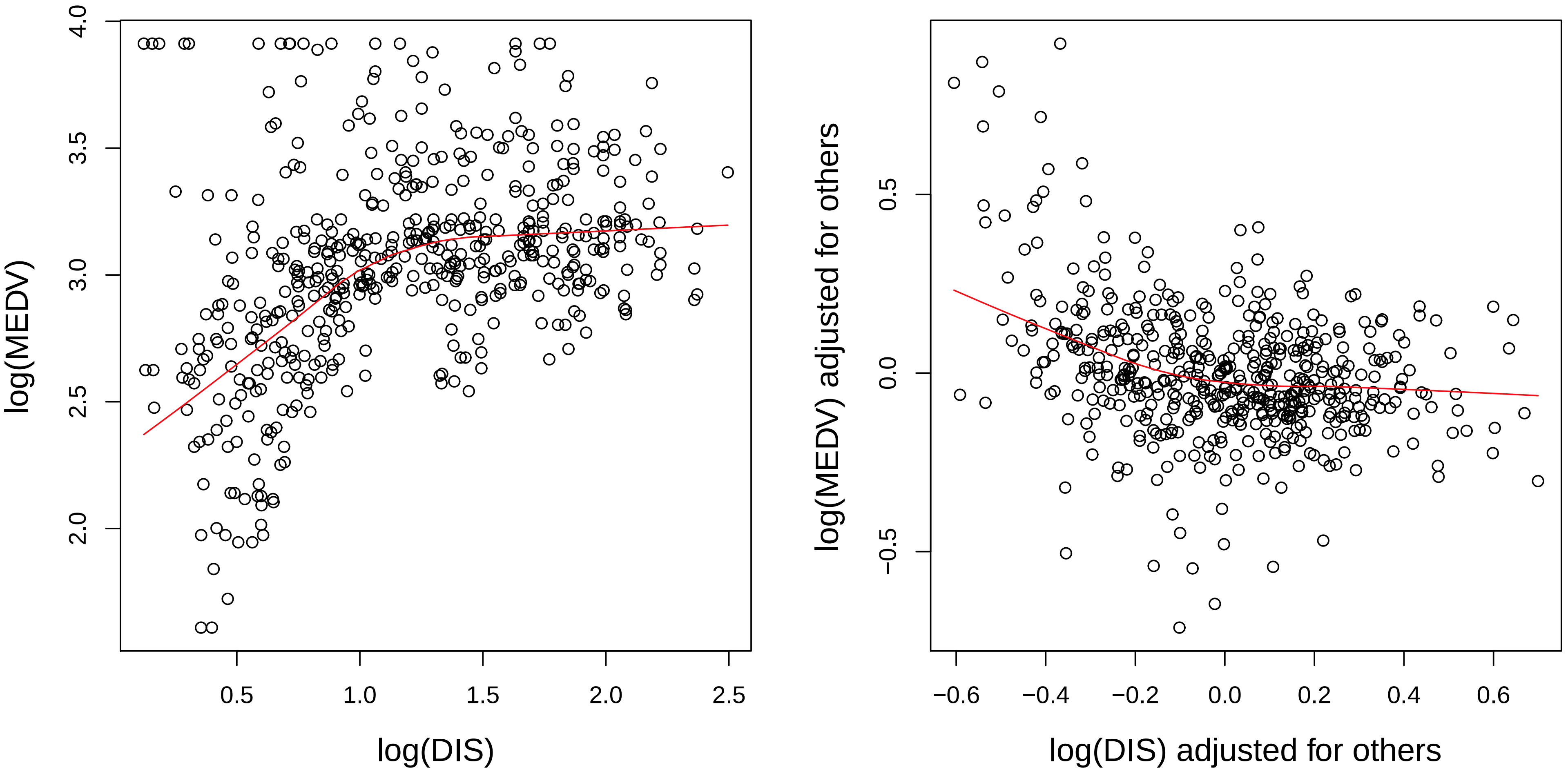
Difficulties in interpreting regression coefficients: Harrison & Rubinfeld model for log(MEDV)

X	β	t	ρ	X	β	t	ρ
Constant	4.6	30.0		AGE	7.1E-5	0.1	-0.5
CRIM	-1.2E-2	-9.6	-0.5	log(DIS)	-2.0E-1	-6.0	0.4
ZN	9.2E-5	0.2	0.4	log(RAD)	9.0E-2	4.7	-0.4
INDUS	1.8E-4	0.1	-0.5	TAX	-4.2E-4	-3.5	-0.6
CHAS	9.2E-2	2.8	0.2	PT	-3.0E-2	-6.0	-0.5
NOX^2	-6.4E-1	-5.7	-0.5	B	3.6E-4	3.6	0.4
RM^2	6.3E-3	4.8	0.6	log(LSTAT)	-3.7E-1	-15.2	-0.8

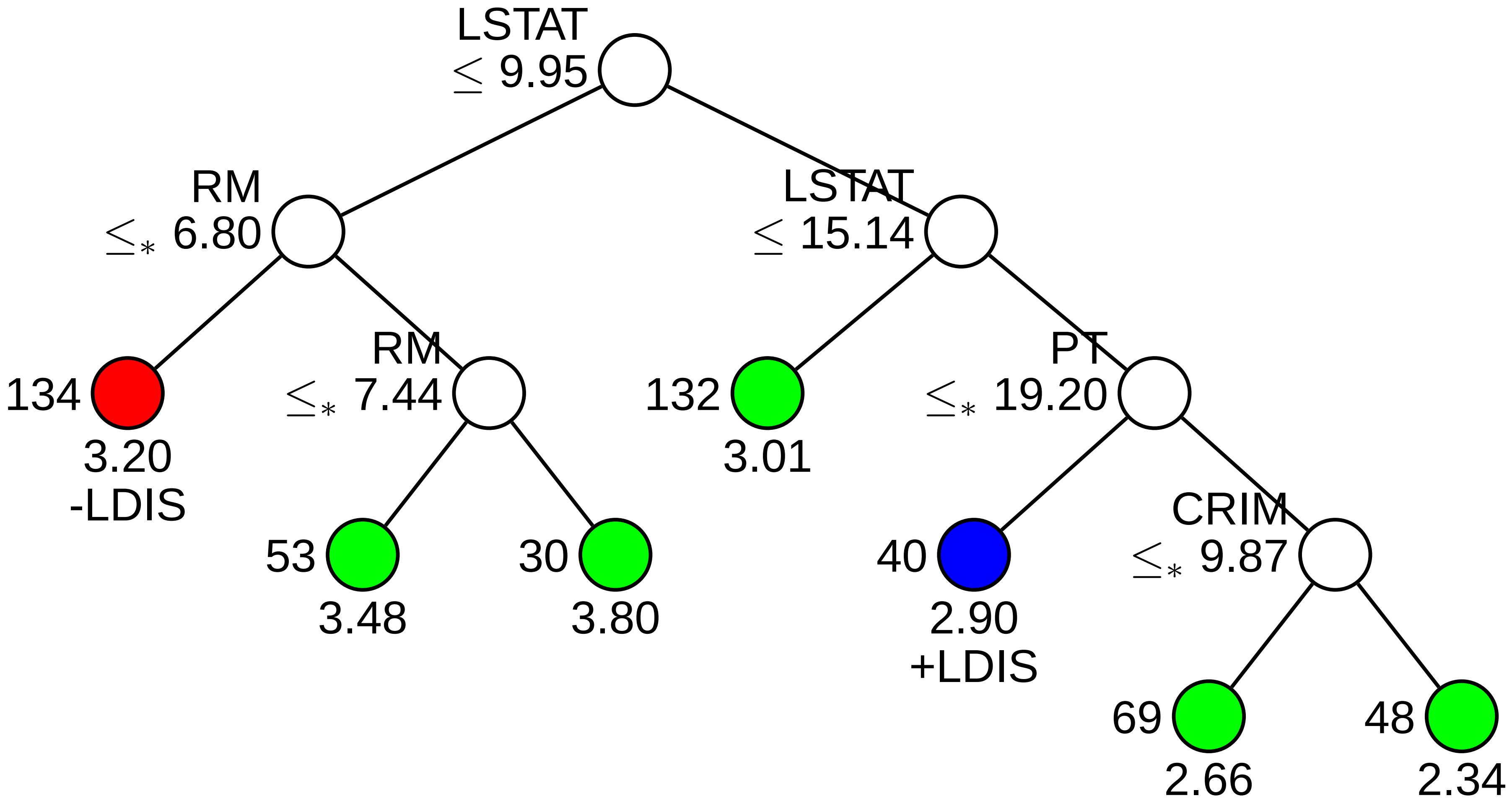
β = coefficient, t = t-statistic, ρ = corr(X, Y)

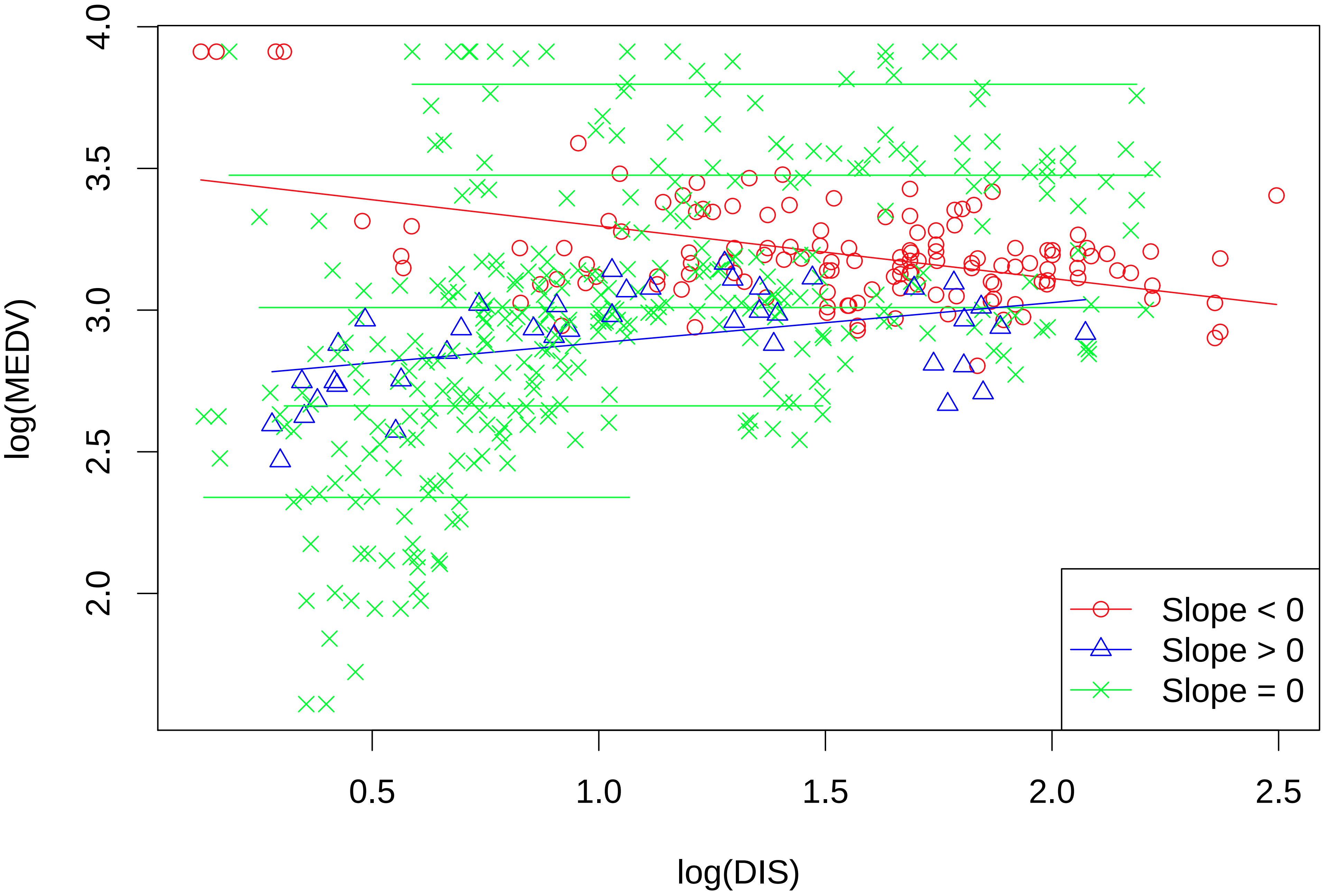
Why do β and ρ have opposite signs for $\log(\mathtt{DIS})$ and $\log(\mathtt{RAD})$?

log(MEDV) vs. log(DIS)

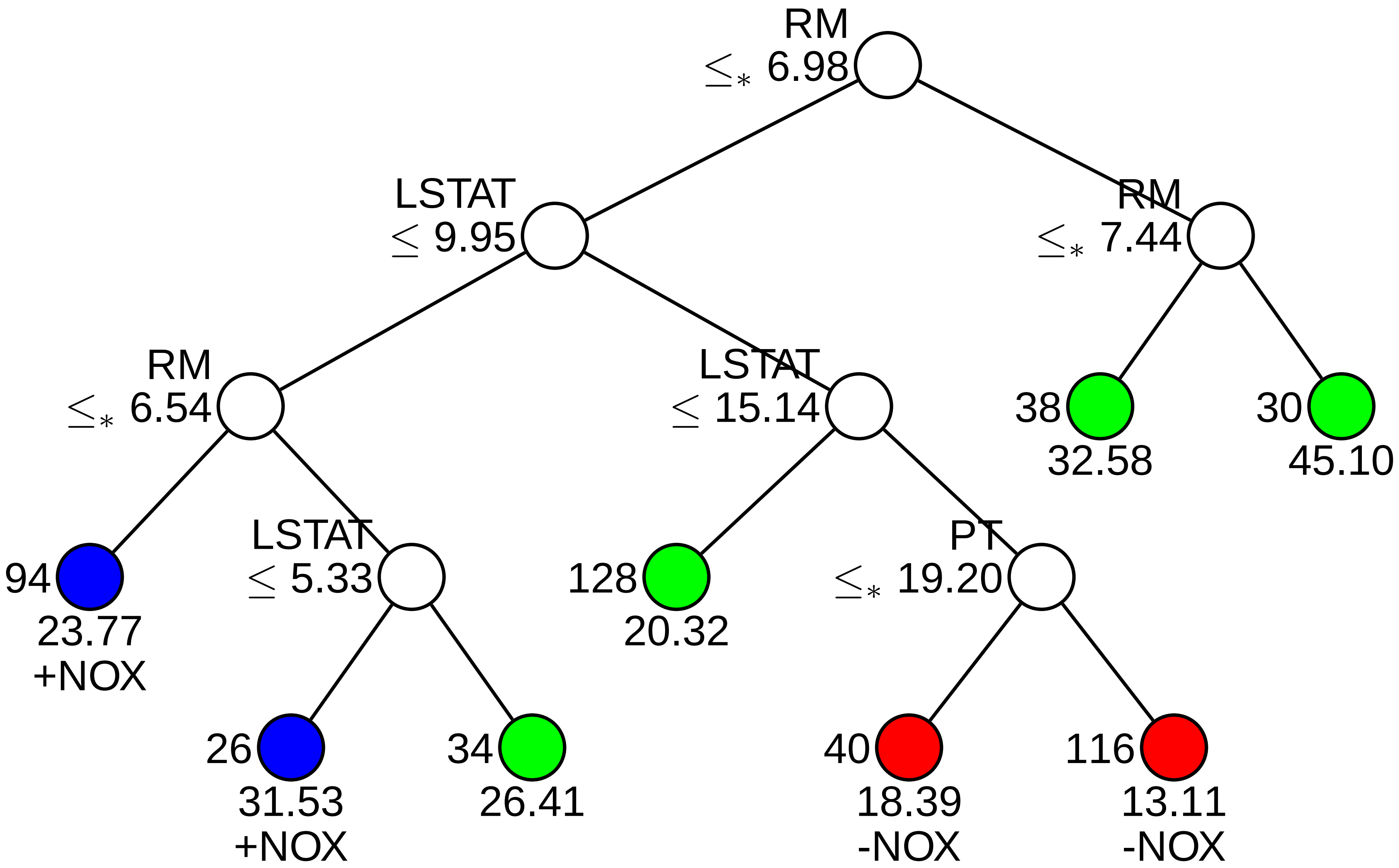


Model for log(MEDV) with log(DIS) as linear predictor

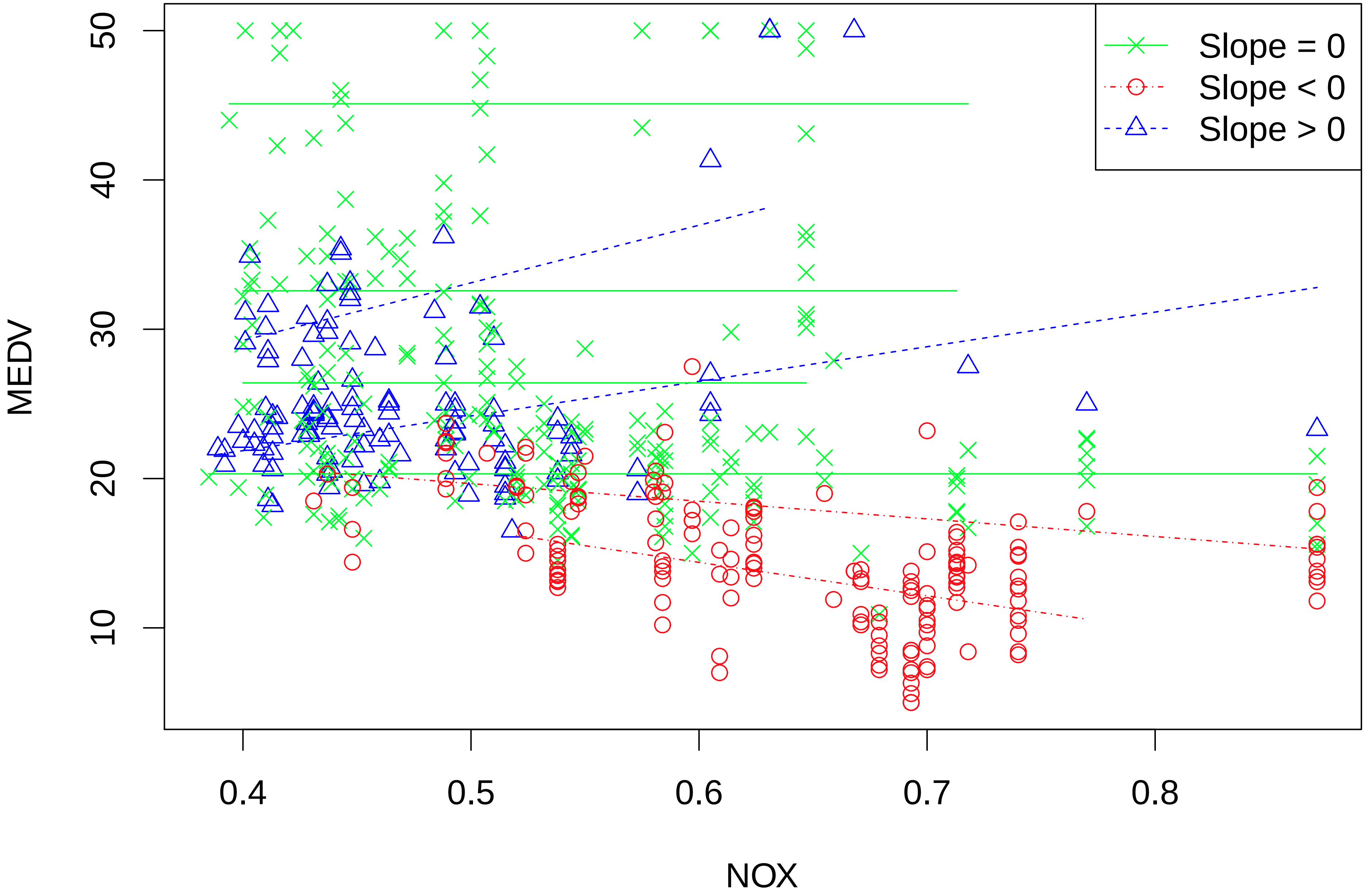



4.0
3.5
3.0
2.5
2.0
log(MEDV)
0.5
1.0
1.5
2.0
2.5
log(DIS)
Slope < 0
Slope > 0
Slope = 0

Model for MEDV with NOX as only linear predictor



MEDV vs NOX



Poisson regression:

Unreplicated $3 \times 2 \times 4 \times 10 \times 3$ soldering experiment

Opening: Amount of clearance around a mounting pad (small, medium, large)

Solder: Amount of solder (thin, thick)

Mask: Type and thickness of solder mask (A1.5, A3, B3, B6)

Pad: Shape and size of mounting pad (D4, D6, D7, L4, L6, L7, L8, L9, W4, W9)

Panel: Each board is divided into three panels (1, 2, 3)

Response: Number of solder skips (0–48)

Ref: Comizzoli et al. (1990), Chambers and Hastie (1992)

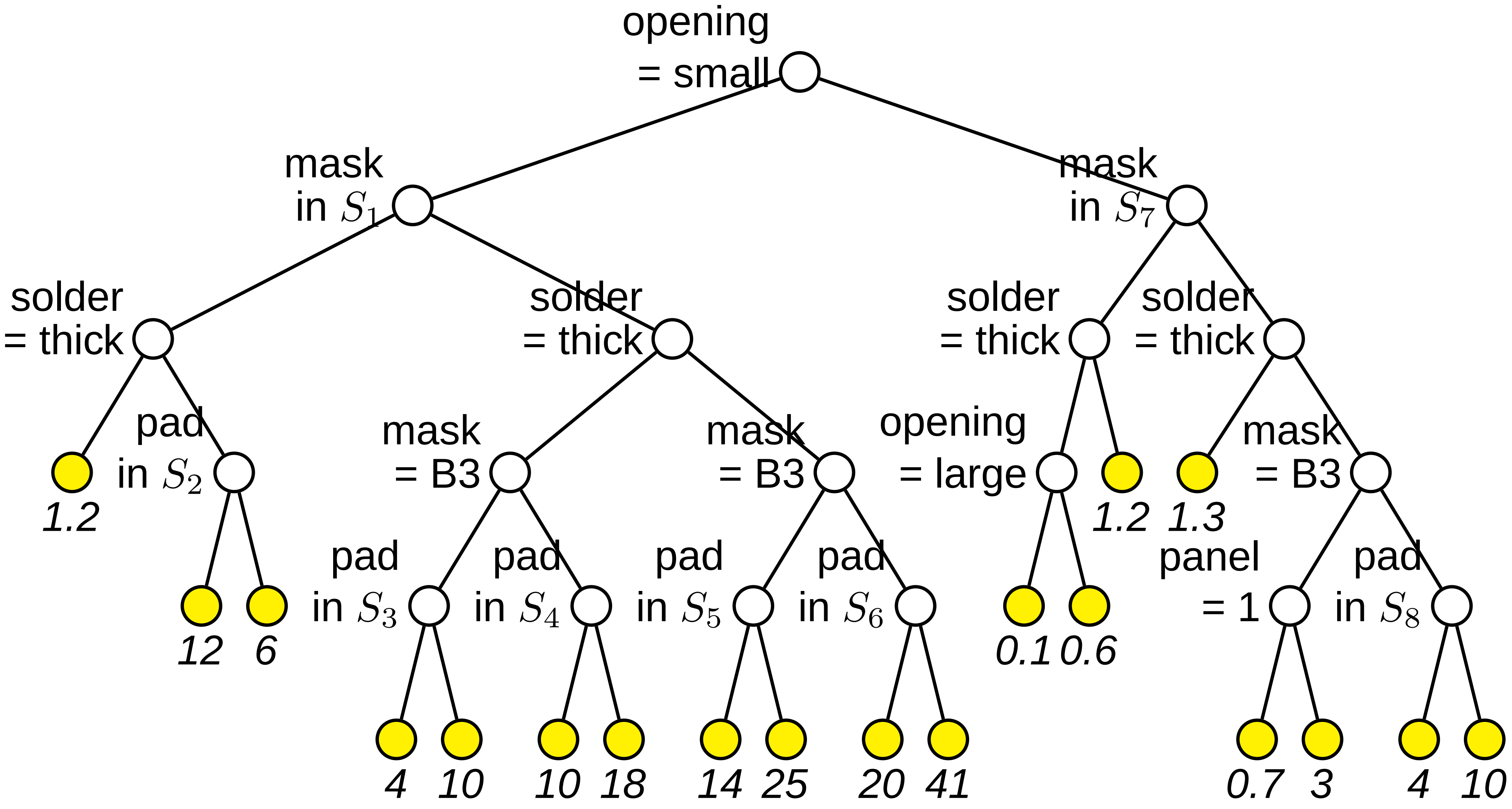
Full 2nd-degree Poisson loglinear model

Term	df	Deviance	P
open	2	2524.6	0.000
solder	1	937.0	0.000
mask	3	1653.1	0.000
pad	9	542.5	0.000
panel	2	68.1	0.000
open:solder	2	28.0	0.000
open:mask	6	71.0	0.000
solder:mask	3	59.8	0.000
open:pad	18	47.4	0.000
open:panel	4	11.2	0.024
solder:pad	9	43.4	0.000
solder:panel	2	6.0	0.050
mask:pad	27	61.5	0.000
mask:panel	6	21.2	0.002
pad:panel	18	13.7	0.748

Chambers & Hastie (1992) model with three 2-factor interactions

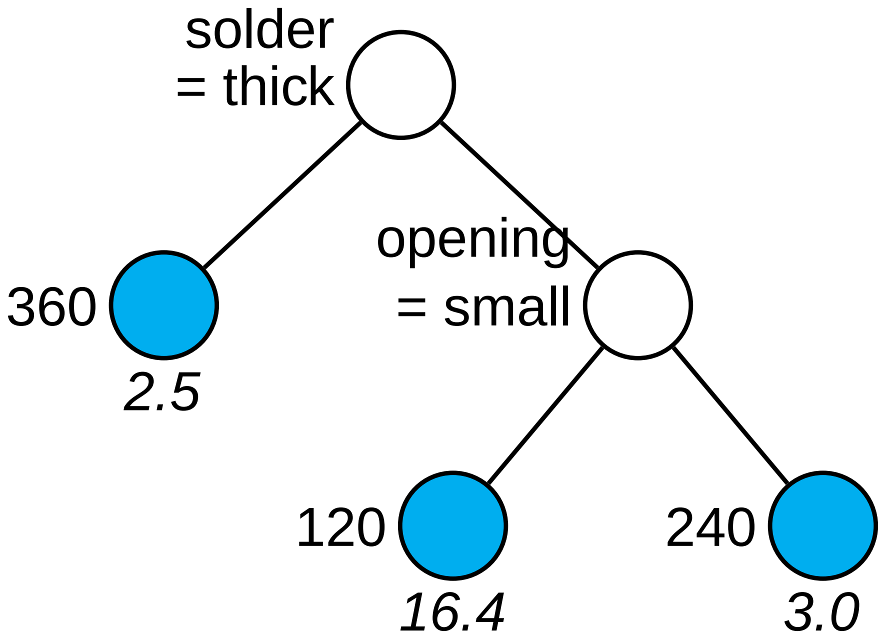
Regressor	Coef	t-stat	Regressor	Coef	t-stat
Constant	-2.668	-9.25			
maskA3	0.396	1.21	openmedium	0.921	2.95
maskB3	2.101	7.54	opensmall	2.919	11.63
maskB6	3.010	11.36	soldthin	2.495	11.44
padD6	-0.369	-5.17	maskA3:openmedium	0.816	2.44
padD7	-0.098	-1.49	maskB3:openmedium	-0.447	-1.44
padL4	0.262	4.32	maskB6:openmedium	-0.032	-0.11
padL6	-0.668	-8.53	maskA3:opensmall	-0.087	-0.32
padL7	-0.490	-6.62	maskB3:opensmall	-0.266	-1.12
padL8	-0.271	-3.91	maskB6:opensmall	-0.610	-2.74
padL9	-0.636	-8.20	maskA3:soldthin	-0.034	-0.16
padW4	-0.110	-1.66	maskB3:soldthin	-0.805	-4.42
padW9	-1.438	-13.80	maskB6:soldthin	-0.850	-4.85
panel2	0.334	7.93	openmedium:soldthin	-0.833	-4.80
panel3	0.254	5.95	opensmall:soldthin	-0.762	-5.13

GUIDE piecewise-constant Poisson model



Estimated mean number of solder skips given under each leaf node

GUIDE piecewise main effects Poisson model

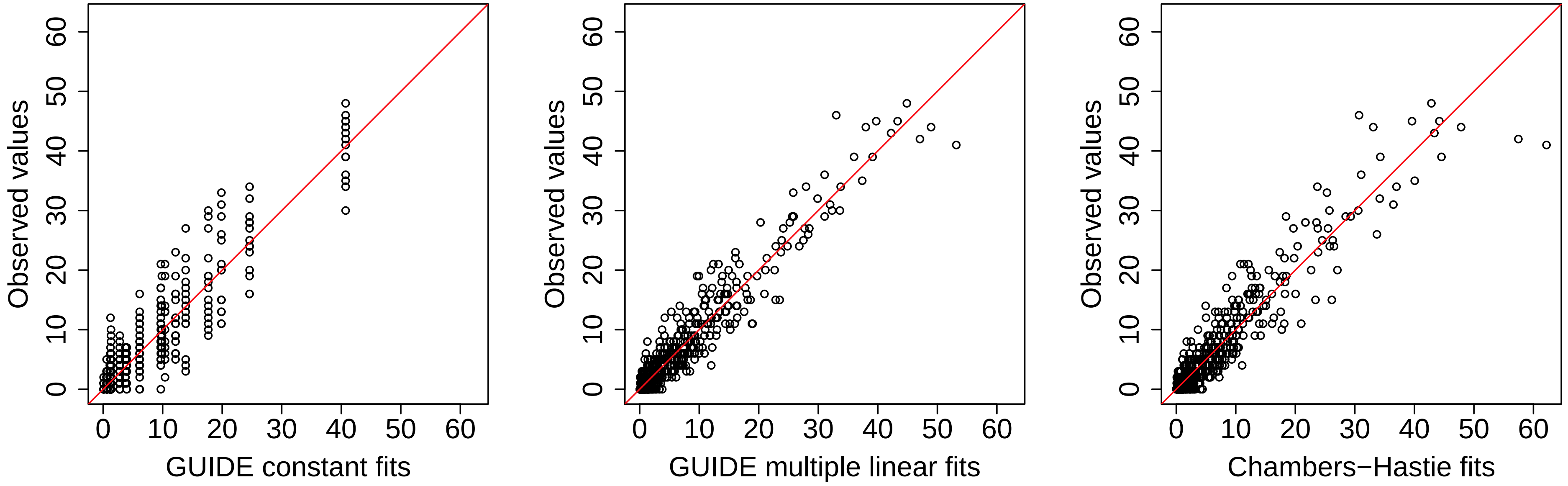


Number in italics below terminal node is sample mean of solder skips.

Number beside terminal node is sample size.

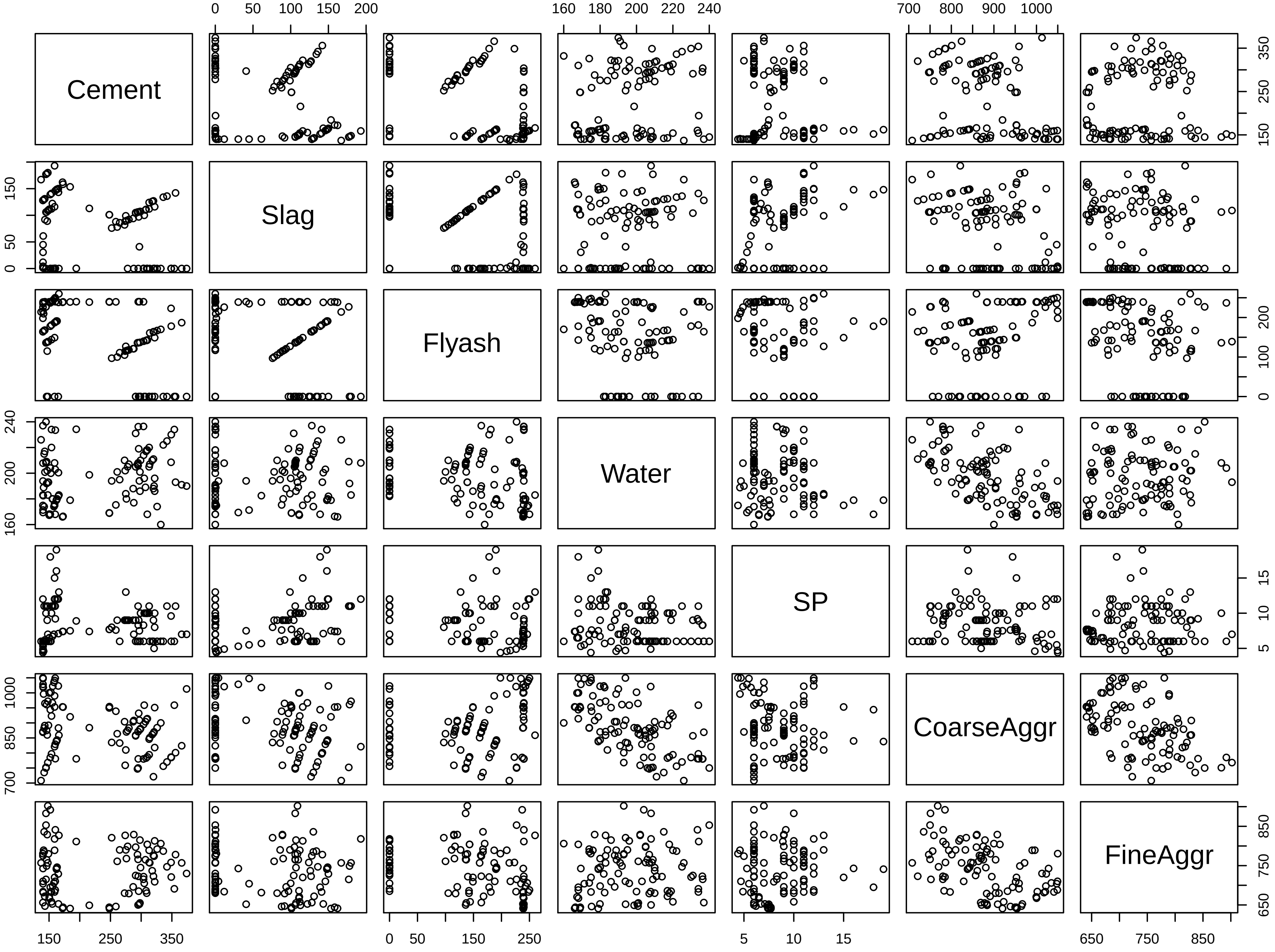
	solder = thick		solder = thin			
			opening = small		medium or large	
Regressor	Coef	t-stat	Coef	t-stat	Coef	t-stat
Constant	-2.43	-10.68	2.08	21.5	-0.37	-1.9
maskA3	0.47	2.37	0.31	3.3	0.81	4.5
maskB3	1.83	11.01	1.05	12.8	1.01	5.8
maskB6	2.52	15.71	1.50	19.3	2.27	14.6
openmedium	0.86	5.57	aliased		0.10	1.4
opensmall	2.46	18.18	aliased		aliased	
panel2	0.22	2.72	0.31	5.5	0.58	5.7
panel3	0.07	0.81	0.19	3.2	0.69	6.9
padD6	-0.32	-2.03	-0.25	-2.8	-0.80	-4.6
padD7	0.12	0.85	-0.15	-1.7	-0.19	-1.3
padL4	0.70	5.53	0.08	1.0	0.21	1.6
padL6	-0.40	-2.46	-0.72	-6.8	-0.82	-4.7
padL7	0.04	0.29	-0.65	-6.3	-0.76	-4.5
padL8	0.15	1.05	-0.43	-4.5	-0.36	-2.4
padL9	-0.59	-3.43	-0.64	-6.3	-0.67	-4.1
padW4	-0.05	-0.37	-0.09	-1.0	-0.23	-1.6
padW9	-1.32	-5.89	-1.38	-10.3	-1.75	-7.0

Observed vs. fitted values



Multiresponse data:
viscosity and strength of concrete (Yeh, 2007)

- 103 observations on seven input variables (kg per cubic meter):
  1. Cement
  2. Slag
  3. Fly ash
  4. Water
  5. Superplasticizer
  6. Coarse aggregate
  7. Fine aggregate
- Three output variables:
  1. Slump (cm)
  2. Flow (cm)
  3. 28-day compressive strength (Mpa)


Cement
Slag
Flyash
Water
SP
CoarseAggr
FineAggr
0 50 100 150 200
160 180 200 220 240
700 800 900 1000
150 250 350
0 50 100 150 200 250
160 200 240
5 10 15
700 850 1000
650 750 850
150 250 350
0 50 150 250
5 10 15
650 750 850

Separate linear models

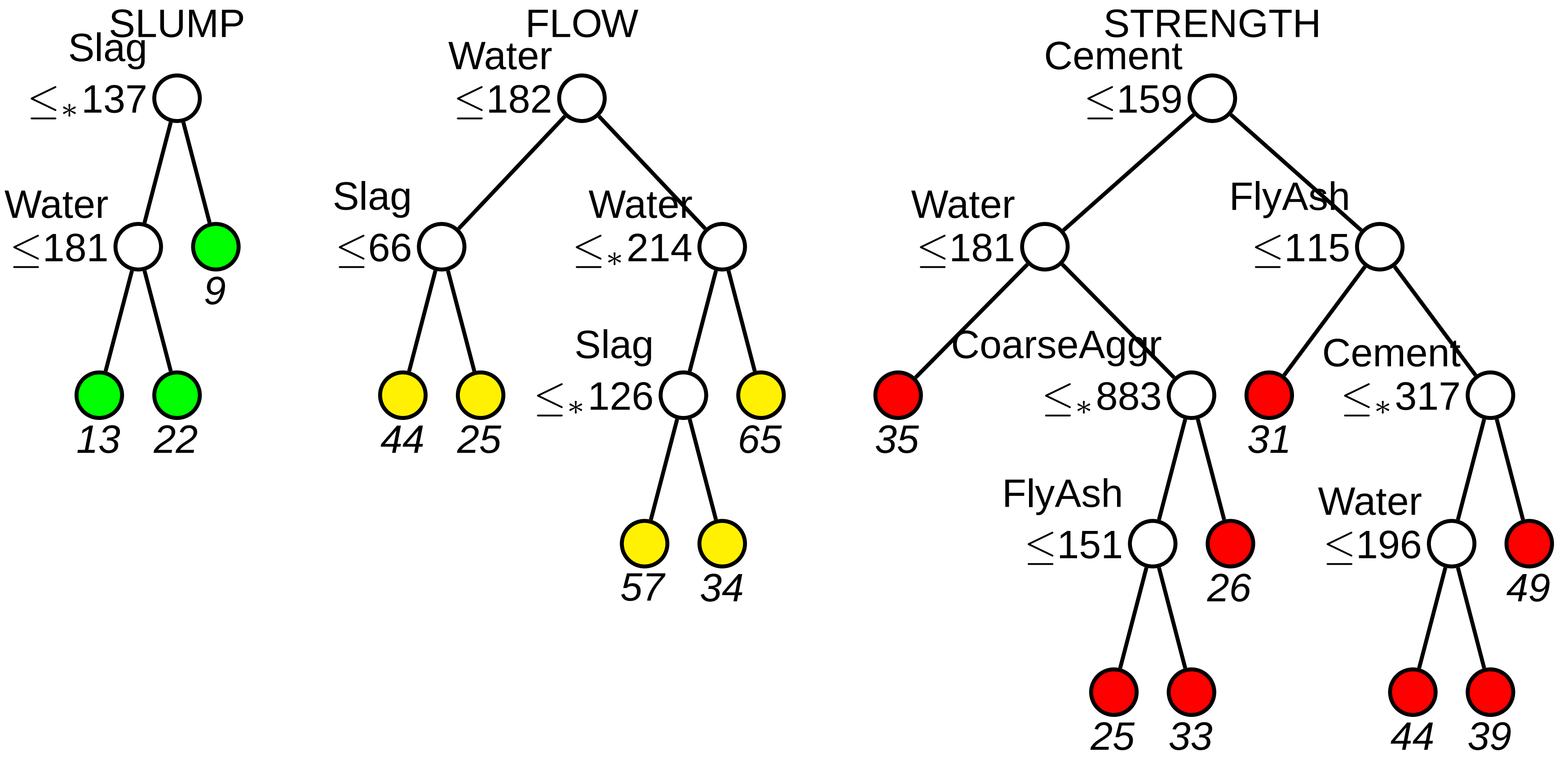
	Slump		Flow		Strength	
	Estimate	P-value	Estimate	P-value	Estimate	P-value
(Intercept)	-88.525	0.66	-252.875	0.472	139.782	0.052
Cement	0.010	0.88	0.054	0.634	0.061	0.008
Slag	-0.013	0.89	-0.006	0.971	-0.030	0.352
Flyash	0.006	0.93	0.061	0.593	0.051	0.032
Water	0.259	0.21	0.732	0.041	-0.23270	0.002
SP	-0.184	0.63	0.298	0.654	0.103	0.445
CoarseAggr	0.030	0.71	0.074	0.587	-0.056	0.045
FineAggr	0.039	0.64	0.094	0.509	-0.039	0.178

Is there really nothing significant for Slump?

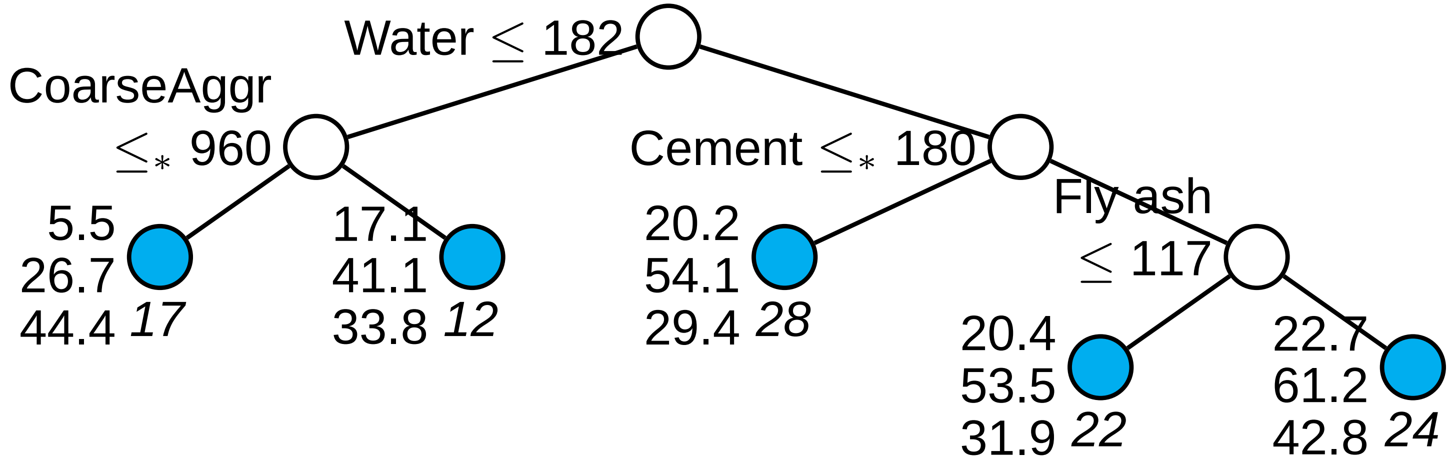
Water and Slag are highly significant for Slump if no other variables are in the model!

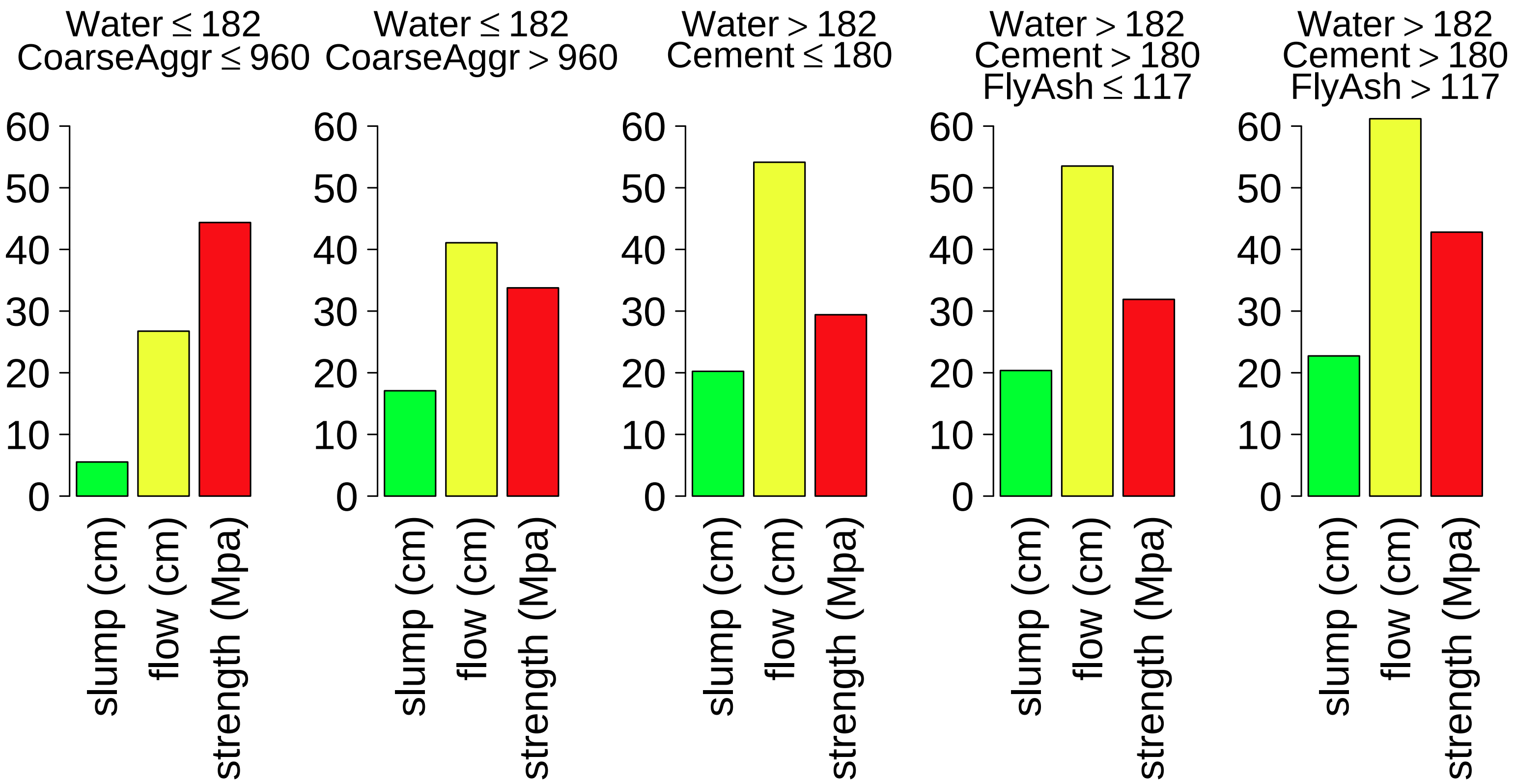
	Estimate	Std. Error	t value	Pr($>|t|$)
(Intercept)	-18.099	7.314	-2.475	0.01502 *
Water	0.199	0.036	5.455	3.56e-07 ***
Slag	-0.039	0.012	-3.227	0.00169 **
(Intercept)	11.370	9.683	1.174	0.243
Water	0.050	0.0486	1.025	0.308
Slag	-0.479	0.104	-4.604	1.23e-05 ***
Water:Slag	0.002	0.001	4.251	4.83e-05 ***

One tree for each response variable



One tree for all response variables





College tuition and graduation rate

- Data on 1134 U.S. colleges and universities for year 1995 from *U. S. News & World Report* (`http://lib.stat.cmu.edu/`)
- Response variables are out-of-state tuition and graduation rate
- 515 complete cases

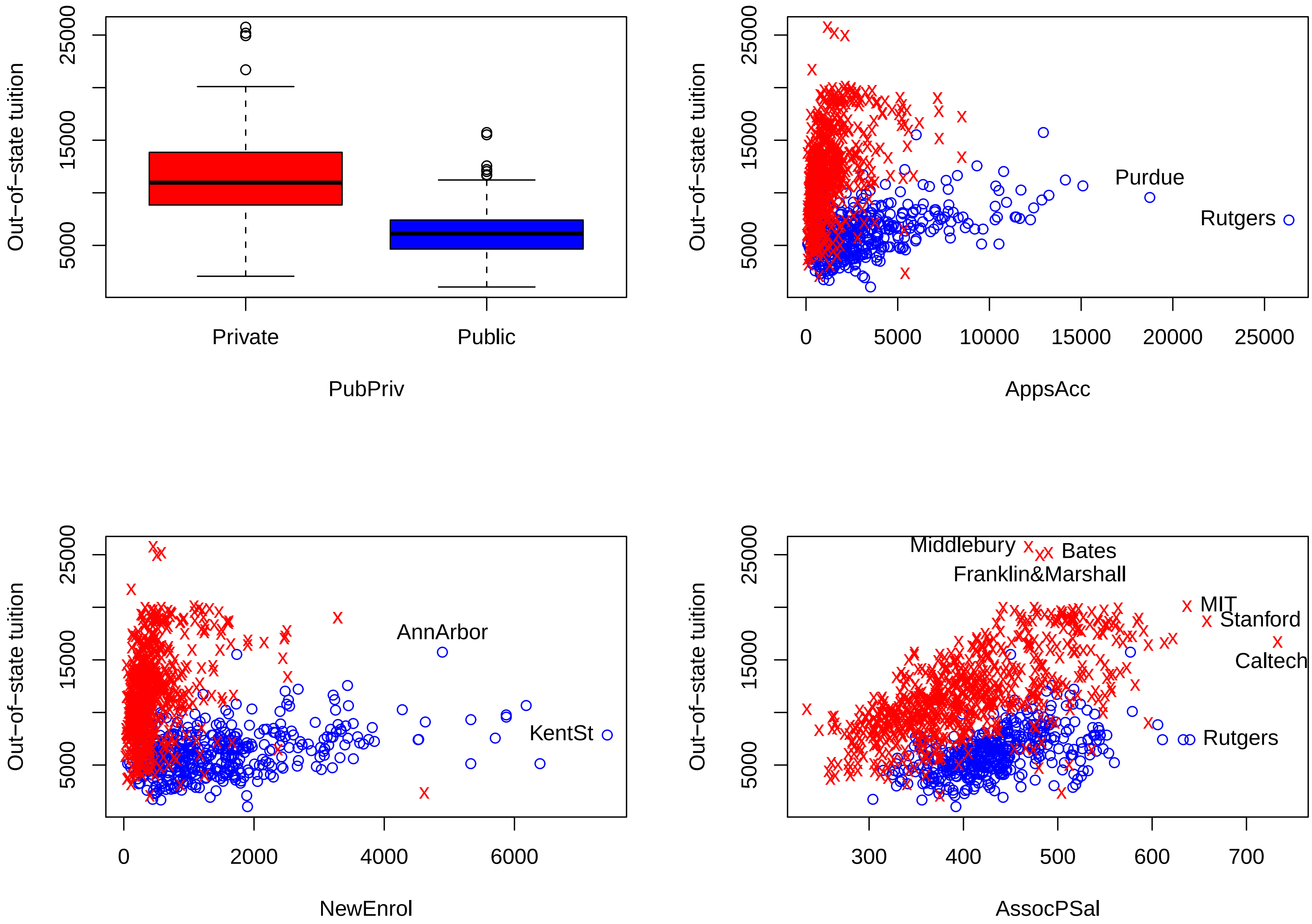
Explanatory variables for college data

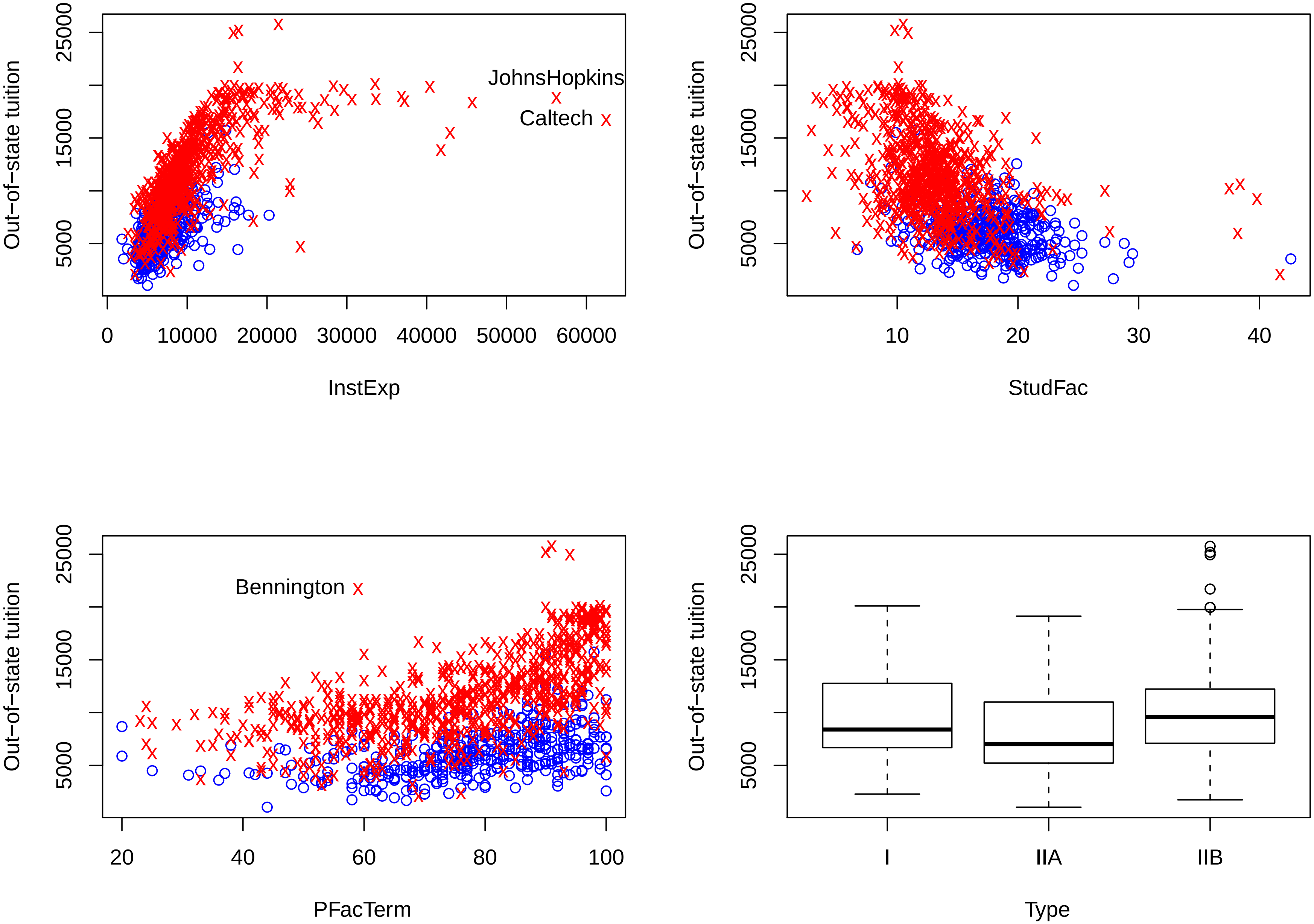
Name	Description	#Missing
PubPriv	Public or private college (binary)	0
CombSAT	Average Combined SAT score	471
AppsRec	Number of applications received	9
AppsAcc	Number of applicants accepted	9
NewEnrol	Number of new students enrolled	5
Top10	Percent new students from top 10% of H.S. class	183
Top25	Percent new students from top 25% of H.S. class	155
FUgrad	Number of fulltime undergraduates	3

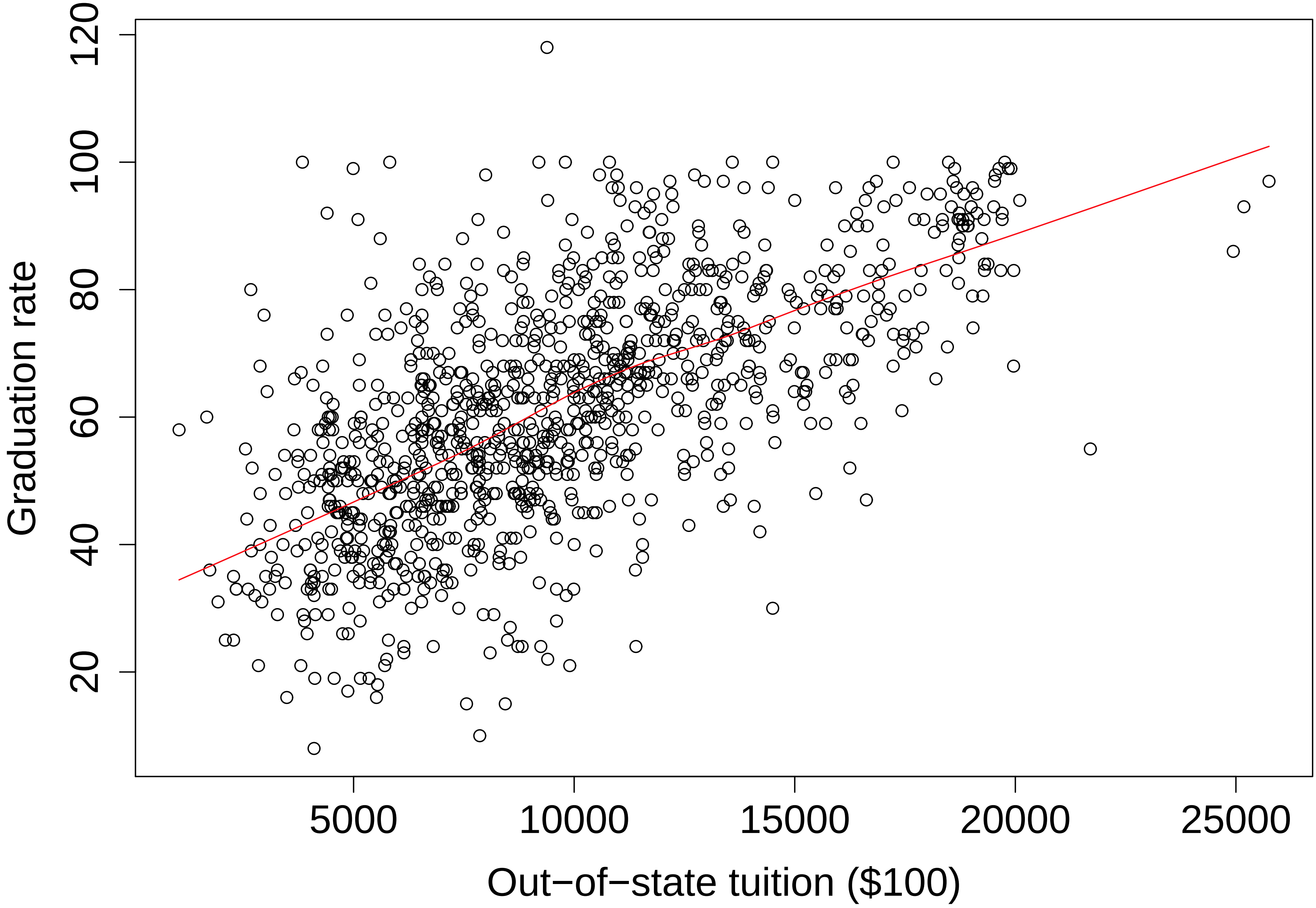
Explanatory variables for college data (cont’d)

Name	Description	#Missing
RnBcost	Room and board costs	57
PFacPhD	Percent of faculty with Ph.D.’s	29
StudFac	Student/faculty ratio	2
InstExp	Instructional expenditure per student	24
GradRate	Graduation rate	69
Type	Type of college (I: PhD, IIA: master, or IIB: bachelor)	0
FullPSal	Average salary—full professors (in $100’s)	61
NFullProf	Number of full professors	0

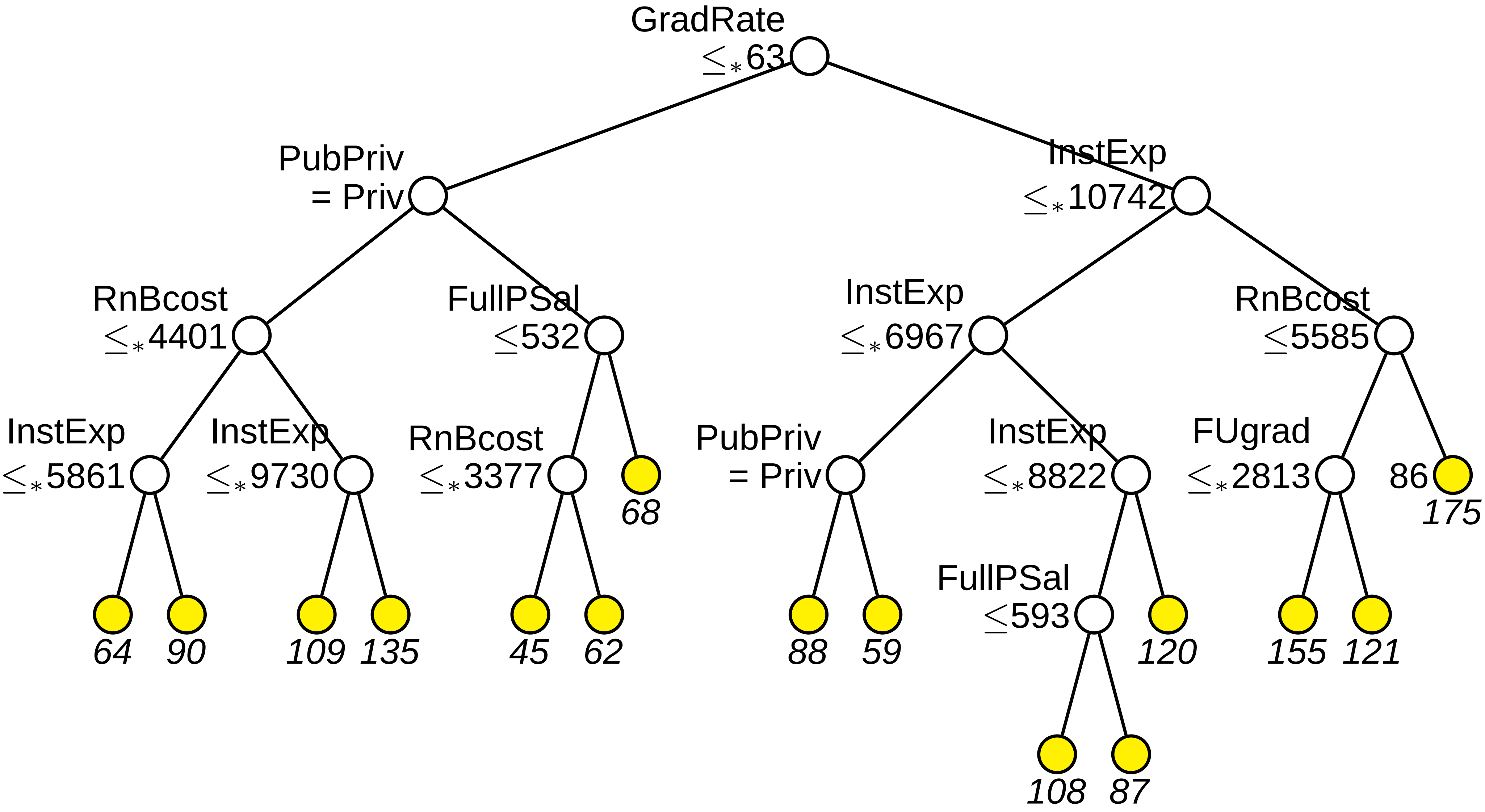
513 cases with complete observations


Out−of−state tuition
5000
15000
25000
Private
Public
PubPriv
0
5000
10000
15000
20000
25000
AppsAcc
Purdue
Rutgers
0
2000
4000
6000
NewEnrol
AnnArbor
KentSt
300
400
500
600
700
AssocPSal
Middlebury
Bates
Franklin&Marshall
MIT
Stanford
Caltech
Rutgers

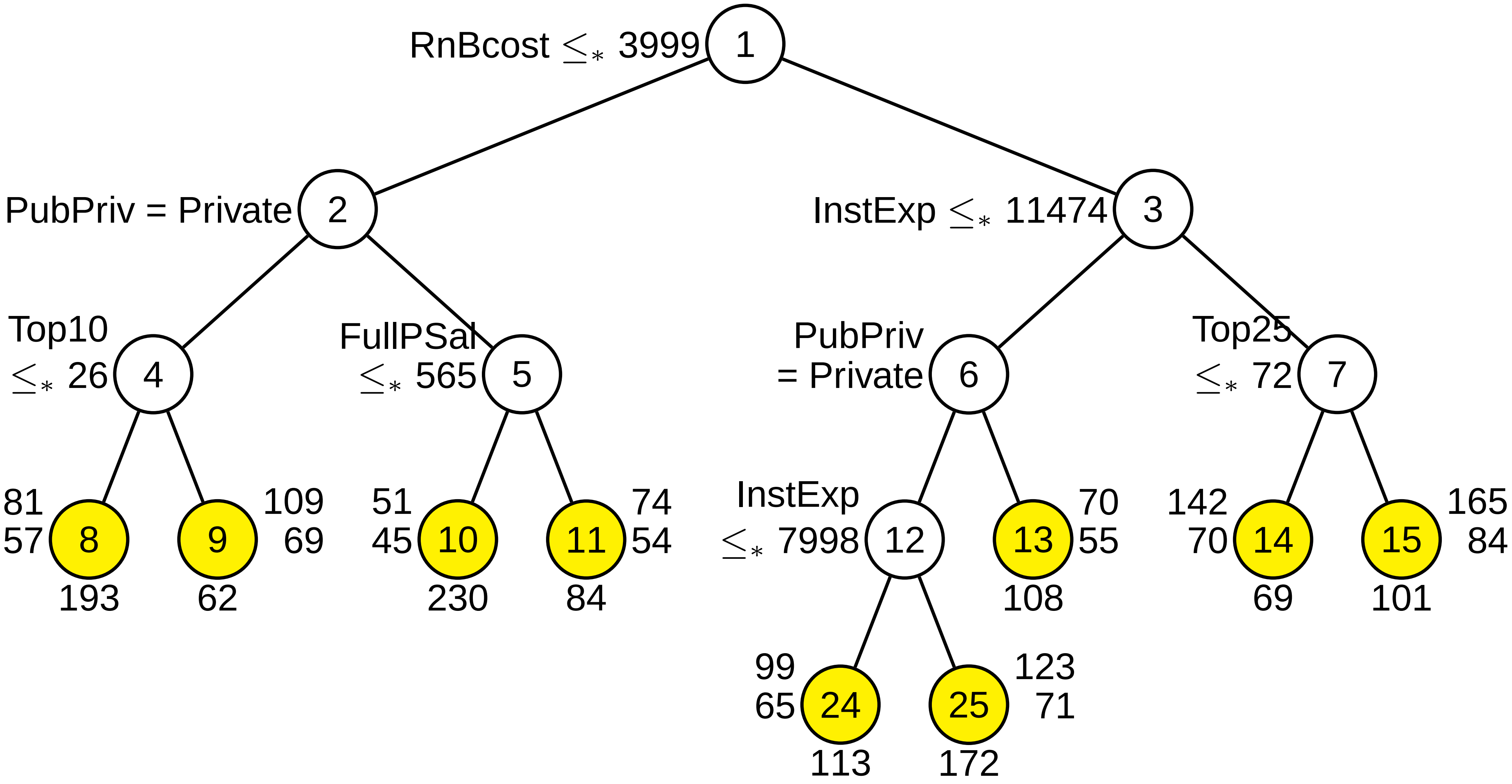

Out−of−state tuition
5000
15000
25000
0
10000
20000
30000
40000
50000
60000
InstExp
JohnsHopkins
Caltech
Out−of−state tuition
5000
15000
25000
10
20
30
40
StudFac
Out−of−state tuition
5000
15000
25000
20
40
60
80
100
PFacTerm
Bennington
Out−of−state tuition
5000
15000
25000
I
IIA
IIB
Type


Graduation rate
20
40
60
80
100
120
5000
10000
15000
20000
25000
Out−of−state tuition ($100)

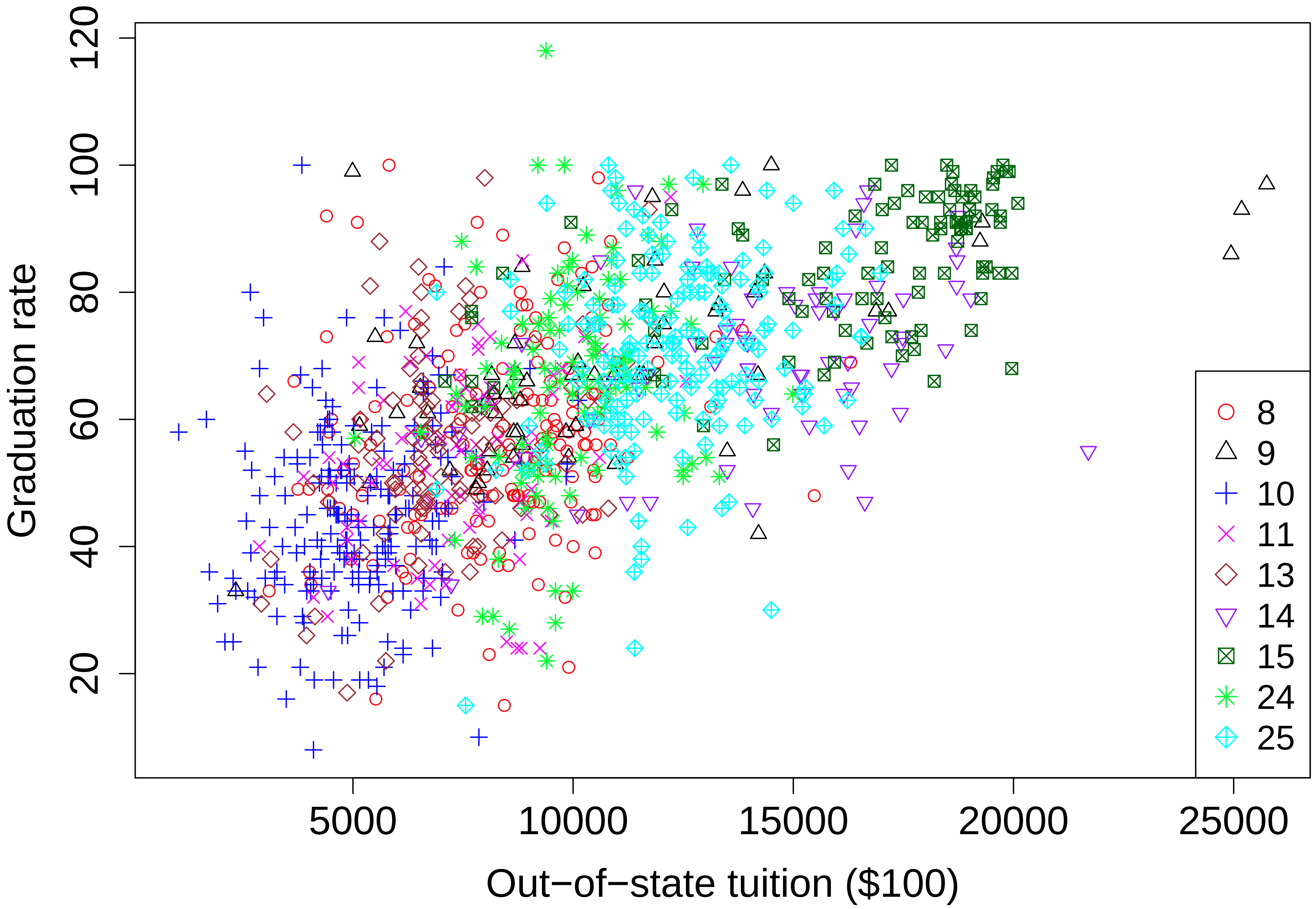
Out-of-state tuition (in $100s)



Out-of-state tuition and graduation rate



Predicted values of OutTuition, GradRate, resp., beside terminal nodes, sample sizes below


Graduation rate
120
100
80
60
40
20
5000
10000
15000
20000
25000
Out−of−state tuition ($100)
○ 8
△ 9
+ 10
× 11
◇ 13
▽ 14
⊠ 15
✳ 24
⊕ 25

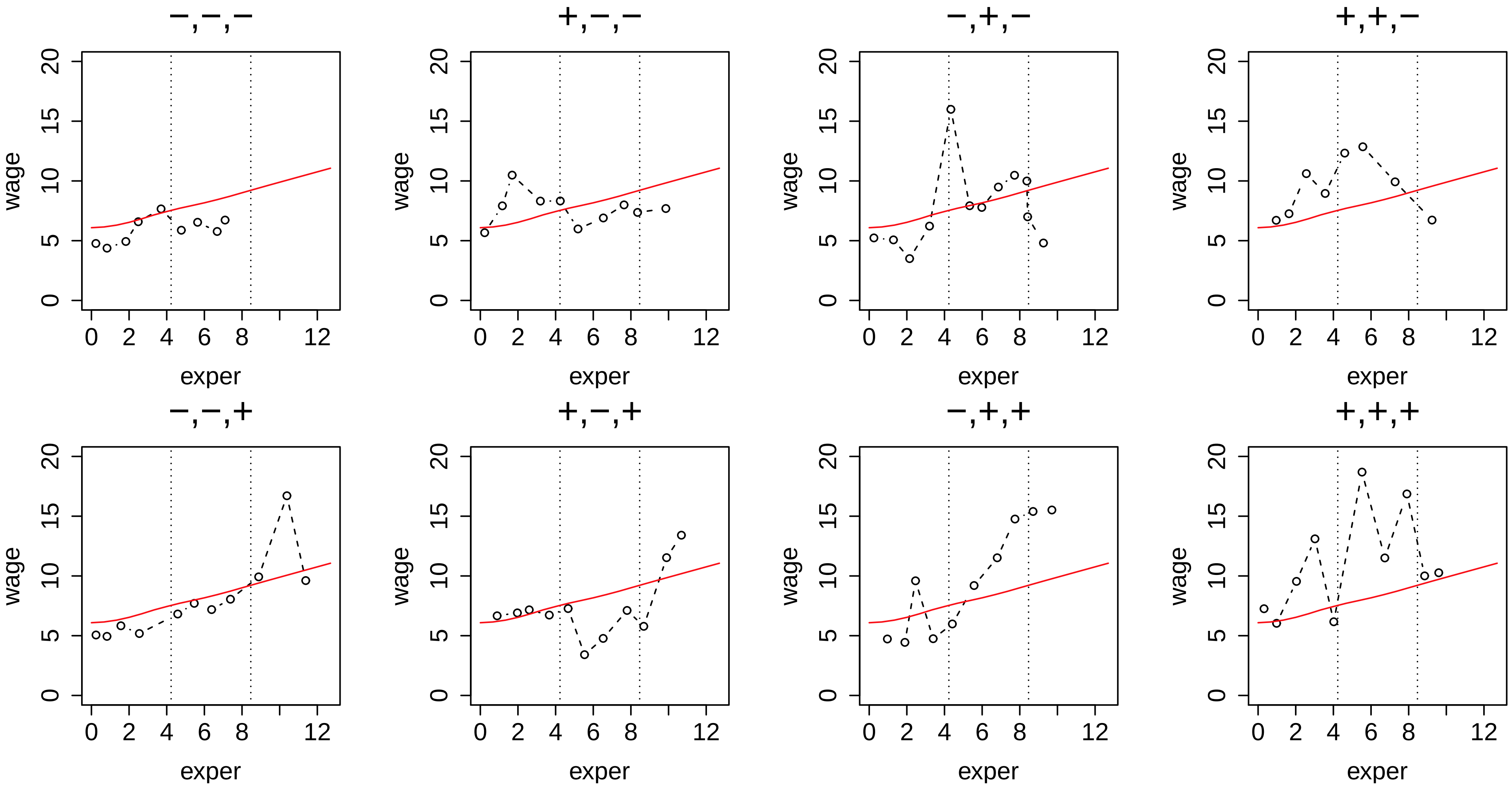
Longitudinal data: Hourly wage of high-school dropouts

- 888 male high-school dropouts (246 Black, 204 Hispanic, 438 White) observed over time
- Response is hourly wage (in 1990 dollars)
- Predictor variables are:
  1. `hgc`: highest grade completed (6–12)
  2. `exper`: years in labor force (0.001–12.7 yrs)
  3. `black`: 1 if Black, 0 otherwise
  4. `hisp`: 1 if Hispanic, 0 otherwise
- Data from the National Longitudinal Survey of Youth
- References: Murnane et al. (1999), Singer and Willett (2003, Sec. 5.2.1)

Design details and complications

1. At first wave of data collection, subjects varied in age from 14–17
2. Some subsequent waves separated by one year, others by two
3. Each wave's interviews conducted at different times in calendar year
4. Subjects observed at random times and random number of times:
   77 have 1–2, 82 have 3–4, 166 have 5–6, 226 have 7–8, 240 have 9–10, and 97 have more than 10 observations
5. Subjects could describe more than one job at each interview
6. Subjects drop out of school and enter labor force at varying times
7. Subjects can change jobs at any time
8. Murnane et al. (1999) clocked time from each subject's first day of work

Some individual trajectories



Questions in analysis of longitudinal data

1. How does the outcome change over time?
2. Can we predict the differences in these changes?

Two popular approaches

Parametric: Fit a *mixed model* (also called *individual growth model, random coefficient model, multilevel model*, and *hierarchical linear model*) and deduce the effect of predictor variables from the regression coefficients

Nonparametric: *Cluster* the subject trajectories, then *test* each predictor variable for its effect on the clusters

Linear mixed model (Singer and Willett, 2003)

$$\begin{aligned}\log(\texttt{wage}) &= \beta_0 + \beta_1\texttt{hgc} + \beta_2\texttt{exper} + \beta_3\texttt{black} + \beta_4\texttt{hisp} \\ &\quad + \beta_5\texttt{exper} \times \texttt{black} + \beta_6\texttt{exper} \times \texttt{hisp} \\ &\quad + b_0 + b_1\texttt{exper} + \epsilon\end{aligned}$$

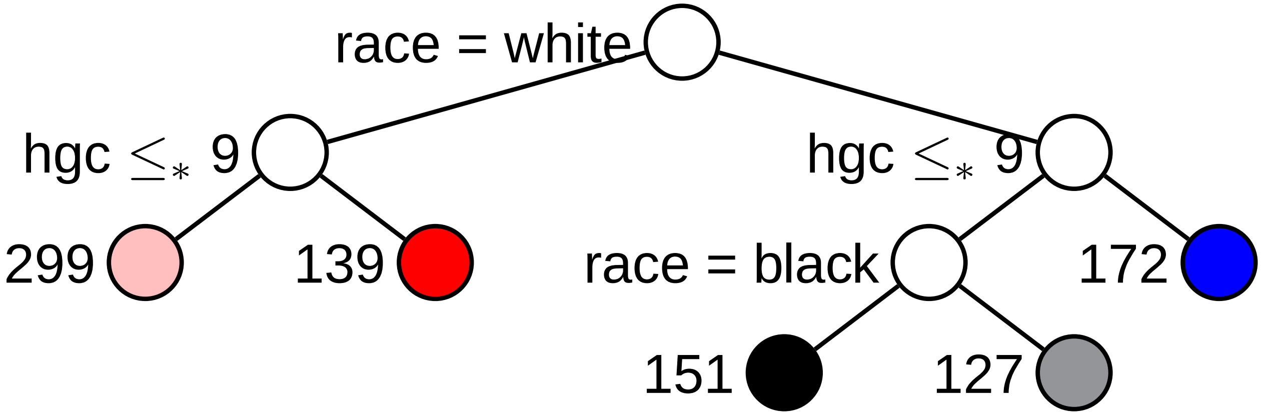
Assumptions/limitations:

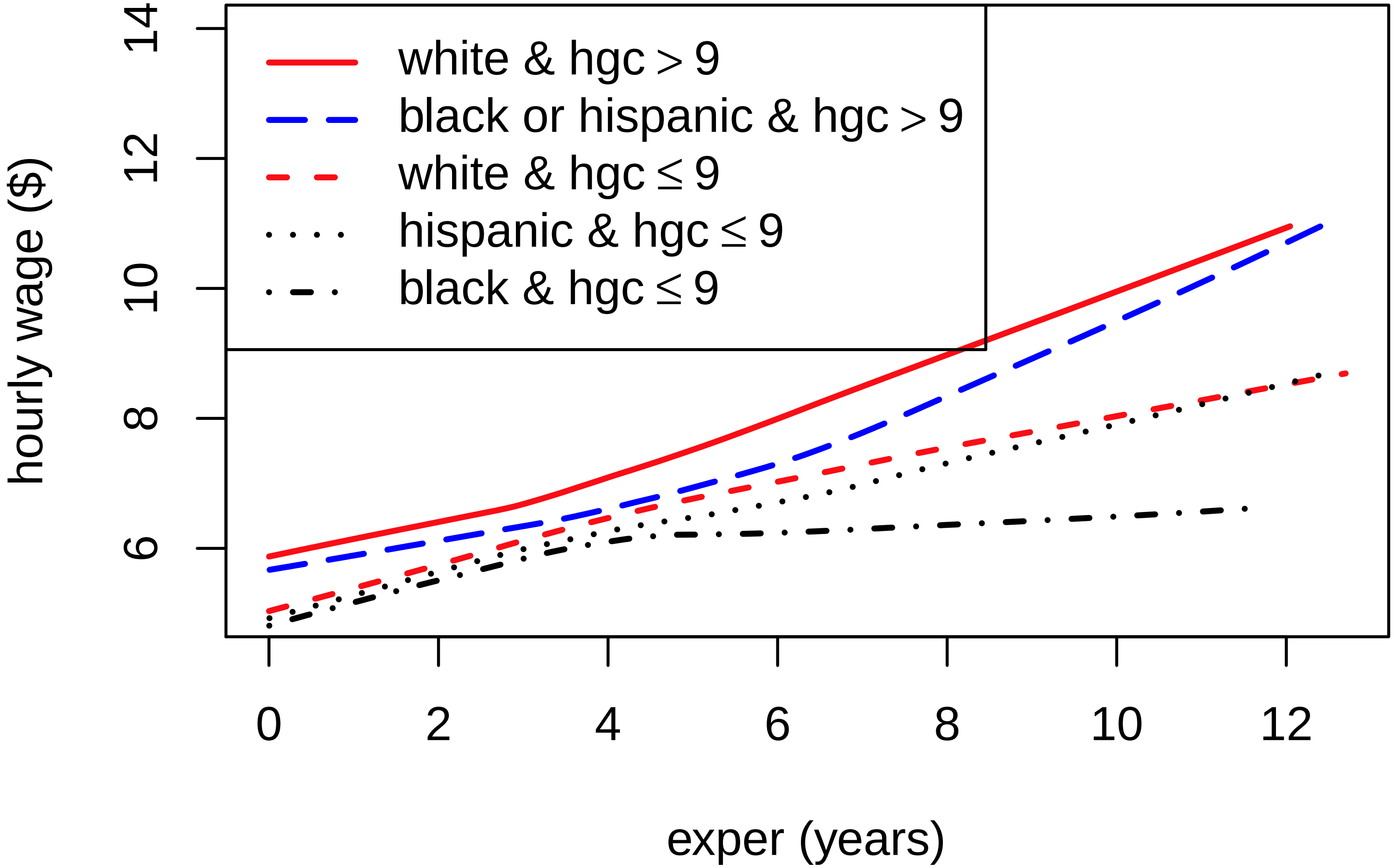
1. Random (subject) intercepts and slopes $b_0 \sim N(0, \sigma_0^2)$ and $b_1 \sim N(0, \sigma_1^2)$; $\epsilon \sim N(0, \sigma^2)$; all independent
2. Log transformation of `wage` to address skewness, linearize individual wage trajectories, and overcome range restriction
3. Predictions of `wage` requires exponentiation of fitted values of $\log(\texttt{wage})$ — least-squares fit on log-dollar scale not best for dollar scale

Coefficients of fixed effect terms

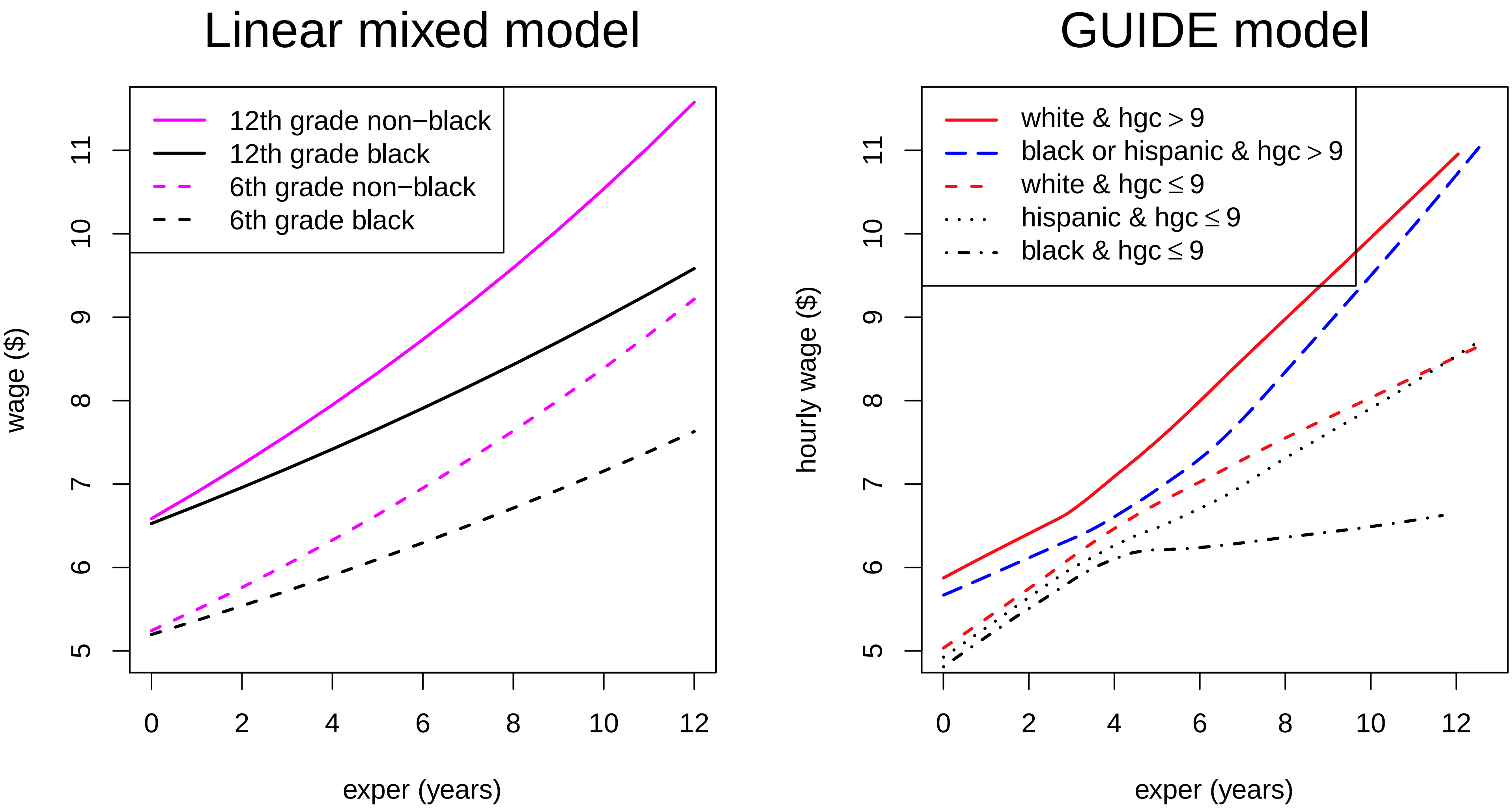
	Value	Std.Error	DF	t-value	p-value
(Intercept)	1.382	0.059	5511	23.43	0.000
hgc	0.038	0.006	884	5.94	0.000
exper	0.047	0.003	5511	14.57	0.000
black	0.006	0.025	884	0.25	0.804
hisp	-0.028	0.027	884	-1.03	0.302
exper×black	-0.015	0.006	5511	-2.65	0.008
exper×hisp	0.009	0.006	5511	1.51	0.131

“Analyses not shown here suggest that we cannot distinguish statistically between the trajectories of Hispanic and White dropouts.” (Singer and Willett, 2003, p. 149)


race = white
hgc $\leq_*$ 9
hgc $\leq_*$ 9
299
139
race = black
172
151
127

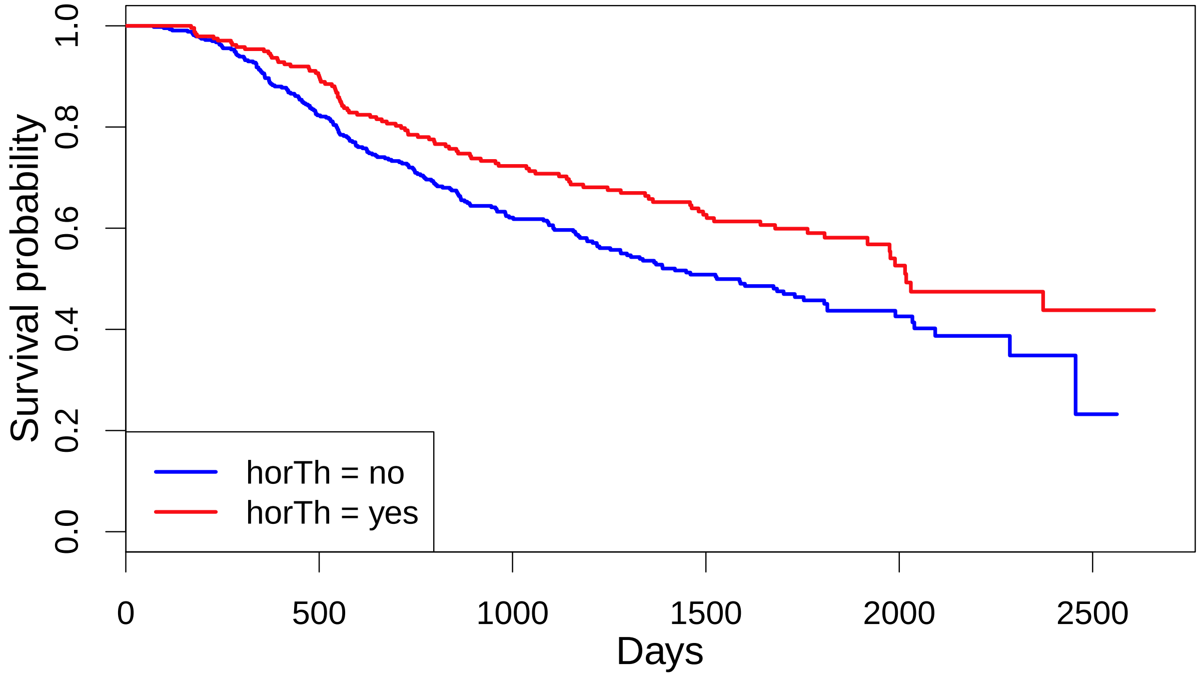

white & hgc > 9
black or hispanic & hgc > 9
white & hgc ≤ 9
hispanic & hgc ≤ 9
black & hgc ≤ 9
hourly wage ($)
6
8
10
12
14
0
2
4
6
8
10
12
exper (years)

LME vs. GUIDE fits



Censored response data: breast cancer

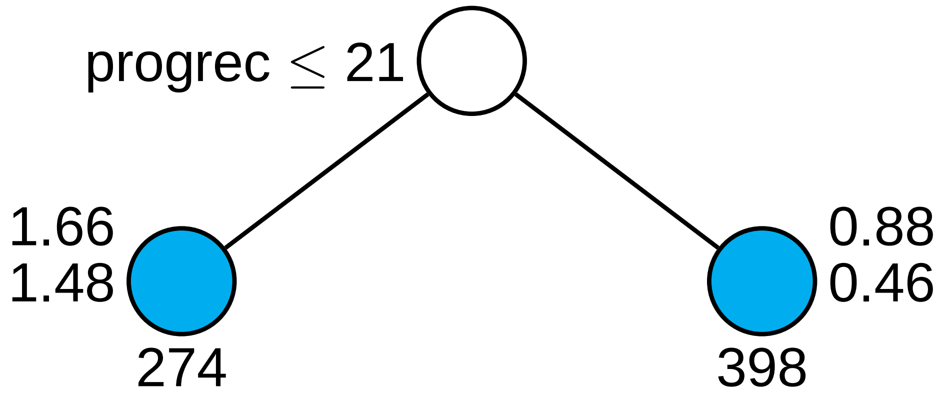
- Randomized clinical trial of 672 subjects with primary node positive breast cancer (Schumacher et al., 1994; data from **ipred** R package; 14 subjects with censored times less than smallest uncensored time excluded)
- Response is recurrence-free survival time (8–2659 days, 299 uncensored, 387 censored)
- 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)

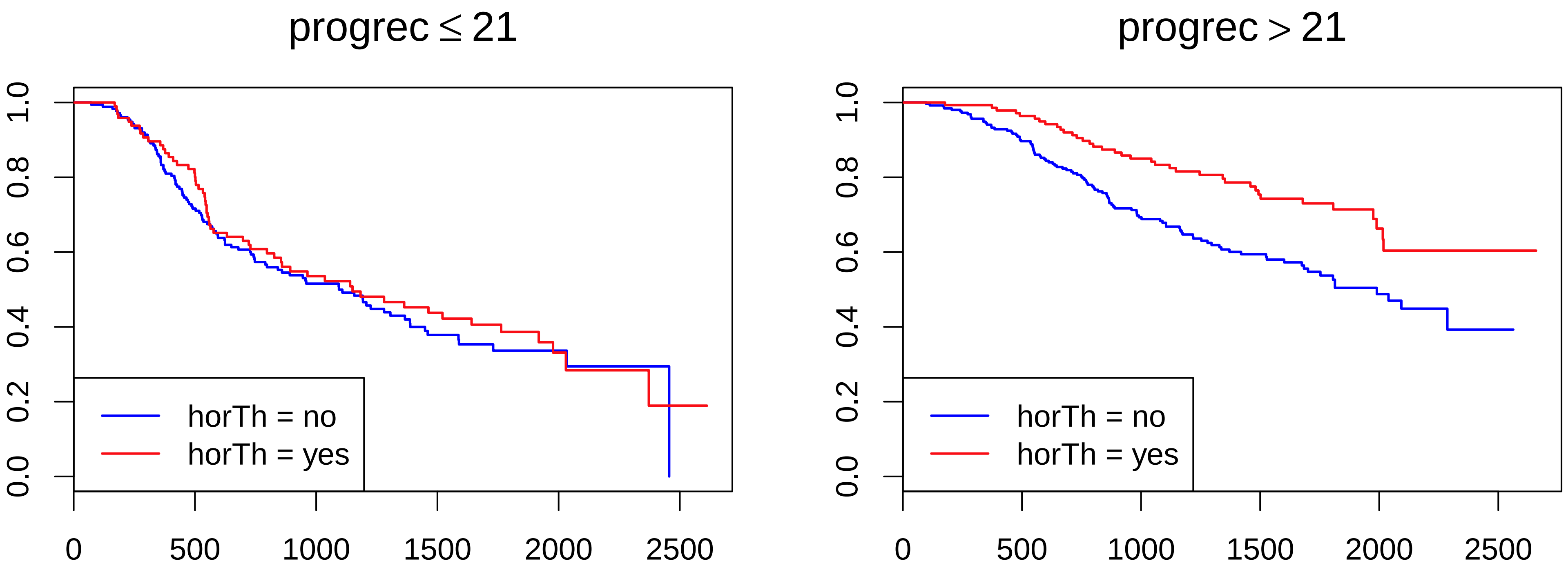


Variable	Coef	p-value	Variable	Coef	p-value
horTh=yes	-0.3463	7.3e-03	tsize	0.0078	4.8e-02
age	-0.0095	3.1e-01	pnodes	0.0488	5.7e-11
meno=Post	0.2585	1.6e-01	progrec	-0.0022	1.1e-04
tgrade.L	0.5513	3.7e-03	estrec	0.0002	6.6e-01
tgrade.Q	-0.2011	9.9e-02			

Is there a subgroup where hormone therapy is ineffective?

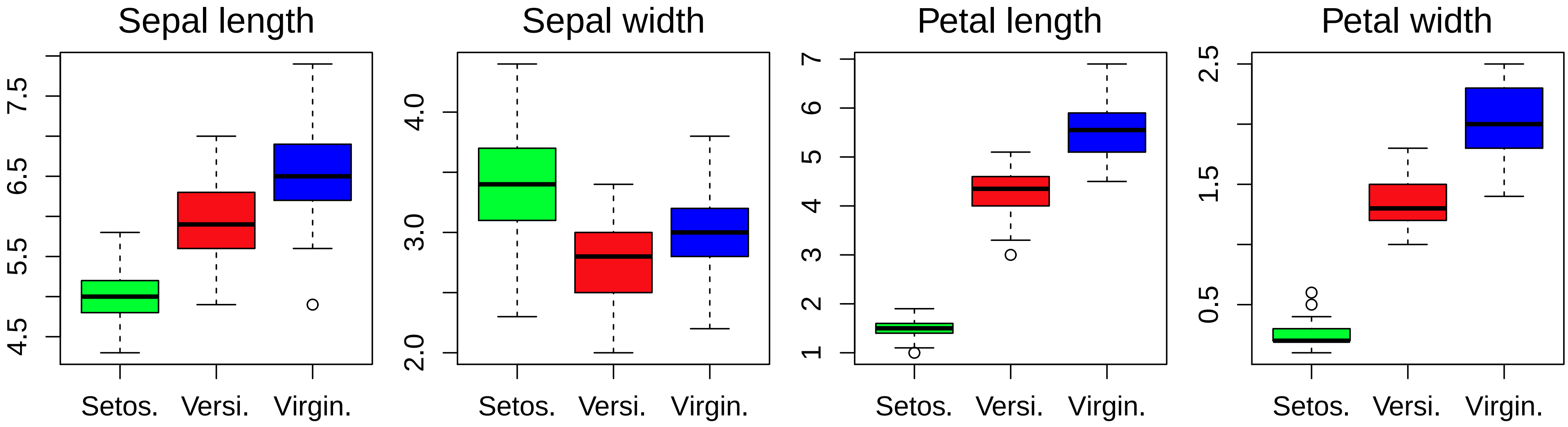
GUIDE tree for subgroup identification



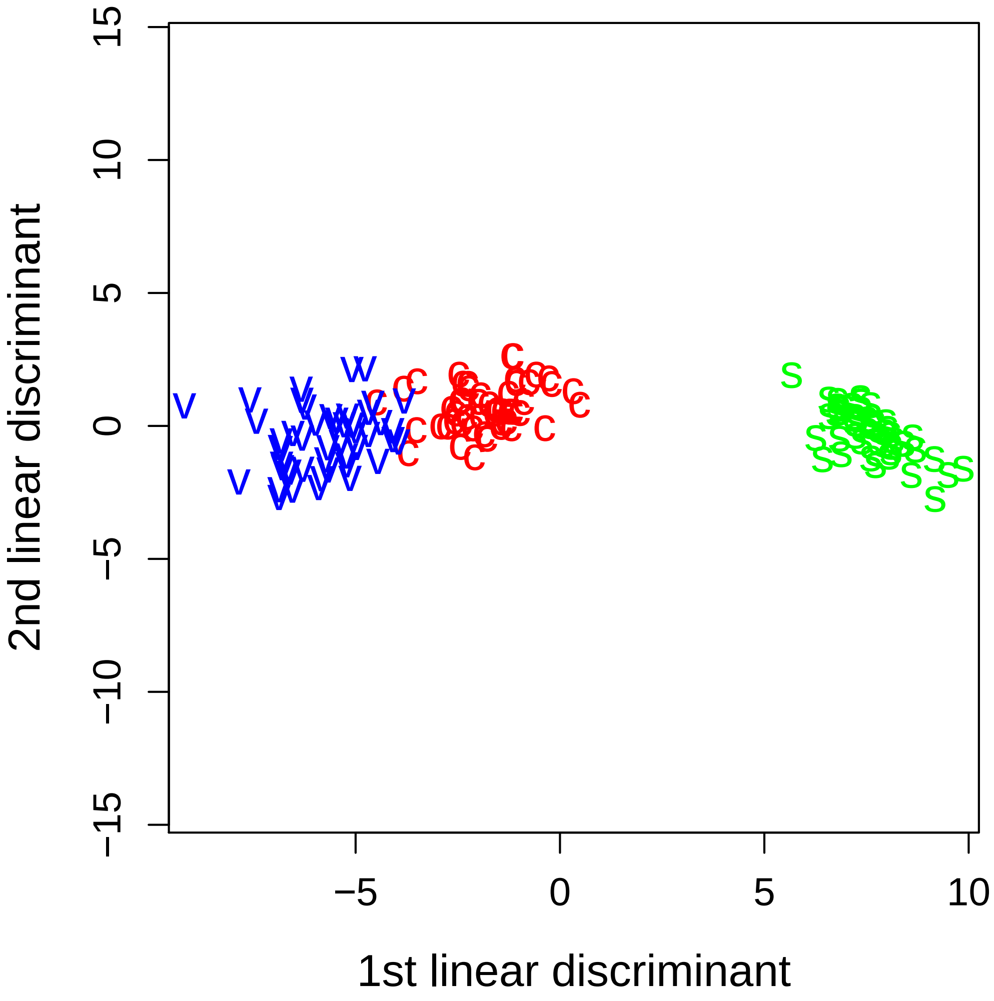


Classification: Fisher's iris data

- 3 classes (Setosa, Versicolour, Virginica)
- 50 observations per class
- 4 predictor variables (petal length and width, sepal length and width)

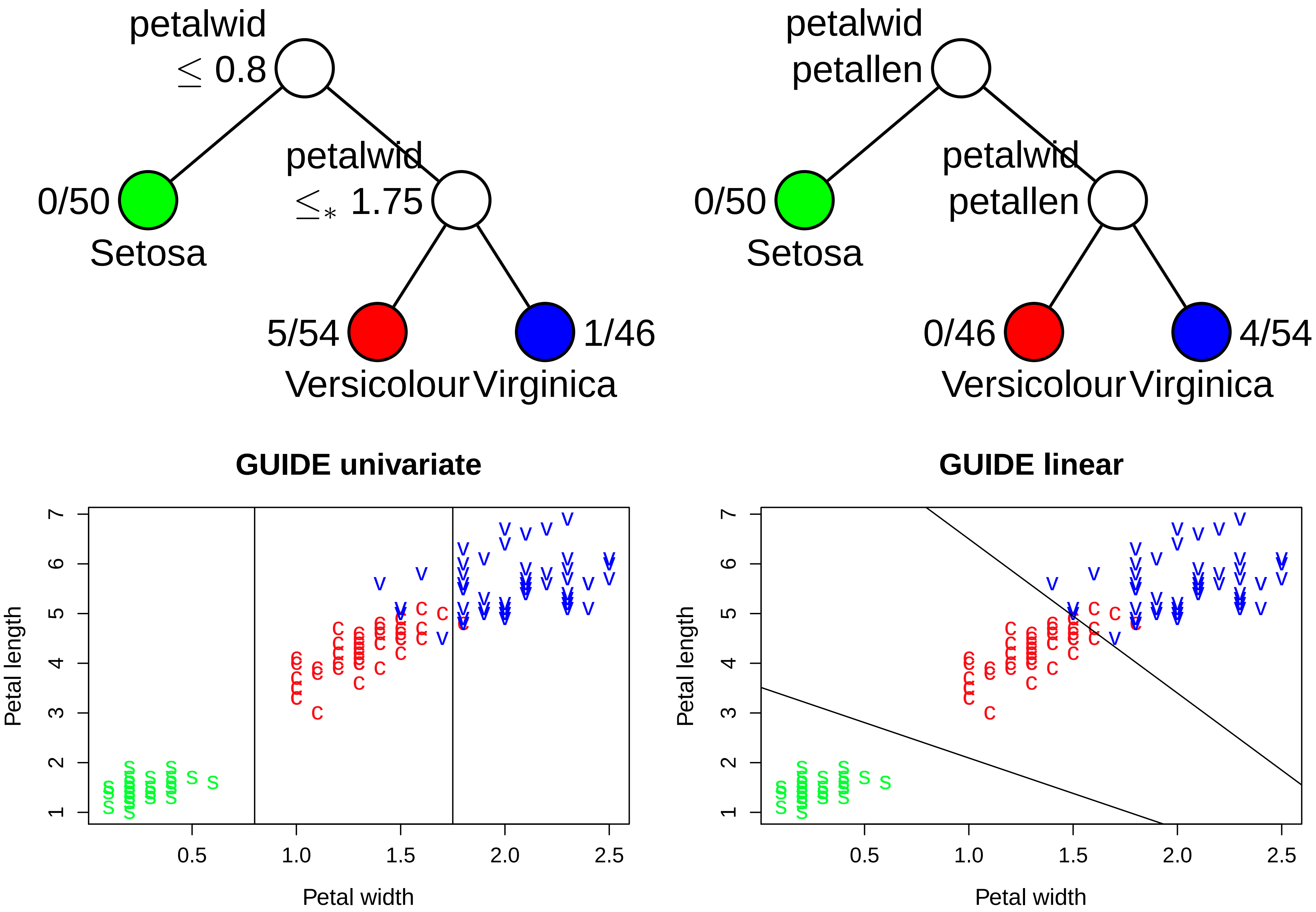


Plot of iris data in first 2 discriminant coords



s = Setosa, c = Versicolour, v = Virginica

Classification trees for iris data



GUIDE univariate

GUIDE linear

Women’s knowledge, attitude, and behavior toward mammography (Hosmer and Lemeshow, 2000)

- Data on 412 women and 3 classes:
  234 had no mammography experience;
  104 had a mammogram within the last year;
  74 had one more than a year ago
- 5 predictor variables: 2 binary; 2 ordered categorical; 1 ordinal

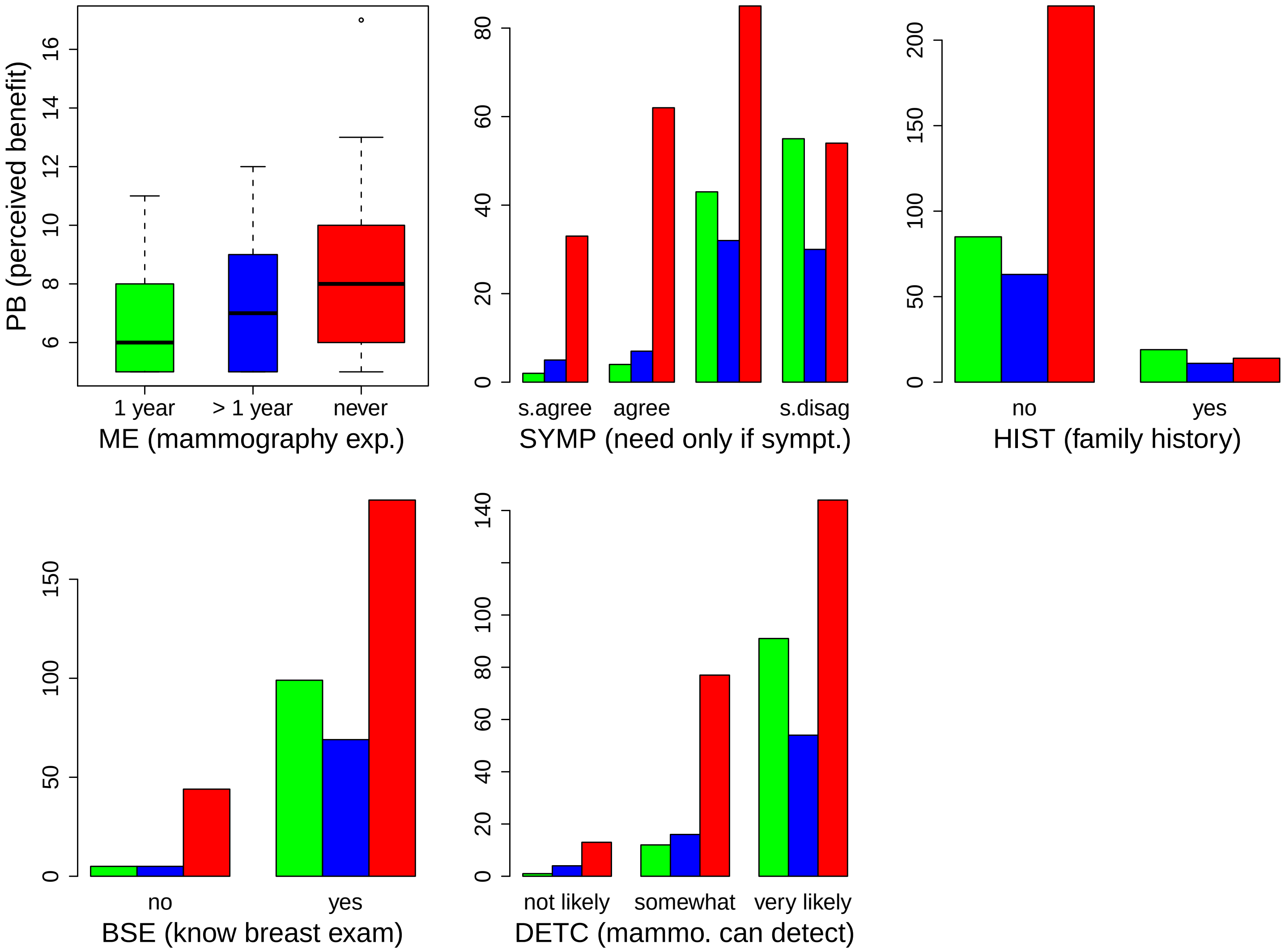
Unequal misclassification costs

	True class		
Predicted	1 ($\leq$ 1 yr)	2 ($>$ 1 yr)	3 (never)
1 ($\leq$ 1 yr)	0	1	2
2 ($>$ 1 yr)	1	0	1
3 (never)	2	1	0

Mammography variables

Name	Description	Values
ME	Mammography experience	within one year (1), over one year ago (2), never (3)
SYMP	You do not need a mammogram unless you develop symptoms	Strongly agree (1), agree (2), disagree (3), strongly disagree (4)
PB	Perceived benefit of mammography	5, 6, . . . , 20 (low values imply greater perceived benefit)
HIST	Mother or sister with history of breast cancer	no (0), yes (1)
BSE	Anyone taught you how to examine your own breasts?	no (0), yes (1)
DETC	How likely is it that a mammogram can find a new case of breast cancer?	Not likely (1), somewhat likely (2), very likely (3)

Distributions of predictor variables



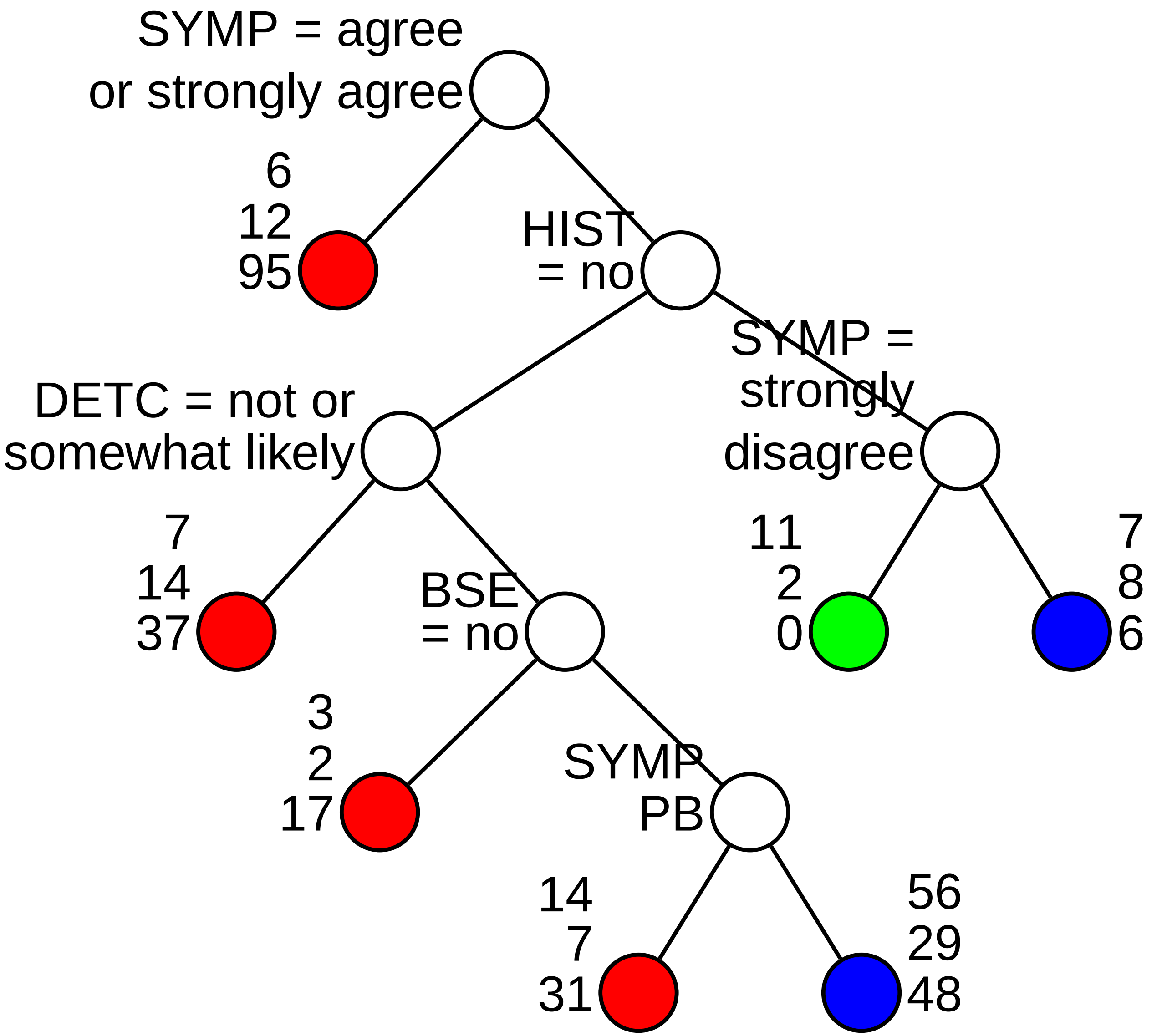
Multinomial logistic regression model with “ME = never” as baseline category

Logit(ME = within 1 year)				Logit(ME = more than 1 year)			
Variable	Coef	SE	P-value	Variable	Coef	SE	P-value
Constant	-2.62	0.93	0.005	Constant	-1.82	0.86	0.033
SYMPD*	2.10	0.46	<0.001	SYMPD*	1.13	0.36	0.002
PB	-0.25	0.07	0.001	PB	-0.15	0.07	0.034
HIST	1.31	0.43	0.003	HIST	1.06	0.45	0.019
BSE	1.24	0.53	0.019	BSE	0.96	0.51	0.056
DETCD**	0.89	0.36	0.019	DETCD**	0.11	0.32	0.720

* SYMPD = 1 if SYMP = “disagree” or “strongly disagree”, SYMPD = 0 otherwise

** DETCD = 1 if DETC = “very likely”, DETCD = 0 otherwise

GUIDE classification tree for mammography data



Within 1 year in green, more than one year in blue, never in red

Highly unbalanced classes: credit card data

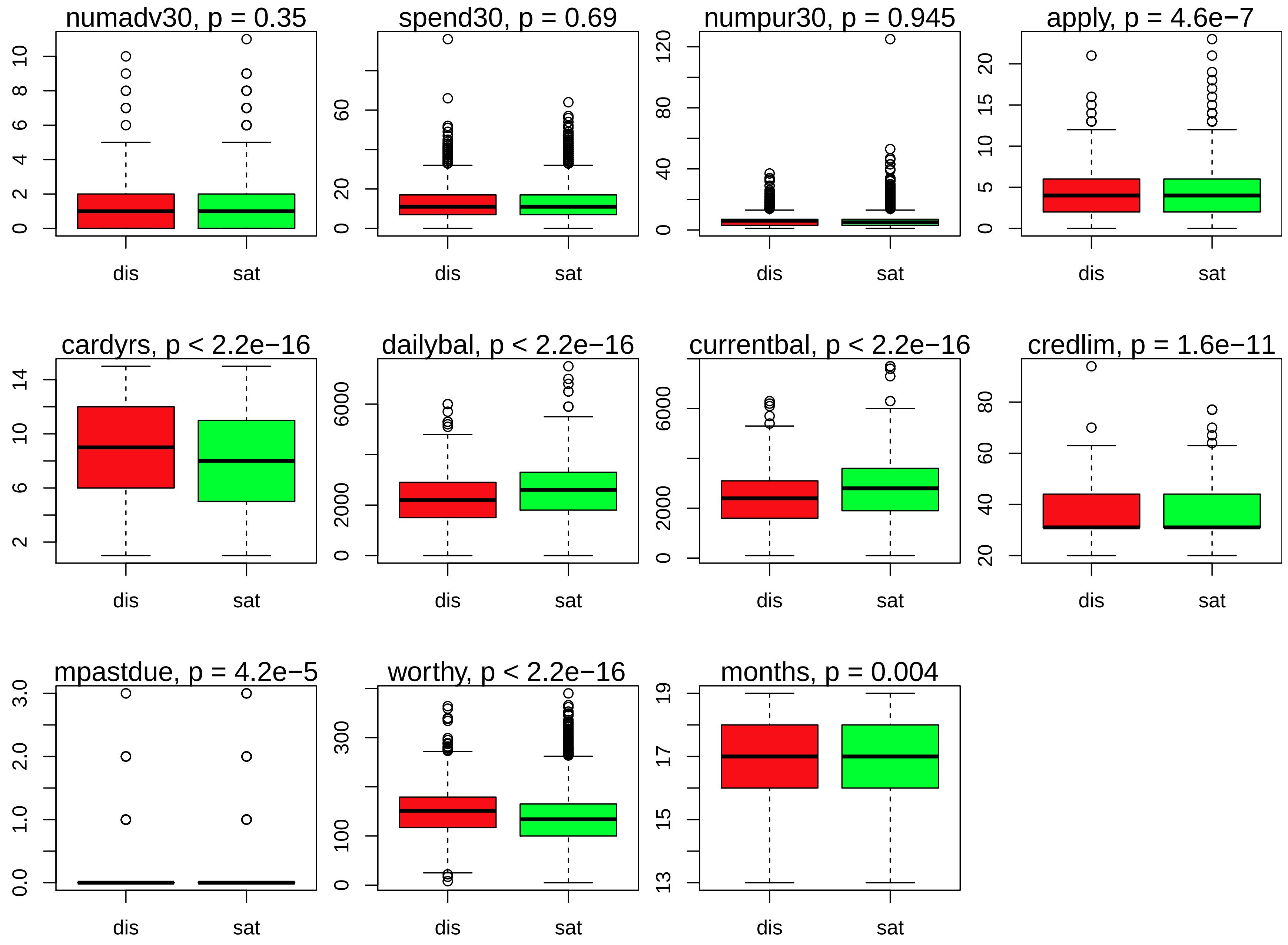
- Goal: A major credit card company wants to find out why 14.8% of its card holders are dissatisfied
- Data: 22,242 card holder records with information on 24 predictor variables
- Missing values: 1,752 records contain one or more missing values; 0.34% missing values overall
- Response variable: whether a card holder is satisfied with the card
- Problem: Low percent of dissatisfied card holders makes most methods classify everyone as “satisfied”—a useless result
- Two solutions: Use equal priors or make cost of misclassifying dissatisfied = 5.5 $\times$ that of satisfied (more emphasis on identifying dissatisfied card holders)

Predictor variables for credit card data

numadv30	How many times did you get cash advances in last 30 days?
spend30	How much money did you spend on purchases in last 30 days? ($)
numpur30	How many times did you make purchases in last 30 days?
over30	Have you gone over limit in last 30 days? (1=yes 0 = no)
otherbal	How much balance do you carry on other bank cards? (0=0K, 1=0–2.5k, 2=2.5K–5K, . . . , 8 = 17.5k–20k, 9 = 20k+)
othercred	How much credit do you have on other bank cards? (0=0K, 1=0–2.5k, 2=2.5K–5K, . . . , 8 = 17.5k–20k, 9 = 20k+)
apply	How many times did you apply for credit card in last year?
joint	Do you have a joint account? (1 = yes 0 = no)
employ	Are you currently employed? (1 = yes 0 = no)
cardyrs	How many years have you had any credit card?

dailybal	The average daily balance, unit in $
currentbal	The current balance, unit in $
credlim	The current credit limit, unit in $100
mpastdue	How many months the customer is past due
apr	The annual percent rate, unit in %
worthy	Historical index, credit worthiness, range [0,400]
months	How many months has the customer had the card?
init	Initial credit limit when account was opened, unit in $100
adv1	Cash advance indicator for month -1, 1 = yes, 0 = no
adv2	Cash advance indicator for month -2, 1 = yes, 0 = no
adv3	Cash advance indicator for month -3, 1 = yes, 0 = no
adv4	Cash advance indicator for month -4, 1 = yes, 0 = no
adv5	Cash advance indicator for month -5, 1 = yes, 0 = no
adv6	Cash advance indicator for month -6, 1 = yes, 0 = no

t-tests on ordered predictors



Chi-squared tests of categorical predictors

	over30 (p = 0.13)		joint (p = 0.47)		employ (p = 0.002)	
Satisfied	No	Yes	No	Yes	No	Yes
Yes	17951	836	3875	15079	2394	16560
No	3132	125	691	2597	351	2937

	otherbal (p = 1.5×10^{-13})								
Satisfied	1	2	3	4	5	6	7	8	9
Yes	9281	4711	1610	1308	497	471	199	194	533
No	1370	947	356	242	98	92	19	34	109

	othercred ($p < 2.2 \times 10^{-16}$)								
Satisfied	1	2	3	4	5	6	7	8	9
Yes	3304	6107	2393	2469	1056	1075	505	522	1435
No	312	915	491	501	227	256	120	110	343

Chi-squared tests of categorical predictors (cont'd.)

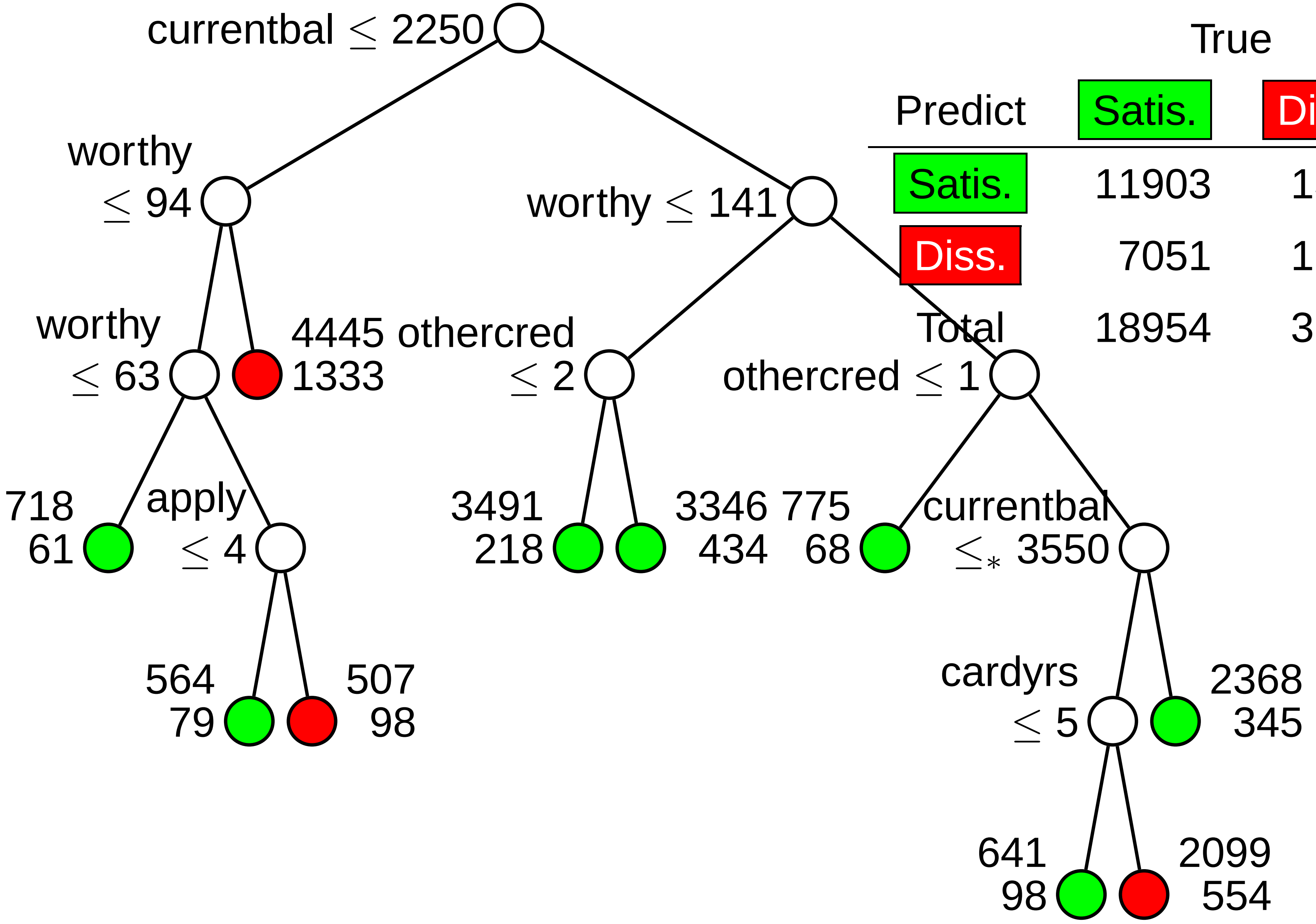
	apr (p = 0.002431)								
Satisfied	4	5	6	7	8	9	10	11	15
Yes	164	5	273	36	459	4	145	17386	482
No	24	6	42	11	59	1	27	3044	74

	init ($p < 2.2 \times 10^{-16}$)			
Satisfied	20	24	31	44
Yes	3375	13	8062	7367
No	773	8	1375	1114

Logistic regression model for Pr(Dissatisfied)

Variable	Estimate	p-value
(Intercept)	-1.802e+00	7.12e-07
numadv30	-1.442e-02	0.517144
spend30	2.661e-03	0.399596
numpur30	3.477e-03	0.594214
over30	6.561e-02	0.529030
otherbal	-7.053e-02	*2.22e-05*
othercred	1.351e-01	*<2e-16*
apply	3.229e-02	*8.97e-05*
joint	-8.693e-02	0.081735
employ	2.313e-01	*0.000356*
cardyrs	3.080e-02	*4.05e-09*
dailybal	-5.665e-05	0.161080
currentbal	-2.623e-04	*1.83e-12*
credlim	4.218e-02	*8.20e-05*
mpastdue	4.479e-01	*3.42e-06*
apr	1.556e-02	0.375681
worthy	5.604e-03	*<2e-16*
months	-4.112e-02	*0.003214*
init	-5.195e-02	*2.19e-06*
adv1	-9.934e-02	0.374672
adv2	-8.055e-03	0.938932
adv3	-3.709e-02	0.752908
adv4	-2.381e-02	0.827685
adv5	1.072e-01	0.310609
adv6	-2.010e-02	0.841265

GUIDE tree with equal priors (or 5.5 to 1 costs)



	True	
Predict	Satis.	Diss.
Satis.	11903	1303
Diss.	7051	1985
Total	18954	3288

Properties of an ideal classifier

High predictive accuracy: classify unseen cases with low error

Intuitive, comprehensible structure: give insight into the roles and relative importance of the predictor variables

Correct, unbiased inference: draw inferences without bias

Fast training time: construct models quickly

Definition

A classifier or classification rule is a function $d(\mathbf{x})$ defined on $\mathcal{X}$ such that for every $\mathbf{x}$, $d(\mathbf{x})$ is equal to one of the numbers $1, 2, \ldots, J$.

A classifier is a partition of the sample space $\mathcal{X}$ such that

$$\begin{aligned} A_j &= \{\mathbf{x} : d(\mathbf{x}) = j\} \\ \mathcal{X} &= \cup_j A_j \end{aligned}$$

Notations

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

AID (Morgan and Sonquist, 1963)

AID is the first published regression tree algorithm. It works as follows.

1. Recursively partition the data with splits of the form “$X \le c$” (ordinal X) and “$X \in S$” (categorical X).

2. At each stage, choose the split that minimizes a measure of node impurity, e.g., sum of squared deviations from mean: $\sum (y_i - \bar{y})^2$.

3. Stop splitting if reduction in impurity is below preset value.

THAID (Messenger and Mandell, 1972)

THAID is the first published classification tree algorithm (categorical Y)

- At each node, count the number of observations in the most frequent Y category (modal category)
- Choose the split that maximizes the sum of observations in the modal categories of the subnodes
- Follow the rest of the AID algorithm

CART (Breiman et al., 1984)

1. Choose the split that maximizes the decrease in node impurity (Gini index for classification, sum of squared errors for regression)

2. For classification, let $C(i|j)$ be cost of misclassifying a class j as class i. Assign terminal node t to class j^* if it minimizes the misclassification cost

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

   For regression, use the sample Y mean in t as predicted value

3. Prune tree using test sample or cross-validation

4. Use surrogates splits to deal with missing values

Estimates of misclassification error

Resubstitution estimate. Use the training data:

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

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

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

V-fold cross-validation estimate.

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

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

3. The V-fold cross-validation estimate is

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

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

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: $i(t) = 1 - \sum_j p^2(j|t)$

- 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 set 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,

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

   (a) If X is ordinal with k unique values, there are $(k-1)$ splits

   (b) If X is categorical with k unique values, there are $(2^{k-1}-1)$ splits

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

Shortcut for categorical splits with 2 classes

Theorem 1 *Let X be a categorical variable taking values in $\{b_1, \ldots, b_L\}$.*
Suppose $i(t) = \phi(p(1|t))$, where ϕ is strictly concave.
Define $(b_{l(i)}; i = 1, \ldots, L)$ such that

$$p(1|X = b_{l(1)}) \ \leq \ p(1|X = b_{l(2)}) \ \leq \ \ldots \ \leq \ p(1|X = b_{l(L)})$$

Then the split on X that maximizes the decrease in impurity is one of the splits:

$$X \in \{b_{l(1)}, \ldots, b_{l(h)}\}, \ \ h = 1, \ldots, L-1$$

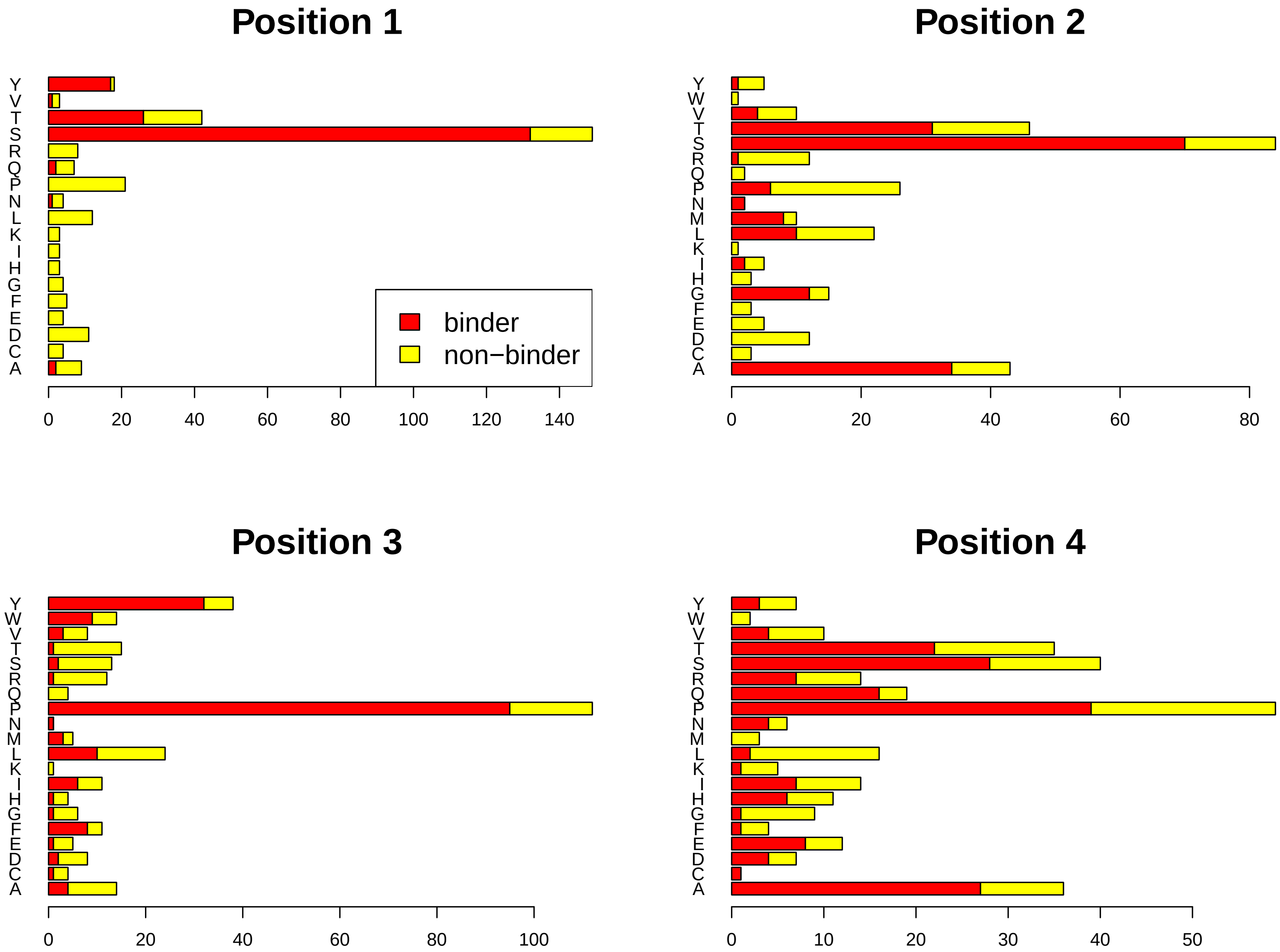
This reduces the search from $(2^{L-1} - 1)$ subsets to $(L - 1)$ subsets
Proof: See Breiman et al. (1984, Sec. 9.4)

Categorical predictors: peptide-binding data

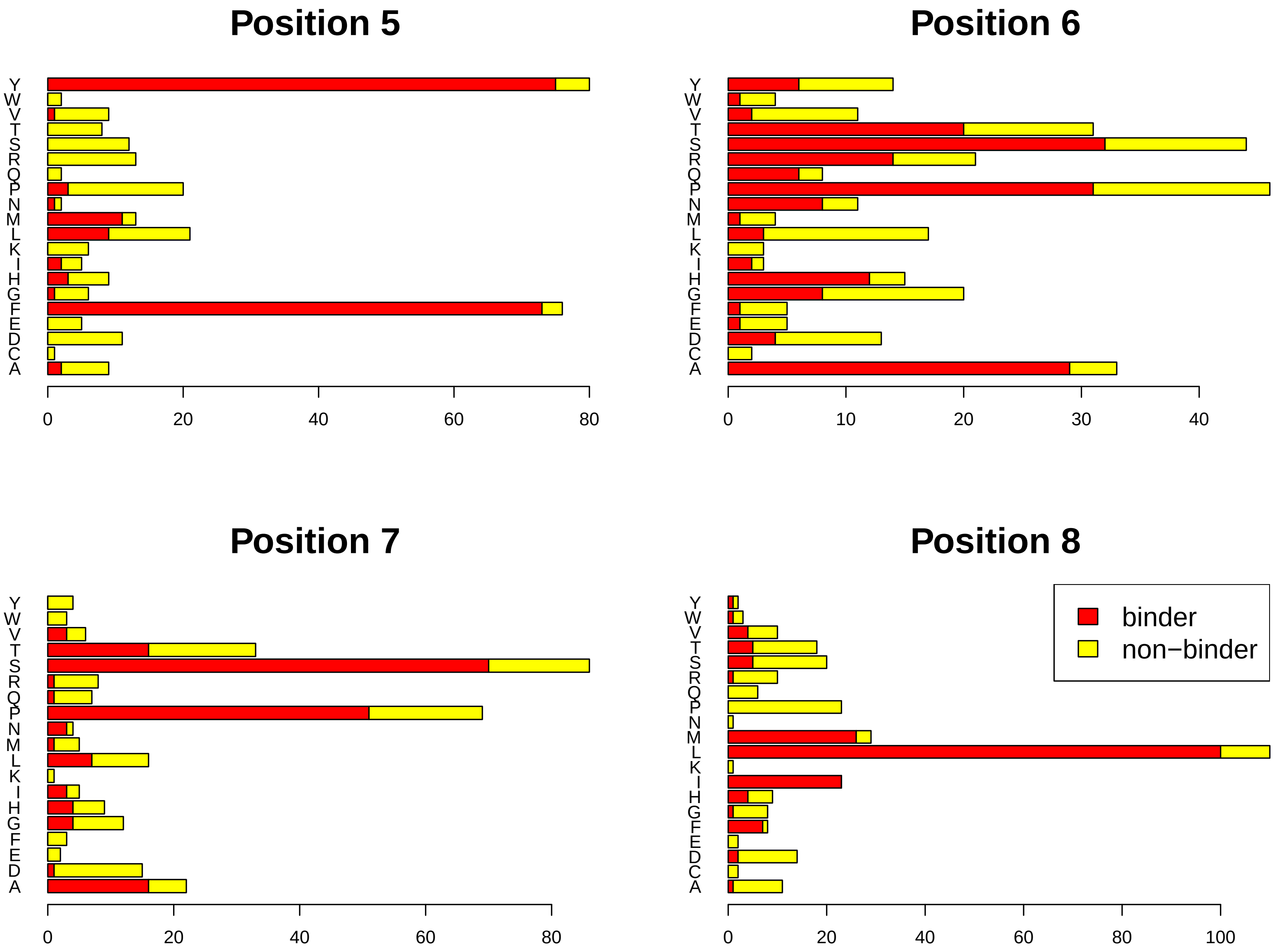
- 310 amino acid sequences of peptides
- 181 bind to a class of MHC molecule, 129 do not
- Each amino acid sequence has length 8
- Each position in a sequence is one of 18–20 amino acids
- Problem: What amino acids in which positions are predictive of binding?
- Milik et al. (1998) convert amino acid info into 104 numerical "property variables" and use neural networks
- Segal et al. (2001) use CART

  `http://repositories.cdlib.org/cbmb/peptide_binding`

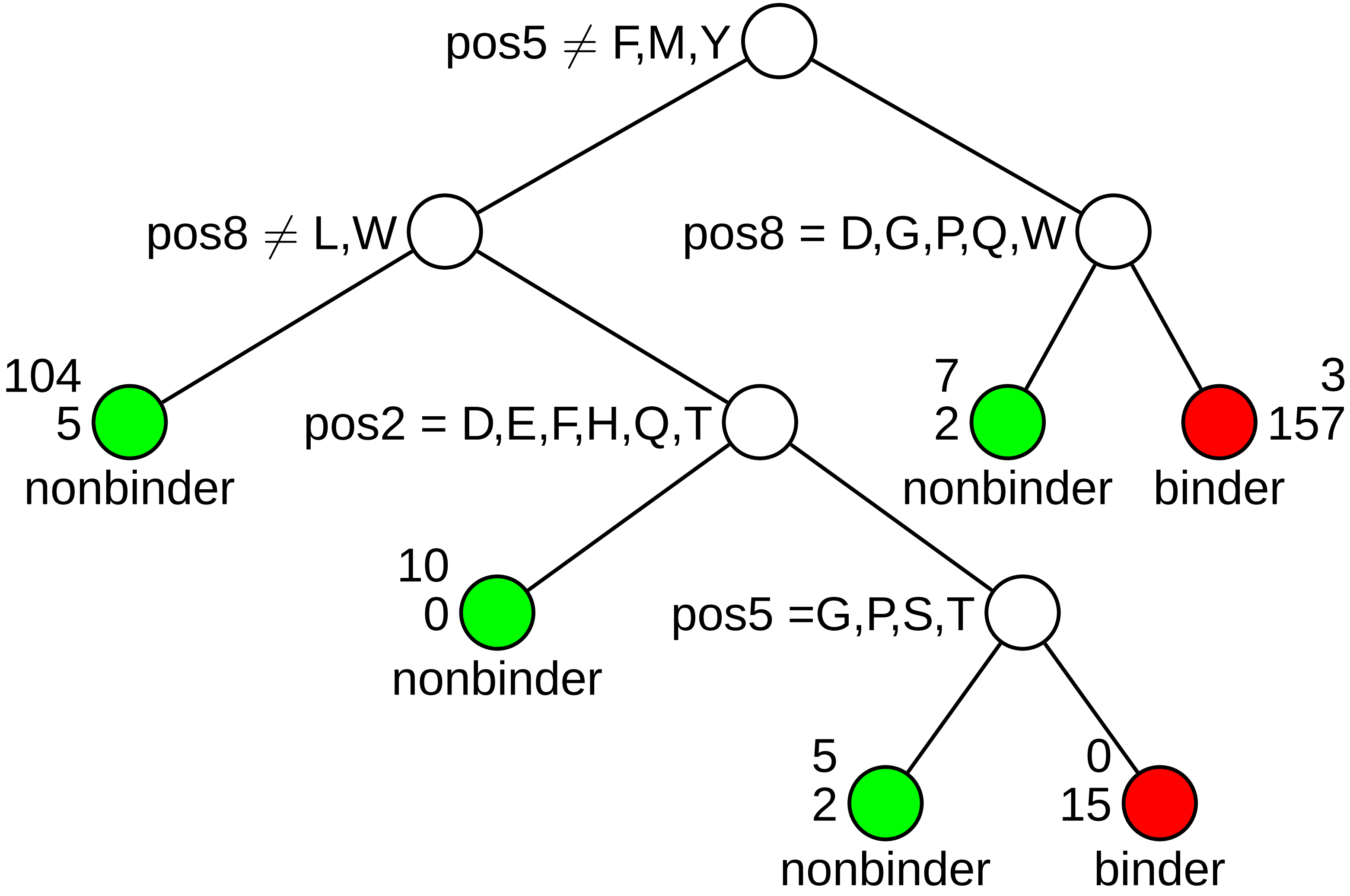
Distributions of peptide-binding data



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



RPART (Therneau and Atkinson, 2012) tree for peptide data



Counts of nonbinder and binder, resp., are beside terminal nodes

Levels of Pos5 ordered by P(Y = 0)

	Class					Class			
Level	0	1	Total	Prop.	Level	0	1	Total	Prop.
F	3	73	76	0.039	V	8	1	9	0.889
Y	5	75	80	0.063	C	1	0	1	1
M	2	11	13	0.154	D	11	0	11	1
N	1	1	2	0.5	E	5	0	5	1
L	12	9	21	0.571	K	6	0	6	1
I	3	2	5	0.6	Q	2	0	2	1
H	6	3	9	0.667	R	13	0	13	1
A	7	2	9	0.778	S	12	0	12	1
G	5	1	6	0.833	T	8	0	8	1
P	17	3	20	0.85	W	2	0	2	1

Resubstitution estimate of misclassification cost

- Let $\pi(j)$ be the prior probability of class j
- Let $N_j(t)$ be the number of class j observations in node t
- Let N_j be the number of class j observations in the training sample
- Let $p(j,t) = \pi(j)N_j(t)/N_j$ be the estimated probability of being in class j and in node t
- Define $p(t) = \sum_j p(j,t)$ and $p(j|t) = p(j,t)/p(t)$
- The resubstitution estimate of expected misclassification cost of node t is

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

- The resubstitution estimate of expected misclassification cost of a tree T is

$$R(T) = \sum_{t \in \tilde{T}} r(t)p(t)$$

Why not use $R(t)$ as impurity function?

- Optimal split is not unique: possible for $R(t) - R(t_L) - R(t_R) = 0$ for some or all splits
- Shortcut algorithm for categorical split is not applicable because $R(t)$ is not a strictly concave function of $\{p(j|t)\}$

CART pruning

1. Given α and tree T, define the cost-complexity $R_\alpha(T) = R(T) + \alpha|\tilde{T}|$
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

$$\begin{aligned} R_\alpha(\{t\}) &= R(t) + \alpha \\ R_\alpha(T_t) &= R(T_t) + \alpha|\tilde{T}_t| \end{aligned}$$

4. $R_\alpha(T_t) = R_\alpha(\{t\})$ when $\alpha = u(t) = [R(t) - R(T_t)]/[|\tilde{T}_t| - 1]$
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

Sequence of minimal cost-complexity trees is a subsequence of the subtrees constructed by finding the minimum cost subtree for a given number of terminal nodes.

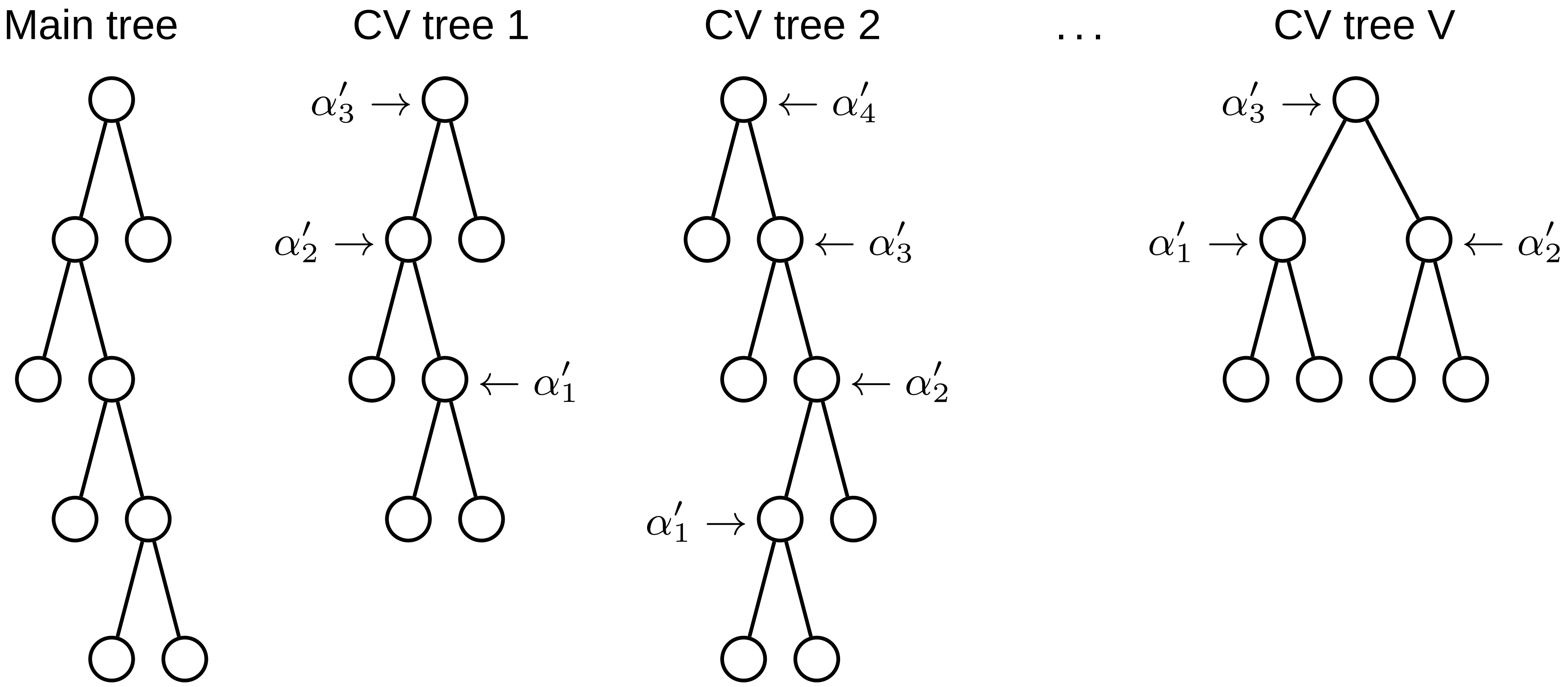
Subtree selection by V-fold cross-validation

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



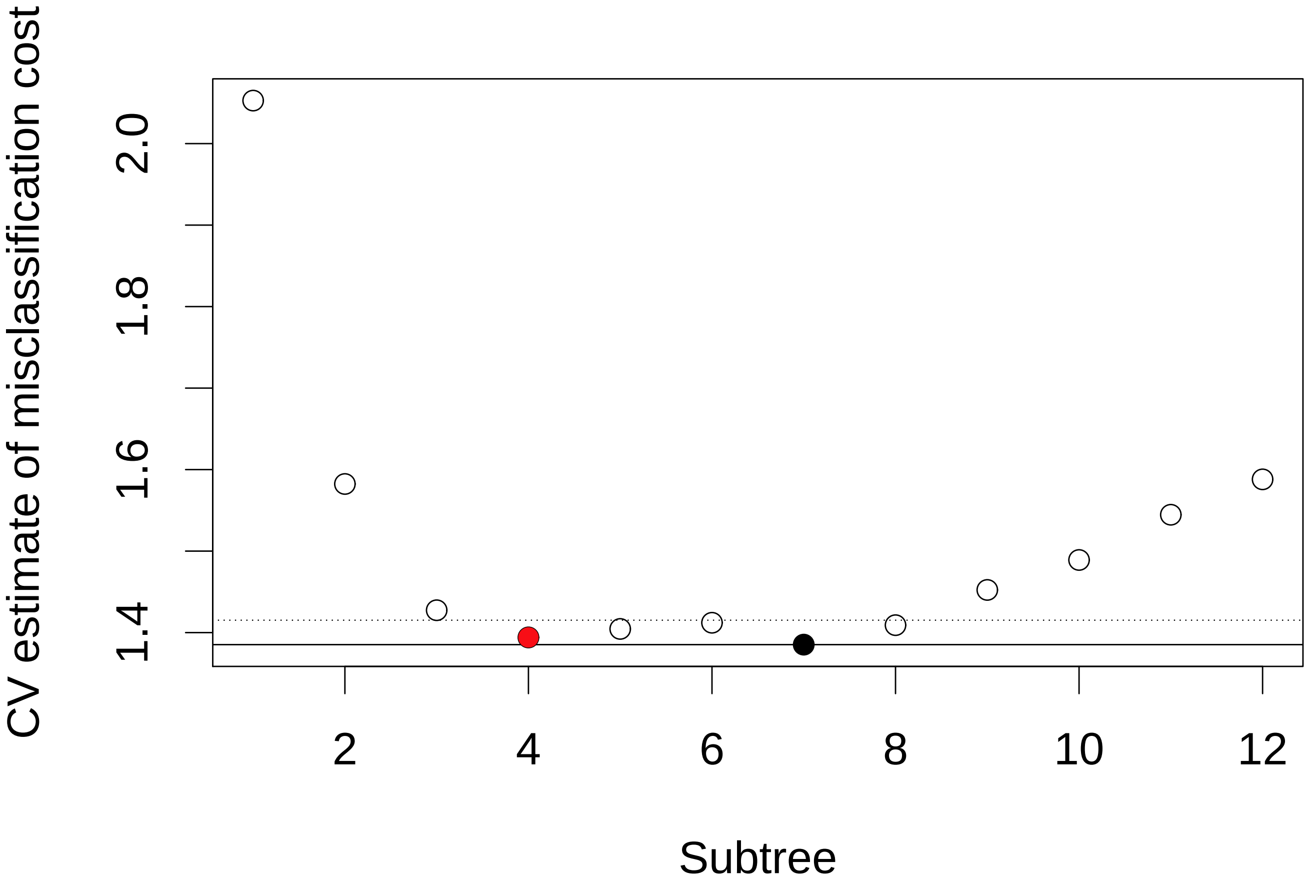
- Main tree is grown using all the data
- Each CV tree is grown using $(V - 1)$ subsets

k-SE rule

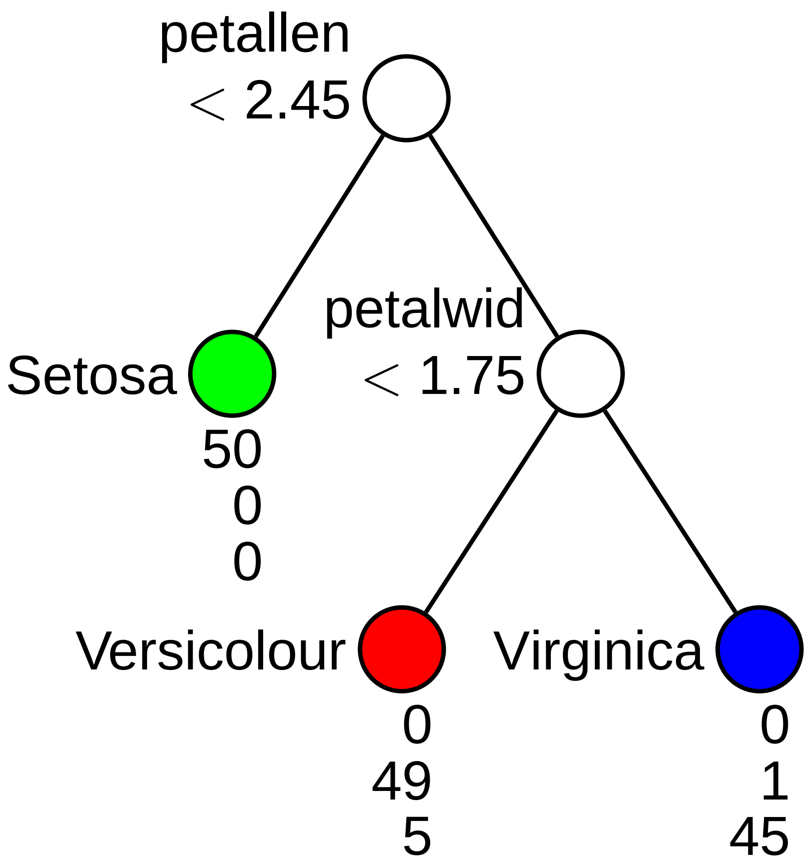
1. Let $\hat{R}(T_j)$ be the CV estimate of misclassification cost of T_j, let T^* be the tree with min. value of $\hat{R}(T_j)$, and let SE be the standard error of $\hat{R}(T^*)$

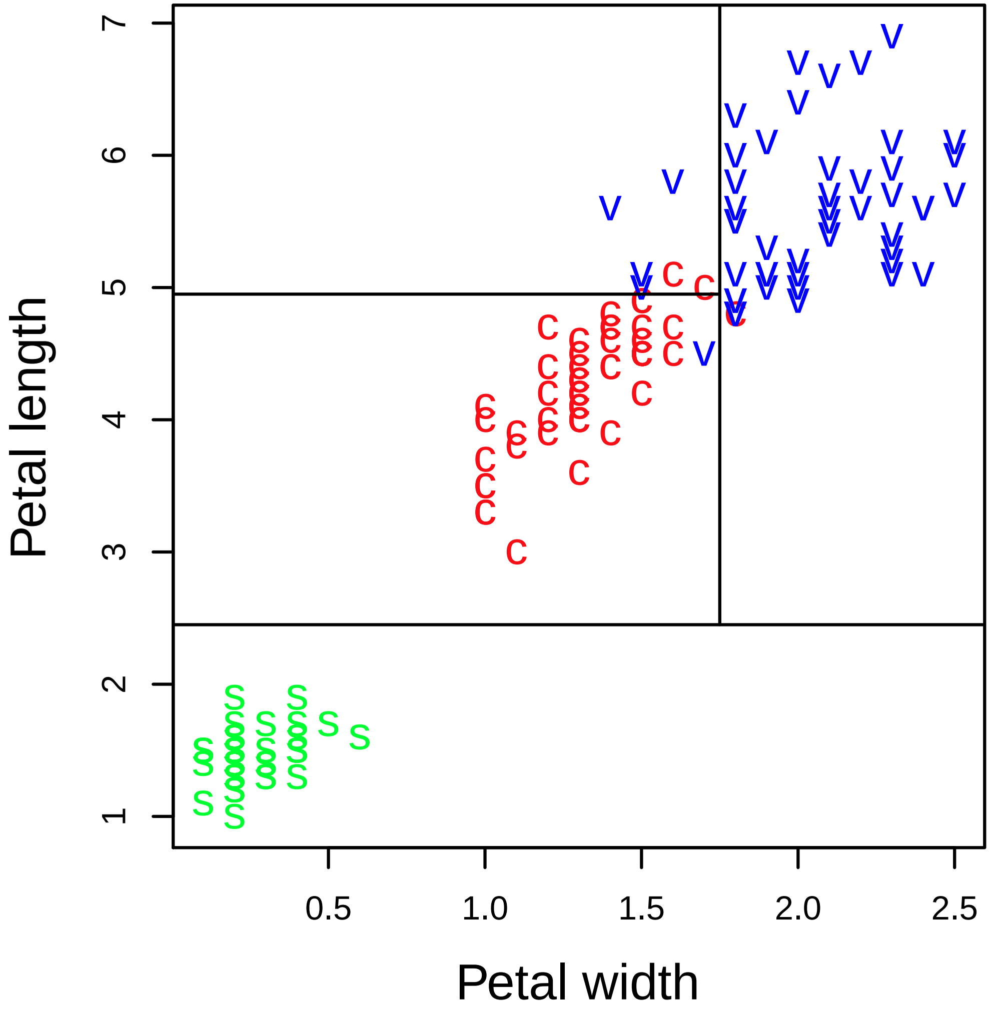
2. The k-SE tree T^{**} is the smallest subtree such that

$$\hat{R}(T^{**}) \leq \min_j \hat{R}(T_j) + k \times \mathrm{SE}$$



RPART tree for iris data





Unequal misclassification costs via Gini

- The Gini index can be generalized to:

$$i(t) = \sum_{i,j} C(i|j)p(i|t)p(j|t)$$

This reduces for $J = 2$ to

$$i(t) = [C(2|1) + C(1|2)]p(1|t)p(2|t)$$

which gives the same split criterion as for unit costs

- Disadvantage: Index symmetrizes the cost matrix

Unequal misclassification costs via altered priors

- Let $\pi(j)$ be the prior probability of class $j \in \mathcal{C}$
- Let $Q(i|j)$ be the proportion of class j cases in $\mathcal{L}$ classified as class i by T
- Resubstitution estimate of T is $R(T) = \sum_{i,j \in \mathcal{C}} C(i|j)Q(i|j)\pi(j)$
- The value of $R(T)$ is the same if $\{\pi'(j)\}$ and $\{C'(i|j)\}$ satisfy

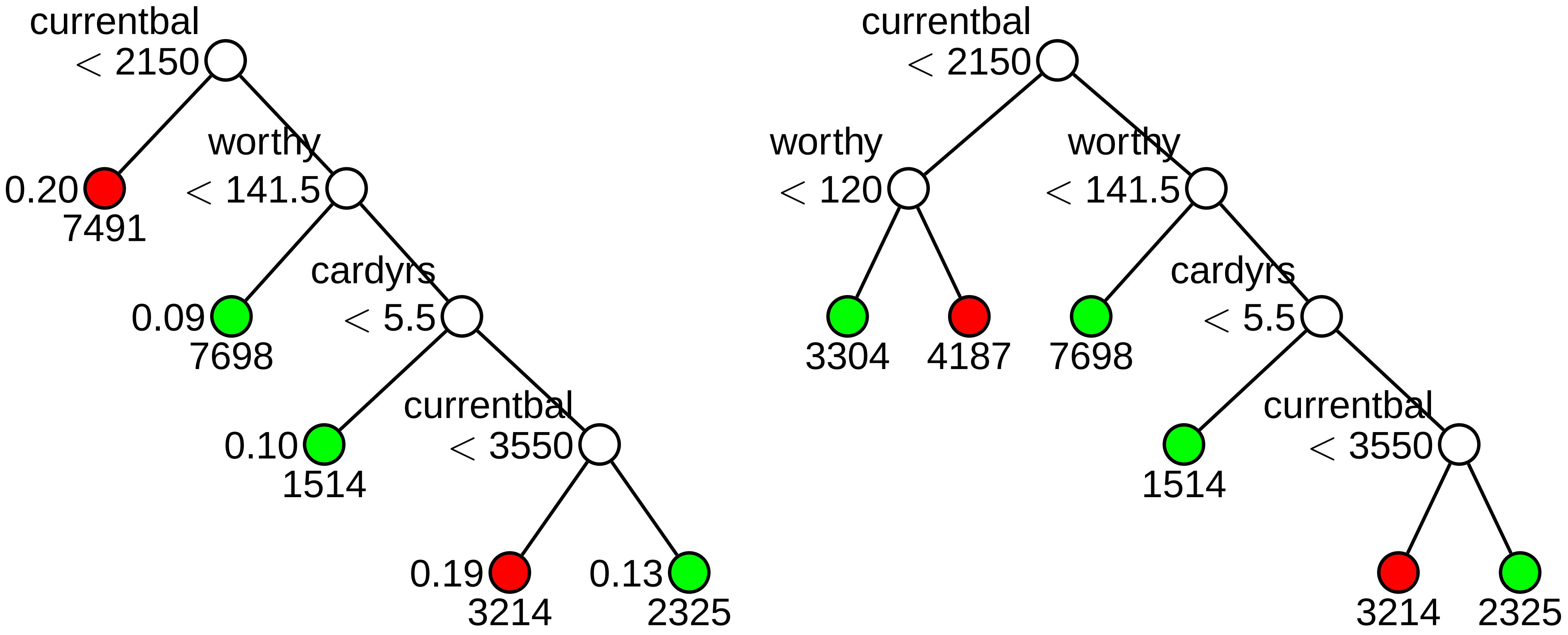
$$C'(i|j)\pi'(j) = C(i|j)\pi(j), \quad i, j \in \mathcal{C}$$

- Thus unequal $C(i|j)$ can be accommodated by altering $\pi(j)$ to $\pi'(j)$
- If $C(i|j) = C(j)$, $i \neq j$ for each j, define $C'(i|j) = 1$, $i \neq j$ and

$$\pi'(j) = \frac{C(j)\pi(j)}{\sum_i C(i)\pi(i)}$$

- Otherwise, use $C(j) = \sum_i C(i|j)$ in the above formula for $\pi'(j)$
- Disadvantage: Only uses the values of $\sum_i C(i|j)$

RPART trees for credit card data: equal priors (left), 5.5:1 costs (right)



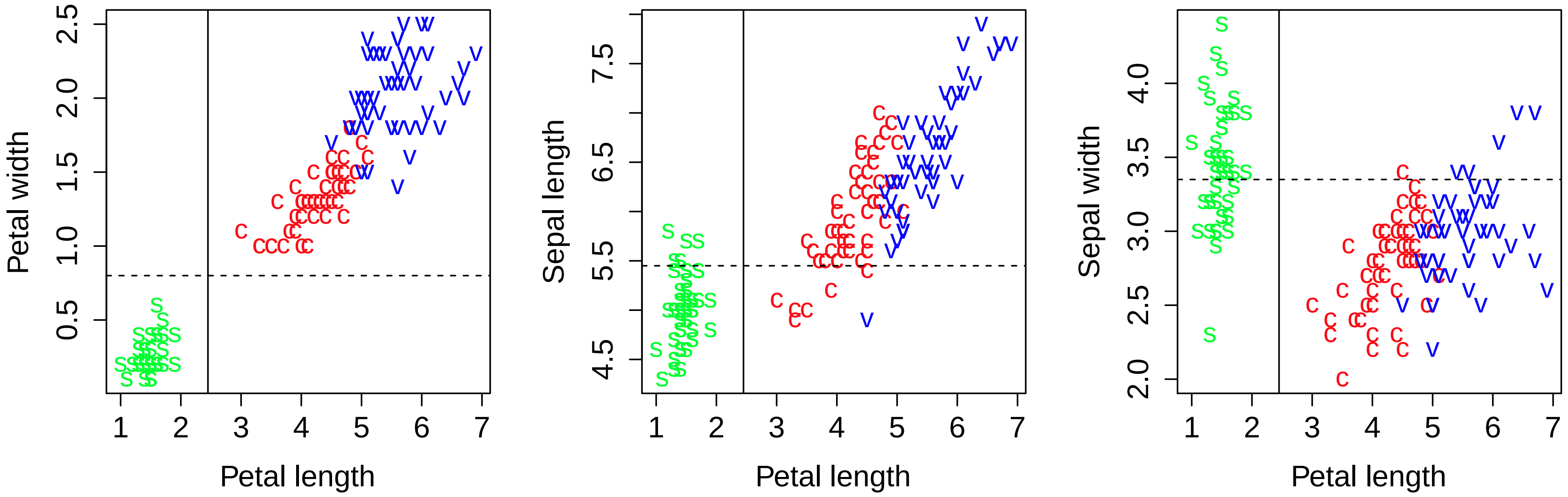
Dissatisfied and satisfied nodes in red and green colors

P(Dissatisfied) beside node in left tree; Sample sizes beneath nodes

Missing values: CART surrogate splits

Suppose $X \in S$ is selected to split node t

1. For each $X_i \neq X$, find the split $X_i \in S_i$ that best predicts $X \in S$ in terms of maximizing the number, M_i, of observations going to the corresponding subnodes
2. Order the X_i in terms of M_i to form a preferential set of surrogate splits



s = Setosa, c = Versicolour, v = Virginica

CART surrogate splits: the details

1. Recall that $p(j,t) = \pi(j)N_j(t)/N_j$ and $p(t) = \sum_j p(j,t)$
2. Let s^* be the best split of t into t_L and t_R
3. For each k, let $\mathcal{S}_k$ be the set of all splits on x_k
4. Let $s \in \mathcal{S}_k$ with subnodes t'_L and t'_R
5. Let $N_j(LL)$ be the number of class j cases in $t_L \cap t'_L$
6. Define $p(t_L \cap t'_L) = \sum_j \pi(j)N_j(LL)/N_j$

- 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 (not available in RPART)

4. Detect masking of variables

CART classification tree construction when there are missing values in the learning sample

Univariate splits: Find the best split s_k^* on each x_k using only cases non-missing in x_k. Select the split s^* that maximizes $\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R)$. Note: $i(t)$ is constant for all splits but p_L, p_R, $i(t_L)$, and $i(t_R)$ are computed from the non-missing values only. This induces a selection bias (Therneau and Atkinson, 2013, pp.18–19).

Linear combination splits: Find the best split s^* using only cases non-missing in all variables

Passing a case with missing values through the split: Let $\tilde{s}_m$ be the surrogate split based on each variable x_m that is nonmissing for the case. Let $\tilde{s}_{m^*}$ be the surrogate split among them with the highest measure of association with s^*. The split $\tilde{s}_{m^*}$ is used on the case in place of s^*.

CART classification of a new case with missing values

- Let s^* be the split at a node. Suppose the new case is missing some variable(s) that are required by s^*.

- Among all nonmissing variables in the case, find the one whose surrogate split $\tilde{s}_k$ (say) has the highest measure of association with s^*.

- Send the case down using $\tilde{s}_k$. If no $\tilde{s}_k$, send the case to the larger node.

Notes on RPART:

1. If a split variable has no missing training values, it has no surrogate splits. In that case, new cases with missing values are sent to the larger node.

2. If a split is on a categorical variable X and a new case has an X value not in the training sample, RPART will return an error.

Importance ranking of predictor variables in CART

- 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

Problems with CART classification

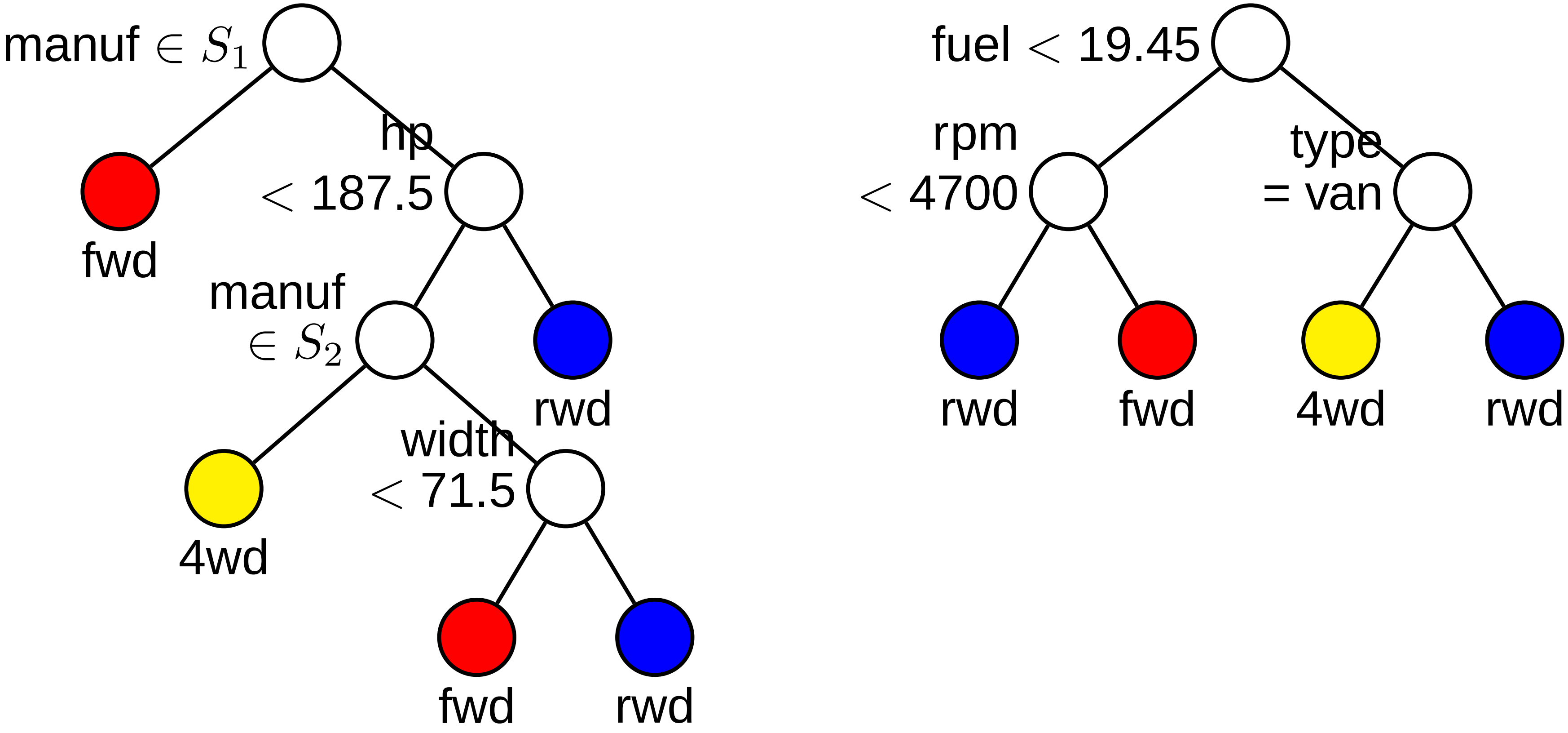
- Biased toward variables with more splits: A k-valued ordered variable has $(k-1)$ splits; a k-valued categorical variable has $(2^{k-1}-1)$ splits.
- Biased toward predictors with more missing values: Split method uses only proportions of nonmissing cases—it ignores the number of missing values. A variable taking a unique value for exactly one case in each class and missing on all other cases yields the largest decrease in impurity. Bias exists for surrogate splits too.
- Computation: Impractical when there are three or more classes and categorical variables with many values. Note: Because CART and RPART encode each categorical variable split with a 32-bit binary integer, they do not properly deal with categorical variables having more than 32 values.
- Prediction accuracy: Often no better than linear discriminant analysis.

Predicting drive train for 1993 model year cars (Lock, 1993)

- 93 cars and 25 variables (3 categorical, 2 binary, 20 ordinal)
- Drive train takes three values: 16 (17.2%) rear (rwd), 67 (72.0%) front (fwd), and 10 (10.8%) four-wheel (4wd) drive

Variable	Description	Variable	Description
manuf	Manufacturer (31 values)	rev	Engine revolutions/mile
type	Type (small, sporty, compact, midsize, large, van)	manual	Manual transmission available (yes, no)
minprice	Minimum price (in $1,000)	fuel	Fuel tank (gallons)
midprice	Midrange price (in $1,000)	passngr	Passenger capacity
maxprice	Maximum price (in $1,000)	length	Length (inches)
citympg	City miles per gallon	whlbase	Wheelbase (inches)
hwympg	Highway miles per gallon	width	Width (inches)
airbag	Air bags standard (0, 1, 2)	uturn	U-turn space (feet)
cylin	Number of cylinders	rseat	Rear seat room (inches)
enginzs	Engine size (liters)	luggage	Luggage capacity (cu. ft.)
hp	Maximum horsepower	weight	Weight (pounds)
rpm	Revolutions per minute	domestic	U.S./non U.S.

RPART trees with (left) and without (right) manuf



- S_1 = {Acura, Audi, Buick, Cadillac, Chrysler, Dodge, Eagle, Geo, Honda, Hyundai, Mitsubishi, Nissan, Oldsmobile, Pontiac, Saab, Saturn, Suzuki, Toyota, VW}
- S_2 = {Mercedes-Benz, Plymouth, Subaru, Volvo}
- S_2^c = {Chevrolet, Ford, Lexus, Lincoln, Mazda, Mercury}
- Trees took 821.6s (13.7m) and 0.023s, respectively, to construct

FACT (Loh and Vanichsetakul, 1988)
Classification trees with two or more splits/node

An approximate, quick, and fairly accurate solution with J splits per node:

1. Replace missing values by means and modes at each node
2. Convert each categorical variable to a dummy vector and then transform to largest discriminant variable (crimcoord)
3. For linear splits, use recursive linear discriminant analysis (LDA)
4. For univariate splits:
   (a) Use one-way ANOVA to choose split variable or crimcoord
   (b) Use LDA on selected variable or crimcoord to split node
   (c) If split is on crimcoord, re-express it as an univariate split $X \in S$
5. Use weighted sums of ANOVA F-statistics as importance scores

FACT method for categorical variable splits

1. Suppose X takes values in the set $\{a_1, \ldots, a_c\}$
2. Define dummy vector $D = (d_1, \ldots, d_{c-1})$ with $d_i = I(X = a_i)$
3. Project the D-data onto the largest discriminant coordinate (crimcoord) $U = \sum_i b_i I(X = a_i)$
4. Search for a split of the form '$U \leq c$'
5. Re-express the split as '$X \in A$' with $A = \{a_i : b_i \leq c\}$

QUEST (Loh and Shih, 1997)
First algorithm with unbiased variable selection

1. If $J > 2$, use 2-means clustering of class means to form 2 superclasses
2. For univariate splits:
   (a) Find p-value of 1-way ANOVA for each ordinal variable
   (b) Find p-value of χ^2 test of independence for each categorical variable
   (c) Select variable with smallest p-value to split node
   (d) Transform each categorical variable to a crimcoord
   (e) Use QDA on selected variable or crimcoord to find split
3. For linear combination splits, use FACT method (LDA on ordinal and crimcoord variables)
4. Use mean/mode imputation for missing values at each node
5. Use CART method to prune the tree

CRUISE (Kim and Loh, 2001, 2003)
First unbiased algorithm with multiple splits

1. Find p-value of χ^2 test of Y vs. each variable, with ordinal variables discretized (replaces F test of QUEST)

2. Find p-value of χ^2 test of Y vs. each **pair** of variables (adds ability to detect local interactions)

3. Select the variable(s) with smallest p-value; if latter is from an interaction test, select the variable with smaller marginal p-value

4. If selected variable is categorical, transform it to a crimcoord

5. Use Box-Cox transformations and LDA to split on selected variable

6. For linear combination splits, use LDA on all variables

7. Use different surrogate split methods for missing values

8. Optionally fit bivariate LDA models in nodes

9. Use CART method to prune the tree

CRUISE ‘alternate variable’ missing value method

1. For univariate splits:
   (a) Compute χ^2 tests using non-missing cases in the respective variables
   (b) For tree construction, impute missing values with class mean/mode
   (c) For predicting new cases, use the next best split at the node to predict the class and then impute with its mean/mode
2. For linear combination splits:
   (a) For tree construction, impute with class mean/mode
   (b) For predicting new cases:
      i. Use best univariate split to predict class; then impute with estimated class mean/mode
      ii. If variable in best univariate split is also missing, impute with grand mean/mode

Prob. of surrogate/alternate variable selection

	CART					CRUISE				
	Percent missing X_1					Percent missing X_1				
	1	2	3	4	25	1	2	3	4	25
X_1	.18	.12	.09	.05	.00	.19	.20	.18	.20	.18
X_2	.25	.25	.26	.24	.30	.18	.22	.18	.19	.19
X_3	.21	.23	.26	.27	.25	.22	.19	.20	.21	.19
X_4	.20	.23	.20	.23	.23	.22	.19	.22	.22	.21
X_5	.17	.17	.19	.21	.22	.20	.20	.22	.18	.23

- $Y \sim$ Bernoulli(1/2), $X_0 \sim N(0.3Y, 1)$, and $X_1, \ldots, X_5$ indep. $N(0, 1)$
- Variable X_1 has missing values but others do not
- Estimates based on 1000 iterations and $n = 200$ in each iteration
- Simulation standard errors about 0.015

GUIDE classification (Loh, 2009)
Improving on FACT, QUEST, and CRUISE

1. Use marginal and interaction χ^2 contingency table tests (as in CRUISE)
2. Use two-deep search to choose variable if split is based on interaction test (more powerful than CRUISE)
3. Allow linear splits on pairs of variables (new; useful for collinearity)
4. Use Bonferroni to control frequencies of interaction and linear splits (corrects CRUISE's propensity to split on interactions)
5. Allow kernel and nearest-neighbor node models (new; reduces tree size and yields predicted probabilities, *à la* logistic regression)
6. Treat missing values as a separate category in split selection (new; replaces imputation and surrogate splits)
7. Use CART method to prune the tree

GUIDE marginal tests for ordinal X

1. Compute the sample mean $\bar{x}$ and SD s of X in t.
2. Define $k = 3$ if $N(t) < 20J_t$; else $k = 4$. Define $b = 2s\sqrt{3}/k$.
3. Divide the range of X into k intervals with boundaries $\bar{x} - s\sqrt{3} + bj$; $j = 1, 2, \ldots, k-1$. Add one “interval” for missing values, if any.
4. Form a table with class values as rows and intervals as columns.
5. Let ν be df of the table. Compute the chi-squared statistic χ^2_ν for testing independence.
6. Convert χ^2_ν to a 1-df chi-squared (Wilson and Hilferty, 1931)

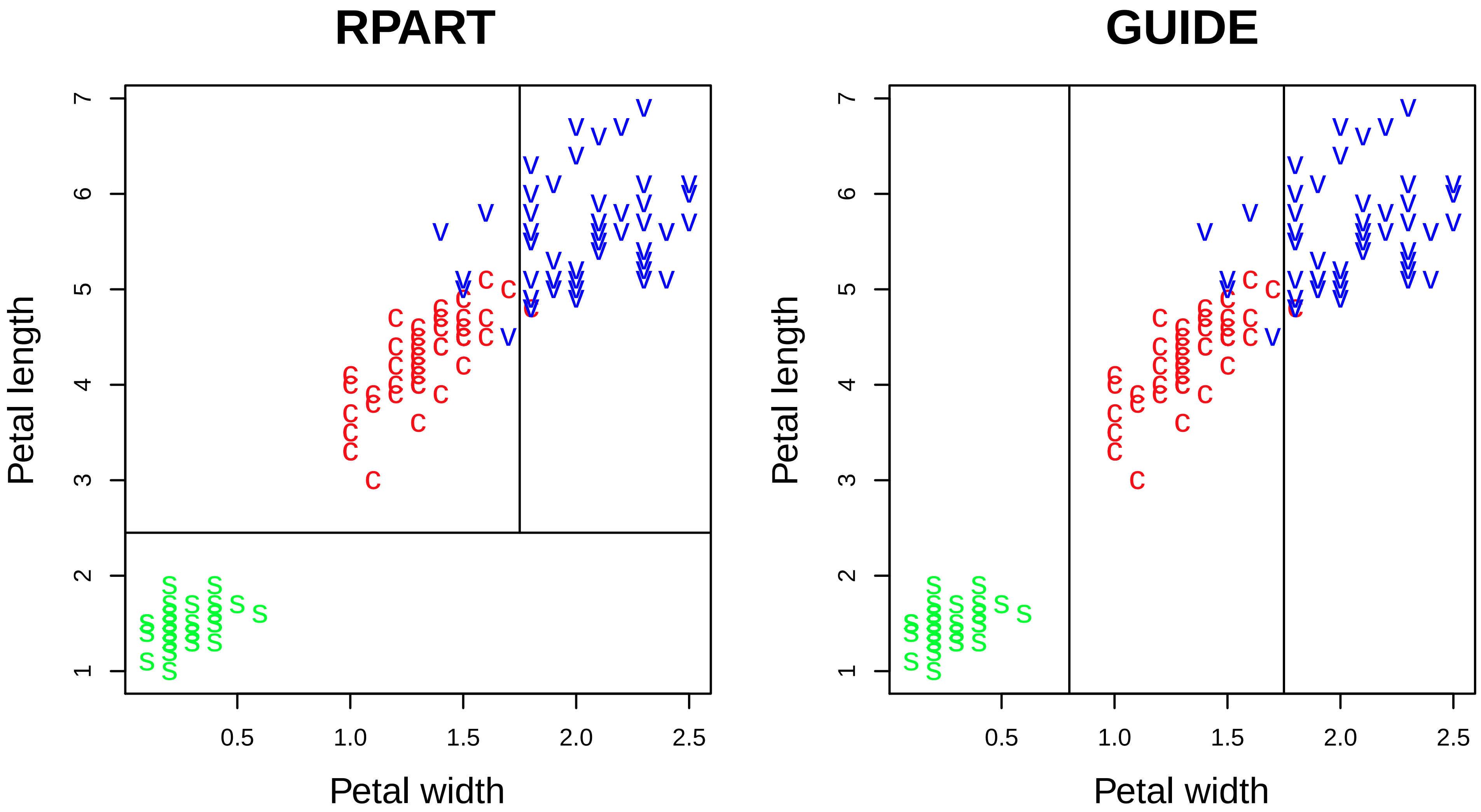
$$W_M(X) = \max\left(0, \left[\frac{7}{9} + \sqrt{\nu}\left\{\left(\frac{\chi^2_\nu}{\nu}\right)^{1/3} - 1 + \frac{2}{9\nu}\right\}\right]^3\right).$$

Note: For categorical X, use its values to form the columns of the table

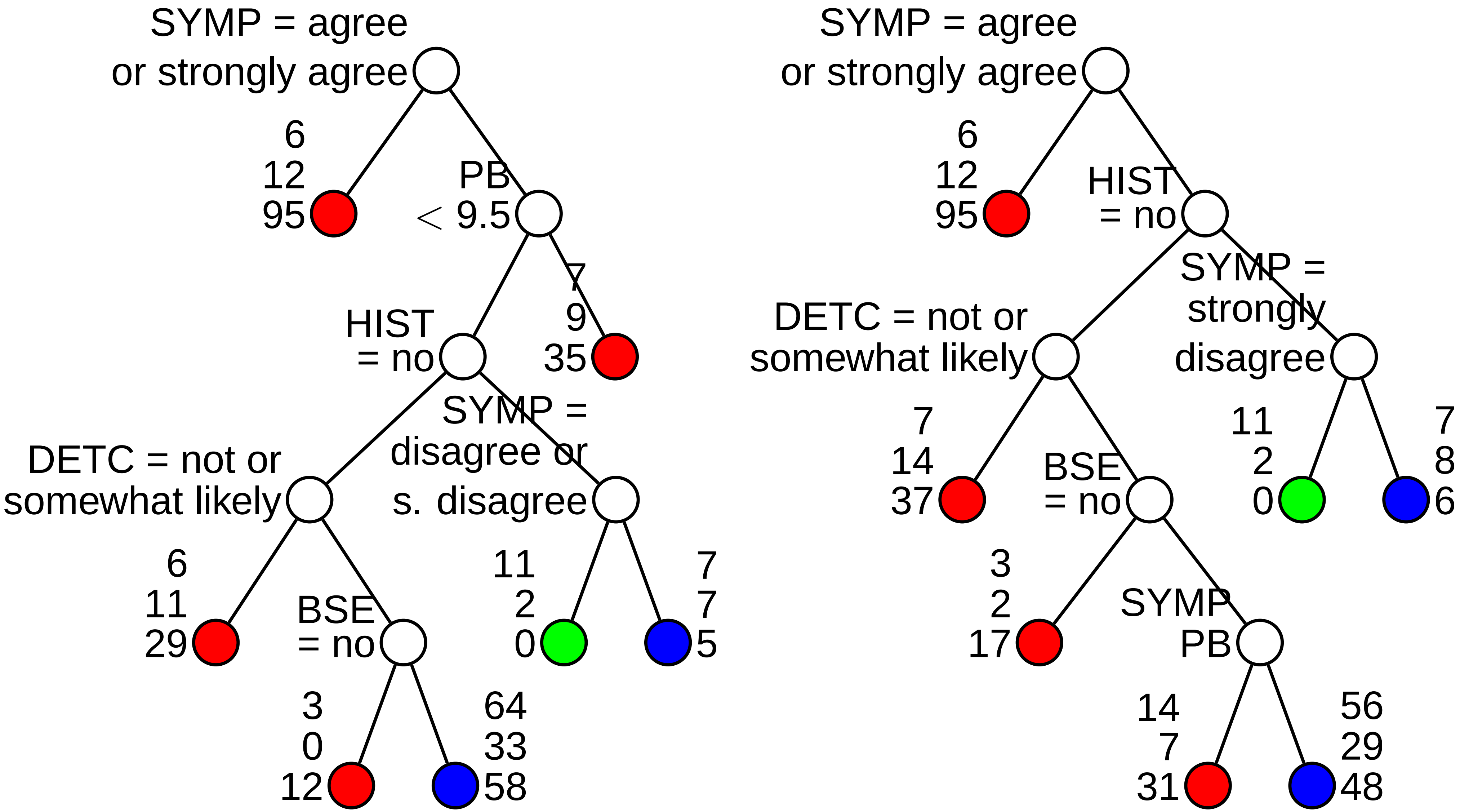
Chi-squared tests

	Petal length (χ^2_6 = 223.9)				Petal width (χ^2_6 = 226.0)			
	$\leq$2.2	(2.2, 3.7]	(3.7, 5.2]	$>$5.2	$\leq$0.5	(0.5, 1.1]	(1.1, 1.8]	$>$1.8
Setosa	50	0	0	0	49	1	0	0
Versicol	0	7	43	0	0	10	40	0
Virginica	0	0	18	32	0	0	16	34

	Sepal length (χ^2_6 = 109.2)				Sepal width (χ^2_6 = 64.6)			
	$\leq$5.1	(5.1, 5.8]	(5.8, 6.5]	$>$6.5	$\leq$2.6	(2.6, 3.0]	(3.0, 3.4]	$>$3.4
Setosa	36	14	0	0	1	7	21	21
Versicol	4	20	18	8	16	26	8	0
Virginica	1	5	22	22	7	26	16	3


RPART
GUIDE
Petal length
Petal width
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5
S
C
V

RPART (left) & GUIDE (right) trees for mammography



Within 1 year in green, more than one year in blue, never in red

Chi-squared tests

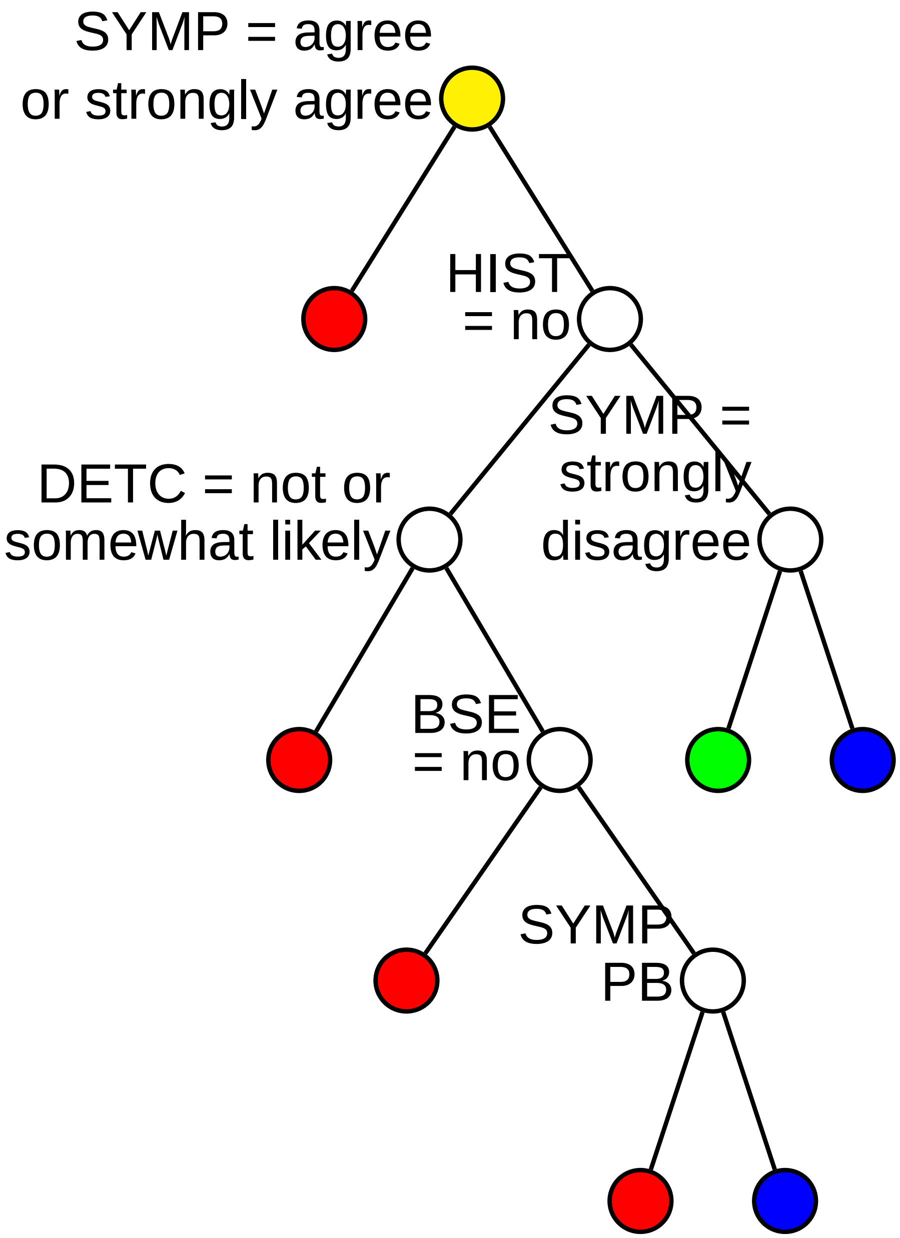
	SYMP (χ_6^2 = 57.2; $\chi_1^2 \approx$ 47)			
	strongly			strongly
ME	agree	agree	disagree	disagree
Never	33	62	85	54
1 year	2	4	43	55
$>$ 1 yr	5	7	32	30

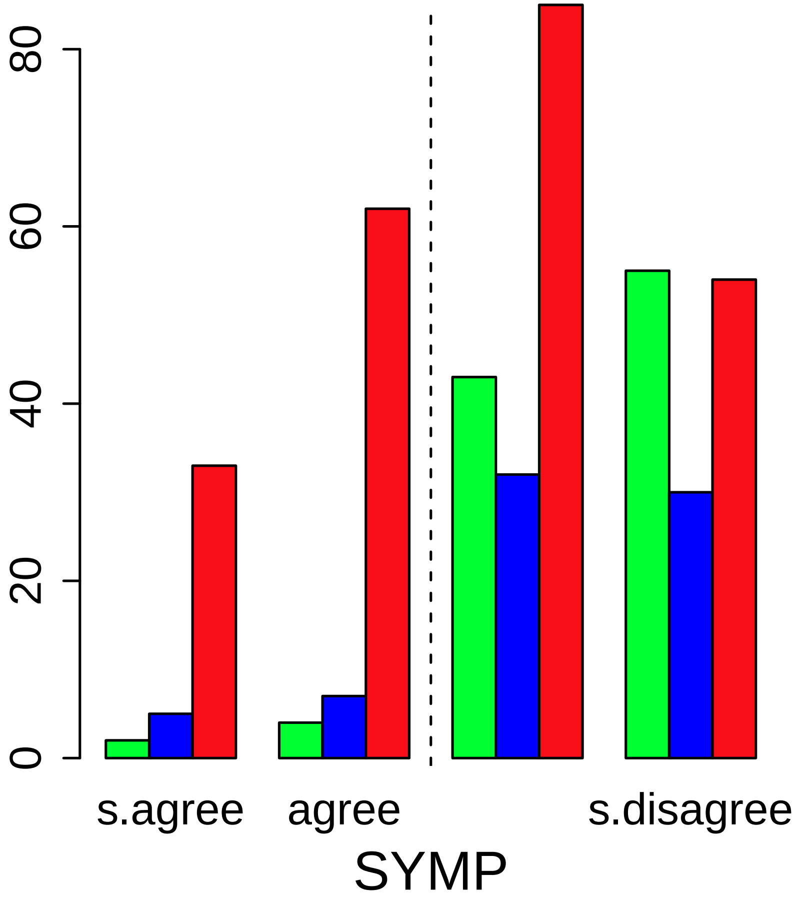
	PB (χ_6^2 = 31.3; $\chi_1^2 \approx$ 19)			
ME	$\leq$ 5.7	(5.7, 7.6]	(7.6, 9.4]	$>$ 9.4
Never	33	68	65	68
1 year	31	43	22	8
$>$ 1 yr	19	25	18	12

	DETC (χ^2_4 = 24.1; $\chi^2_1 \approx$ 16)		
	not	somewhat	very
ME	likely	likely	likely
Never	13	77	144
1 year	1	12	91
$>$ 1 yr	4	16	54

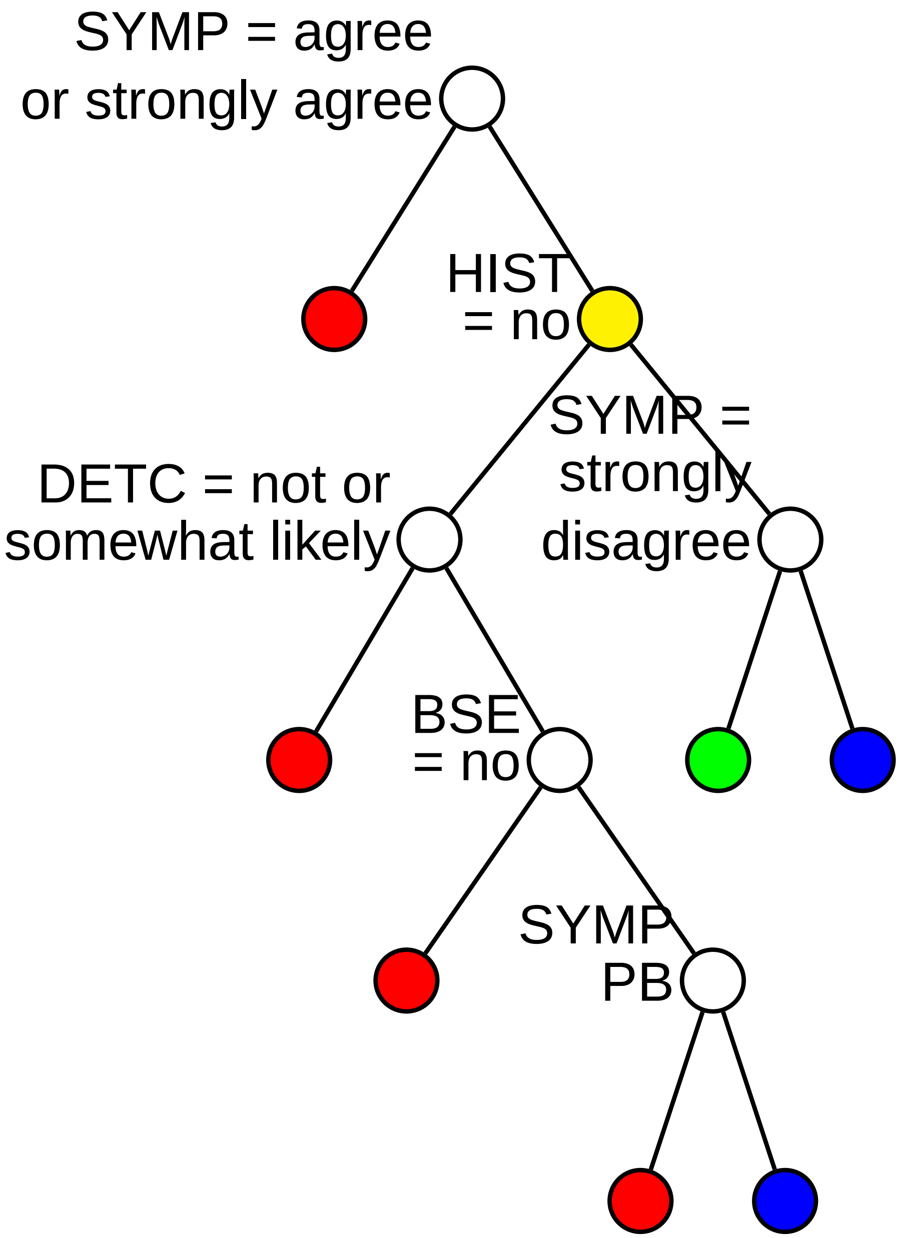
	BSE (χ^2_2 = 15.6, $\chi^2_1 \approx$ 13)		HIST (χ^2_2 = 13.1, $\chi^2_1 \approx$ 10)	
ME	no	yes	no	yes
Never	44	190	220	14
1 year	5	99	85	19
$>$ 1 yr	5	69	63	11

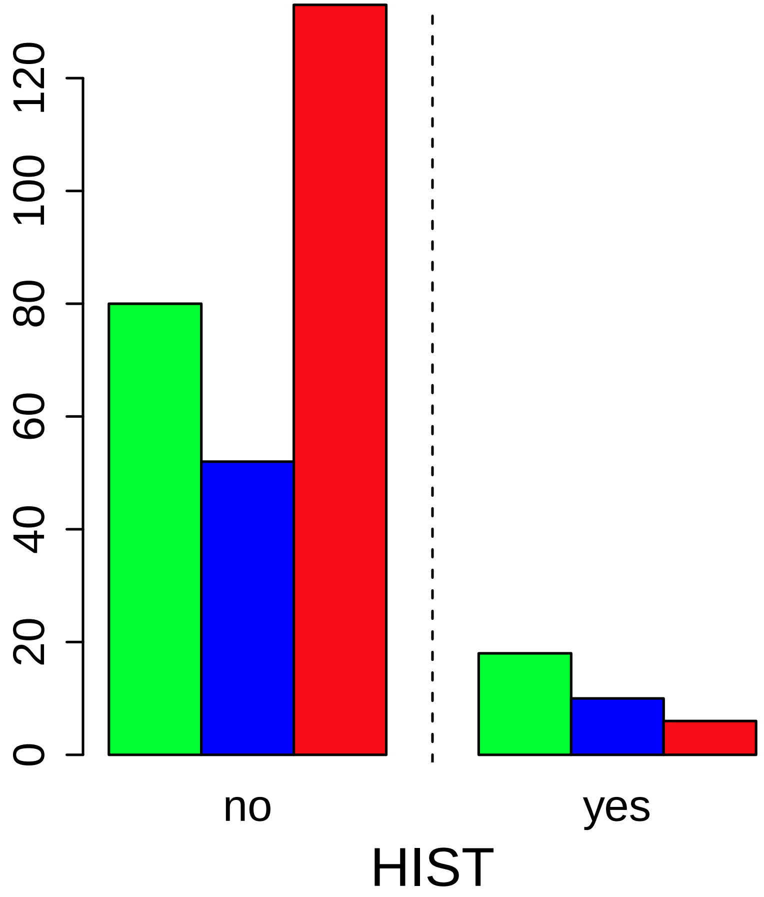
1st split


SYMP = agree
or strongly agree
HIST
= no
DETC = not or
somewhat likely
SYMP =
strongly
disagree
BSE
= no
SYMP
PB

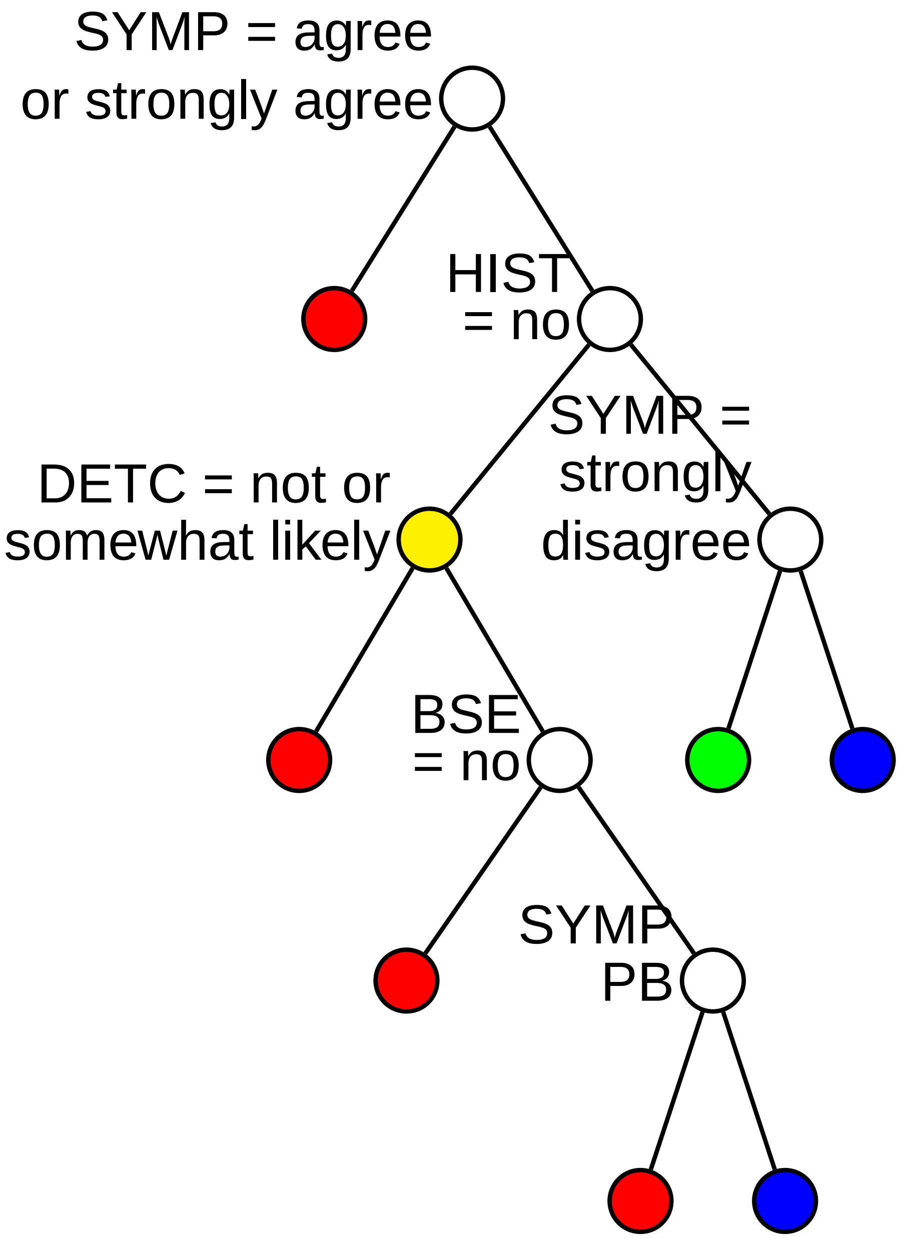

0
20
40
60
80
s.agree
agree
s.disagree
SYMP

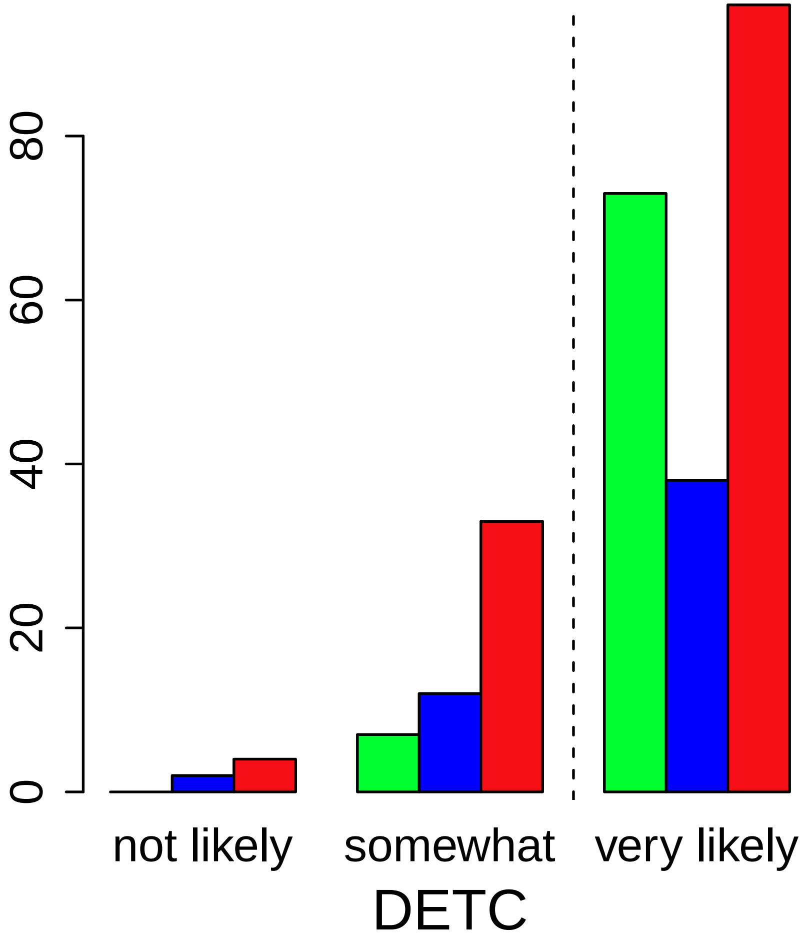
2nd split


SYMP = agree
or strongly agree
HIST
= no
DETC = not or
somewhat likely
SYMP =
strongly
disagree
BSE
= no
SYMP
PB

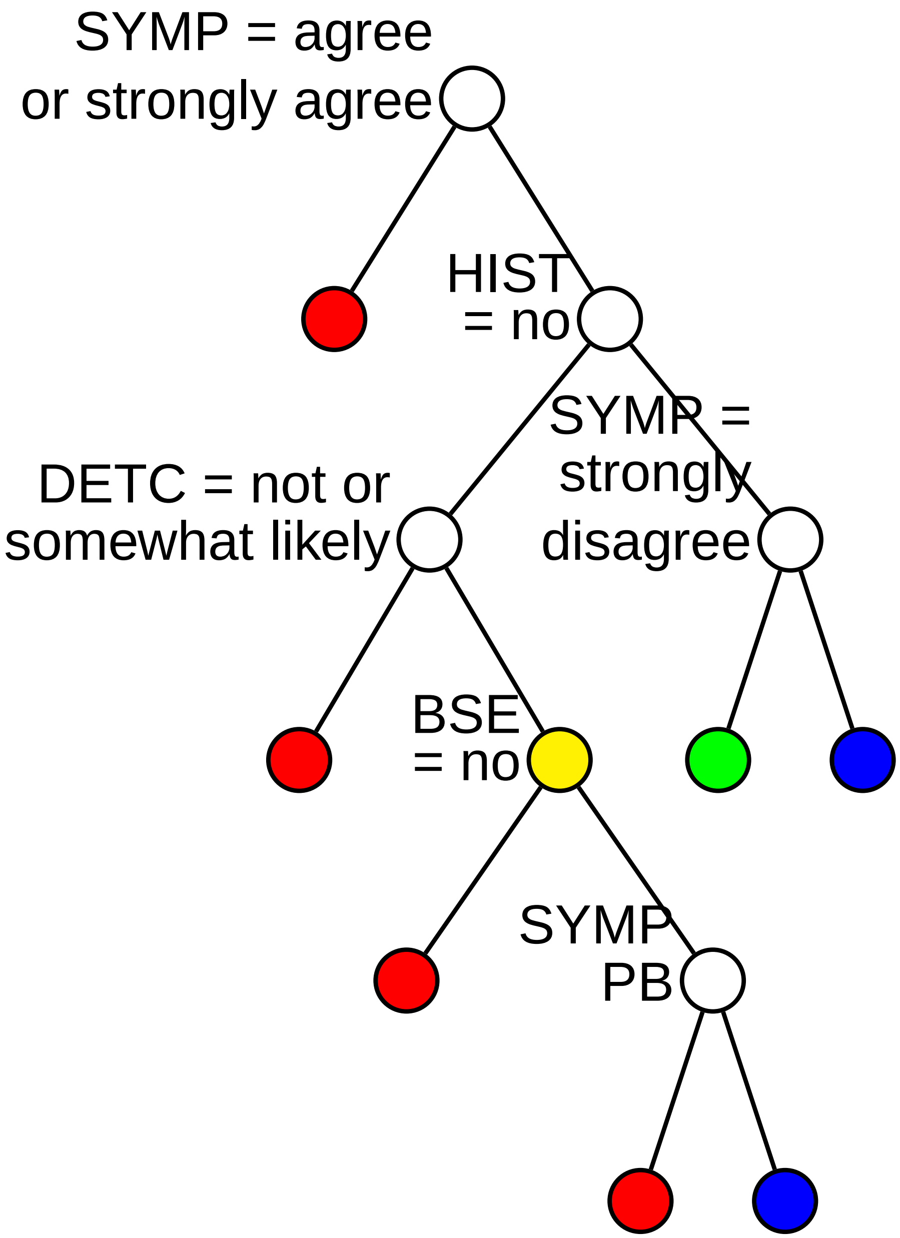

0
20
40
60
80
100
120
no
yes
HIST

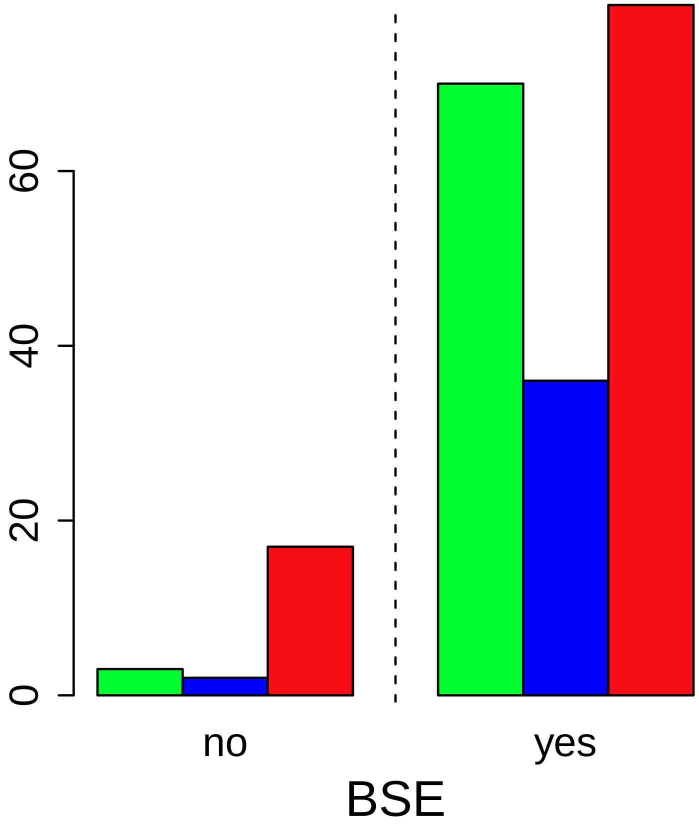
3rd split


SYMP = agree
or strongly agree
HIST
= no
DETC = not or
somewhat likely
SYMP =
strongly
disagree
BSE
= no
SYMP
PB

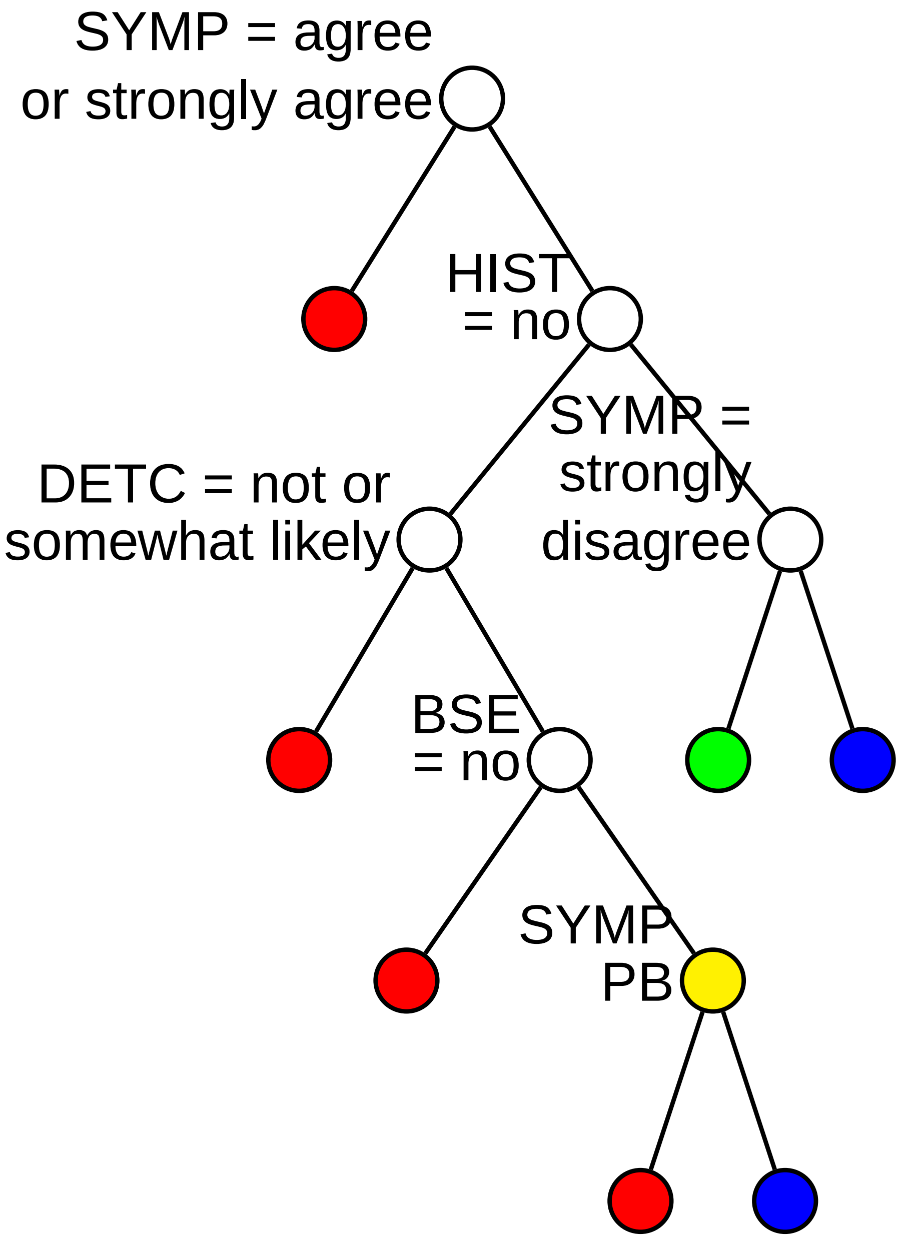

0
20
40
60
80
not likely
somewhat
very likely
DETC

4th split


SYMP = agree
or strongly agree
HIST
= no
DETC = not or
somewhat likely
SYMP =
strongly
disagree
BSE
= no
SYMP
PB


0
20
40
60
no
yes
BSE

5th split



SYMP = agree or strongly agree

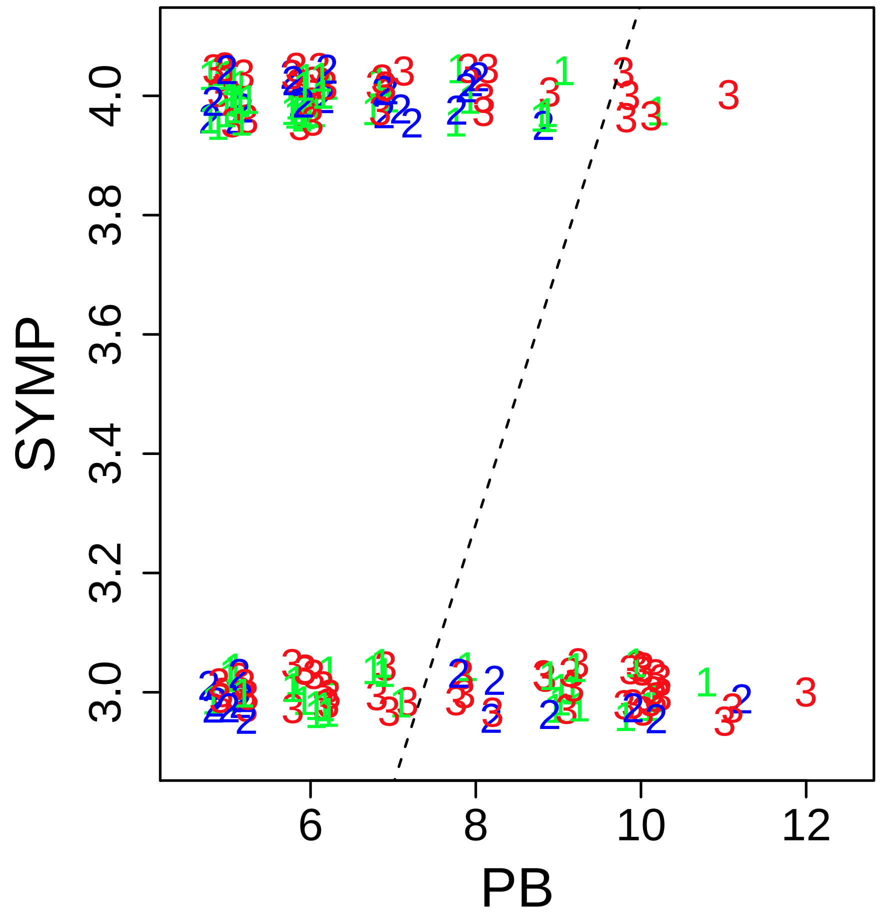
HIST = no

DETC = not or somewhat likely

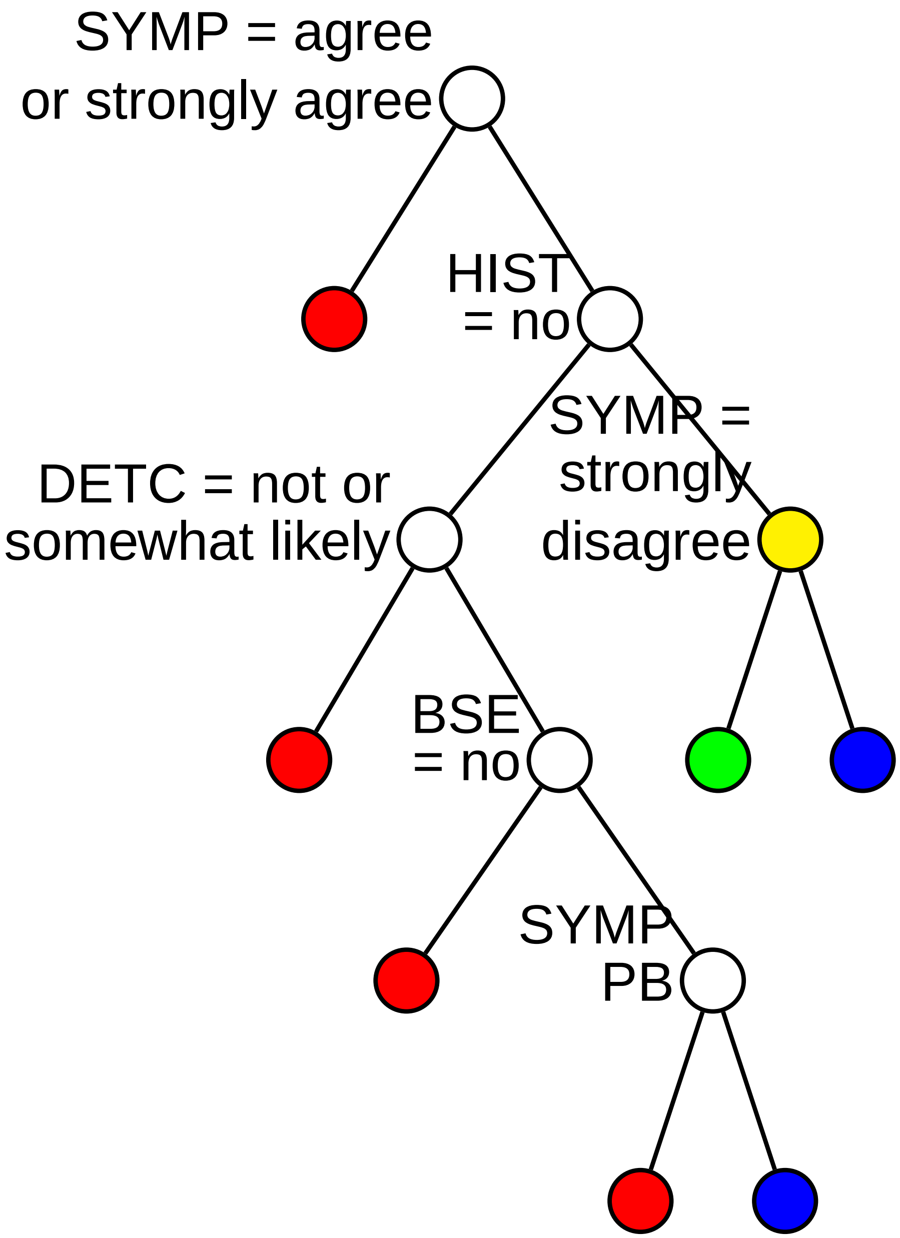
SYMP = strongly disagree

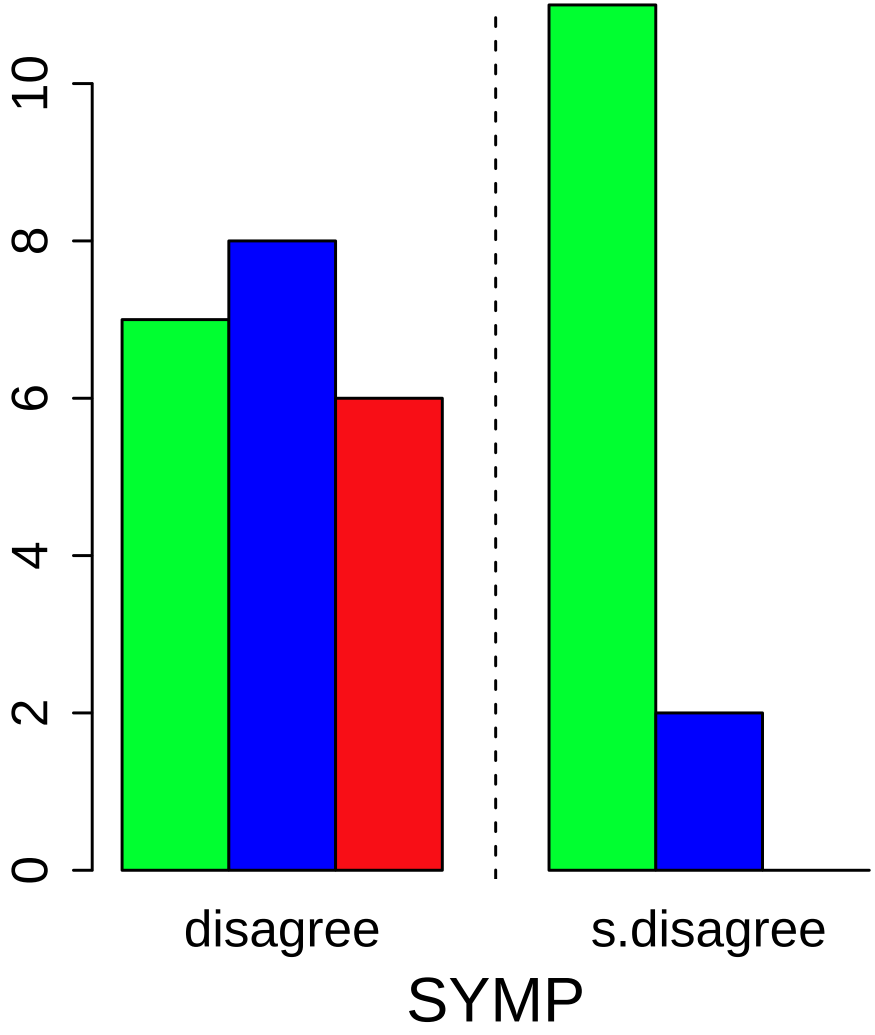
BSE = no

SYMP PB

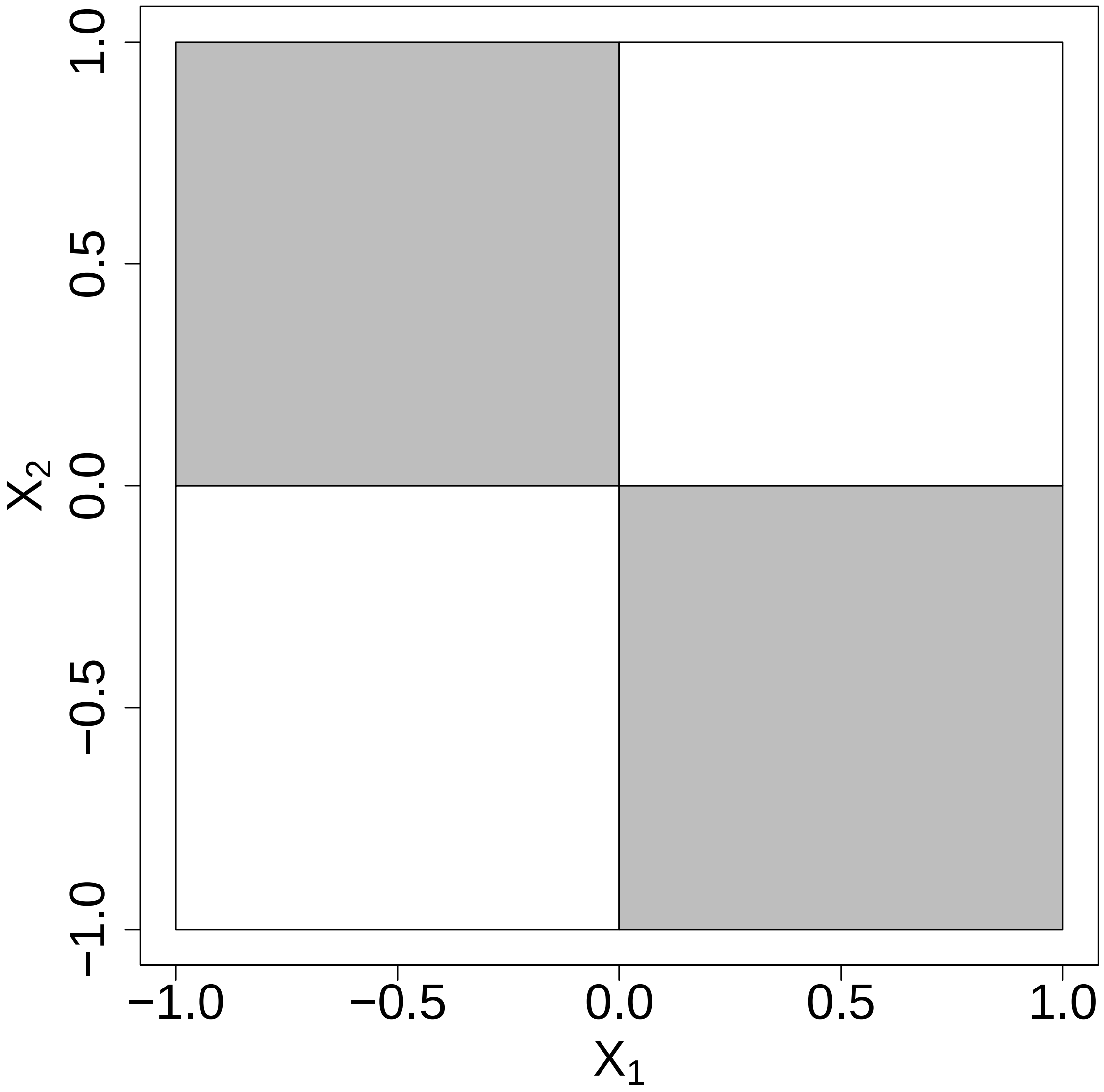


6th split


SYMP = agree
or strongly agree
HIST
= no
DETC = not or
somewhat likely
SYMP =
strongly
disagree
BSE
= no
SYMP
PB


0
2
4
6
8
10
disagree
s.disagree
SYMP

Two-class problem with interaction



GUIDE split variable selection: interaction tests for X_1, X_2

1. Divide the (X_1, X_2)-space into sets

$$B_{k,m} = \{(x_1, x_2) : x_1 \in A_{1k}, x_2 \in A_{2m}\}, \quad k, m = 1, 2, \ldots$$

   where A_{1k} and A_{2m} are the respective intervals or categories

2. Form a contingency table with class labels as rows and $\{B_{k,m}\}$ as columns

3. Compute chi-squared statistic and use Wilson-Hilferty approximation to convert it to a 1-df chi-squared value $W_I(X_1, X_2)$

SYMP-BSE interaction test

	SYMP							
	strongly agree		agree		strongly disagree		disagree	
	BSE		BSE		BSE		BSE	
ME	no	yes	no	yes	no	yes	no	yes
0	6	27	15	47	15	70	8	46
1	1	1	0	4	0	43	4	51
2	1	4	0	7	2	30	2	28

χ^2_{14} = 72, χ^2_1 = 45, p = 9×10^{-10}

GUIDE split variable selection

1. Let K be the number of non-constant predictor variables in node t.
2. Let $\chi^2_{\nu,\alpha}$ be the upper-α quantile of the chi-squared distribution with ν df and define
$$\alpha = \frac{0.05}{K}, \quad \beta = \frac{0.1}{K(K-1)}$$
3. Find $W_M(X_i)$ for each X_i.
4. (a) If $\max_i W_M(X_i) > \chi^2_{1,\alpha}$, select the variable with the largest $W_M(X_i)$.
   (b) Otherwise, find $W_I(X_i, X_j)$ for each pair of predictor variables.
      i. If $\max_{i \neq j} W_I(X_i, X_j) > \chi^2_{1,\beta}$, select pair with largest $W_I(X_i, X_j)$.
      ii. Otherwise, select variable with largest $W_M(X_i)$.

Split set selection for ordinal X

Search all splits of the form $X \leq c$ to minimize misclassification cost

Split set selection for categorical X

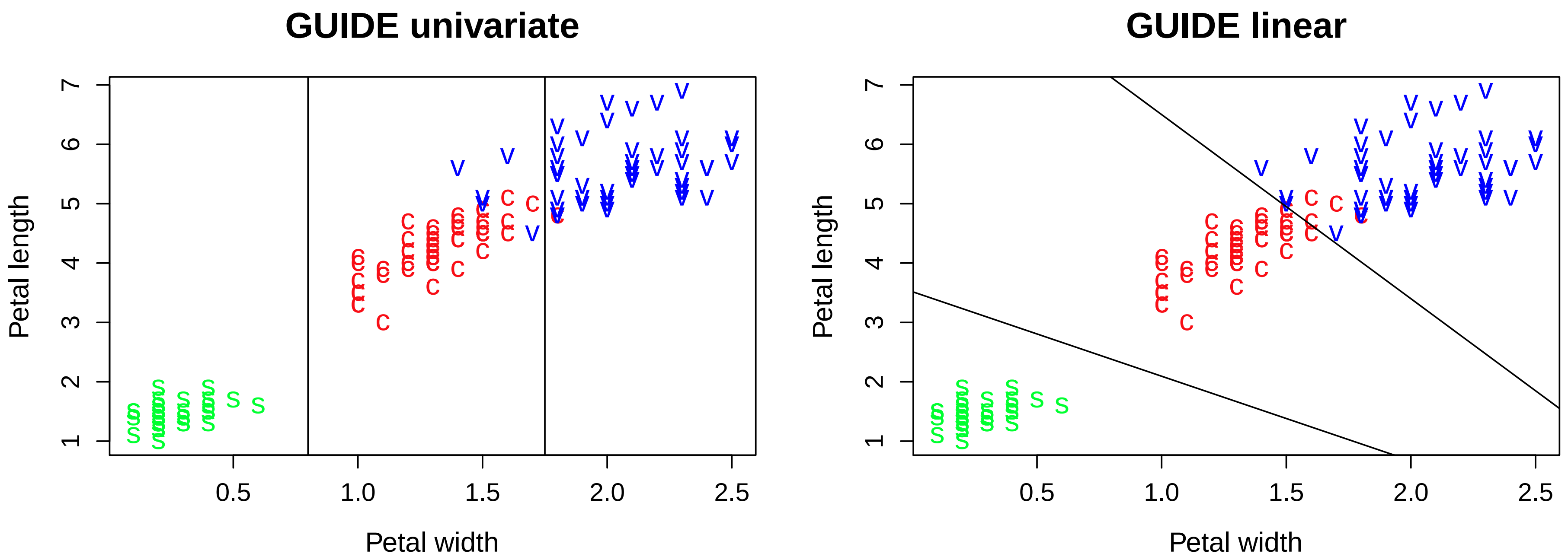
Suppose X takes distinct values $\{a_1, a_2, \ldots, a_n\}$ in node t

1. If $J = 2$ or $n \leq 11$, search all subsets S to find $t_L = \{X \in S\}$
2. If $J \leq 11$ and $n > 20$, let class j_i minimize misclassification cost in $t \cap \{X = a_i\}$
   (a) Define $X' = \sum_i j_i \, I(X = a_i)$ and search for the split based on X' that minimizes the decrease in impurity
   (b) Express the split as $t_L = \{X \in S\}$
3. Otherwise, use linear discriminant analysis on the dummy vectors

Bivariate linear discriminant split

Let X_1 and X_2 be two s variables in node t

1. For the jth class and each X_i, find node class mean $\bar{x}_{i,j}$ and SD $s_{i,j}$
2. Find trimmed set S_j of class j samples in t such that $|X_i - \bar{x}_{i,j}| \leq 2.5 s_{i,j}$
3. Find the larger linear discrim. coord. Z from observations in $S_1 \cup \ldots \cup S_J$
4. Project the data in t onto the Z-axis
5. Compute the Wilson-Hilferty 1-df chi-squared $W_L(X_1, X_2)$ from the Z values



Summary of GUIDE classification split selection

Let K be the number of non-constant predictor variables and let K_1 ($< K$) be the number that are ordinal. Define

$$\alpha = 0.05/K, \quad \beta = 0.1/\{K(K-1)\}, \quad \gamma = 0.1/\{K_1(K_1-1)\}$$

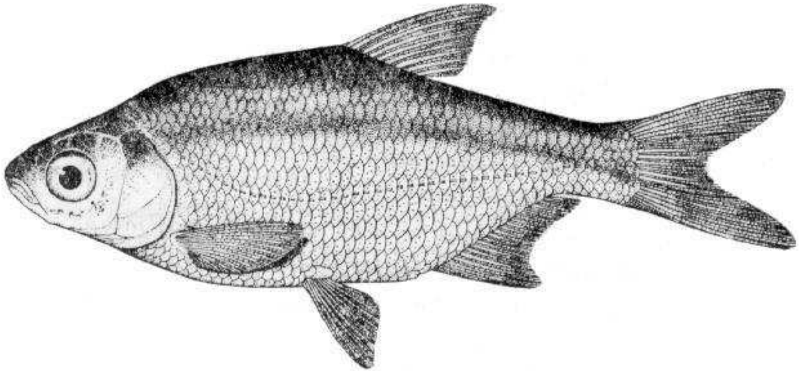
1. Split on the X_i with the largest marginal χ^2, if it is significant at level α
2. Otherwise,
   (a) If (X_i, X_j) has largest interaction χ^2 and is significant at level β, use a two-deep search to find the best split on X_i or X_j
   (b) Otherwise, compute linear discriminant χ^2 for all ordinal pairs
      i. Use most significant linear split if it is significant at level γ
      ii. Otherwise, choose the X_i with largest marginal χ^2

Fish classification

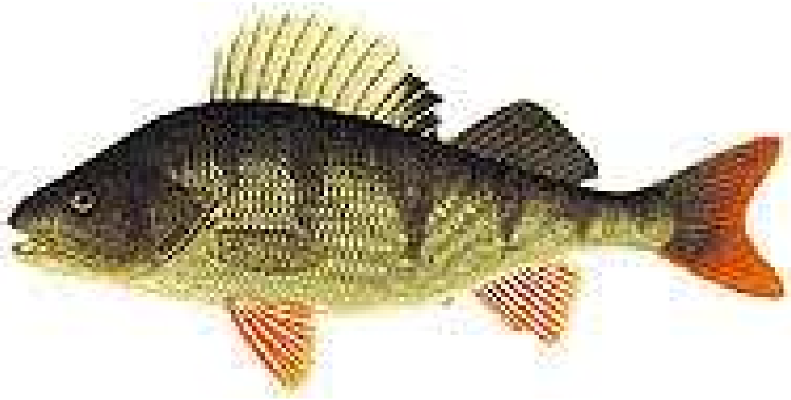
- 159 fish caught from the same lake near Tampere, Finland
- The fish are from 7 species: (1) 35 Bream, (2) 11 Parkki, (3) 56 Perch, (4) 17 Pike, (5) 20 Roach, (6) 14 Smelt, (7) 6 Whitefish

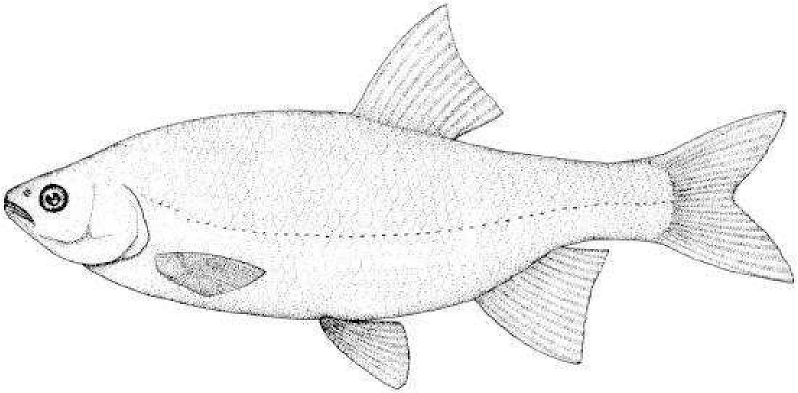
Predictor	Definition
Weight	Weight of the fish (in grams); one missing value
Length1	Length from the nose to the beginning of the tail (in cm)
Length2	Length from the nose to the notch of the tail (in cm)
Length3	Length from the nose to the end of the tail (in cm)
Height	Maximal height as % of Length3
Width	Maximal width as % of Length3
Sex	female, male, unknown

Bream (left) and Parkki (right)

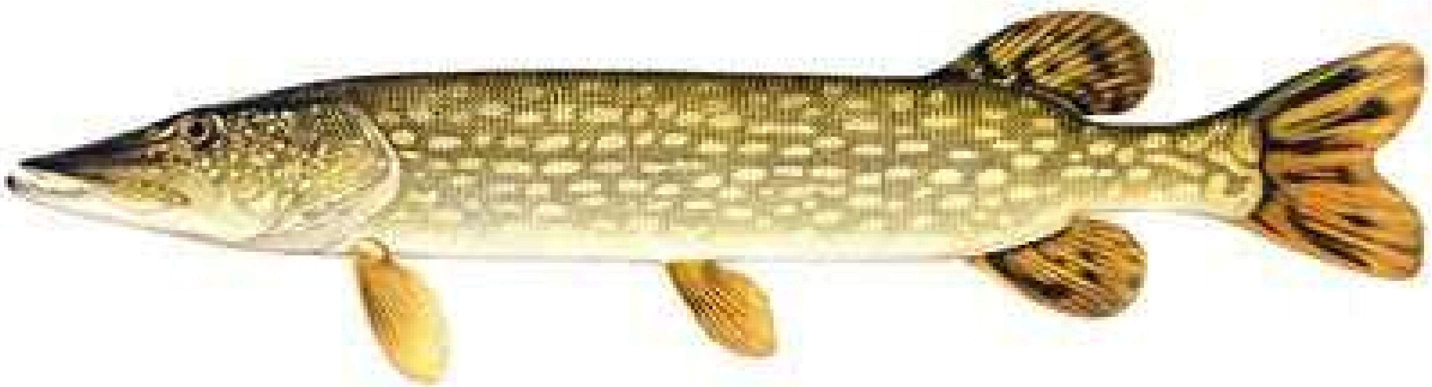


Perch (left) and Whitefish (right)

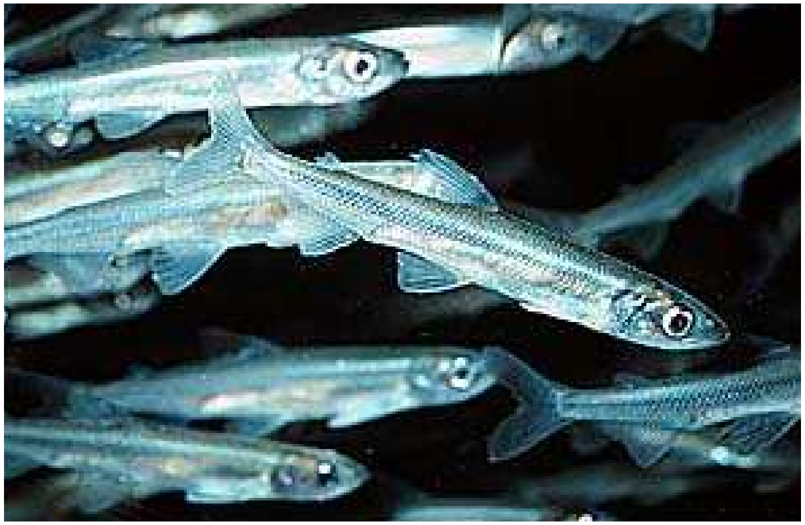




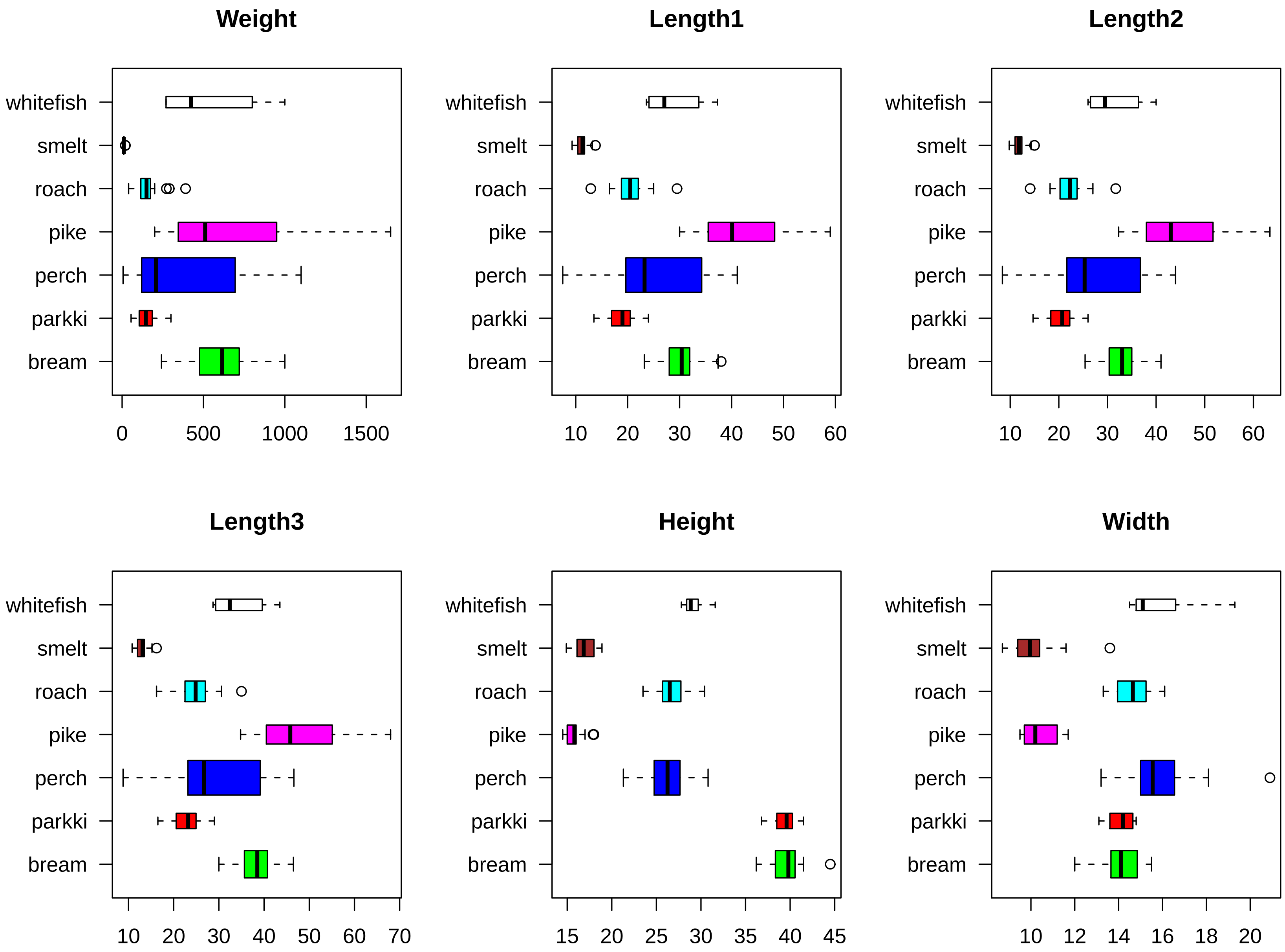
Pike



Roach (left) and Smelt (right)



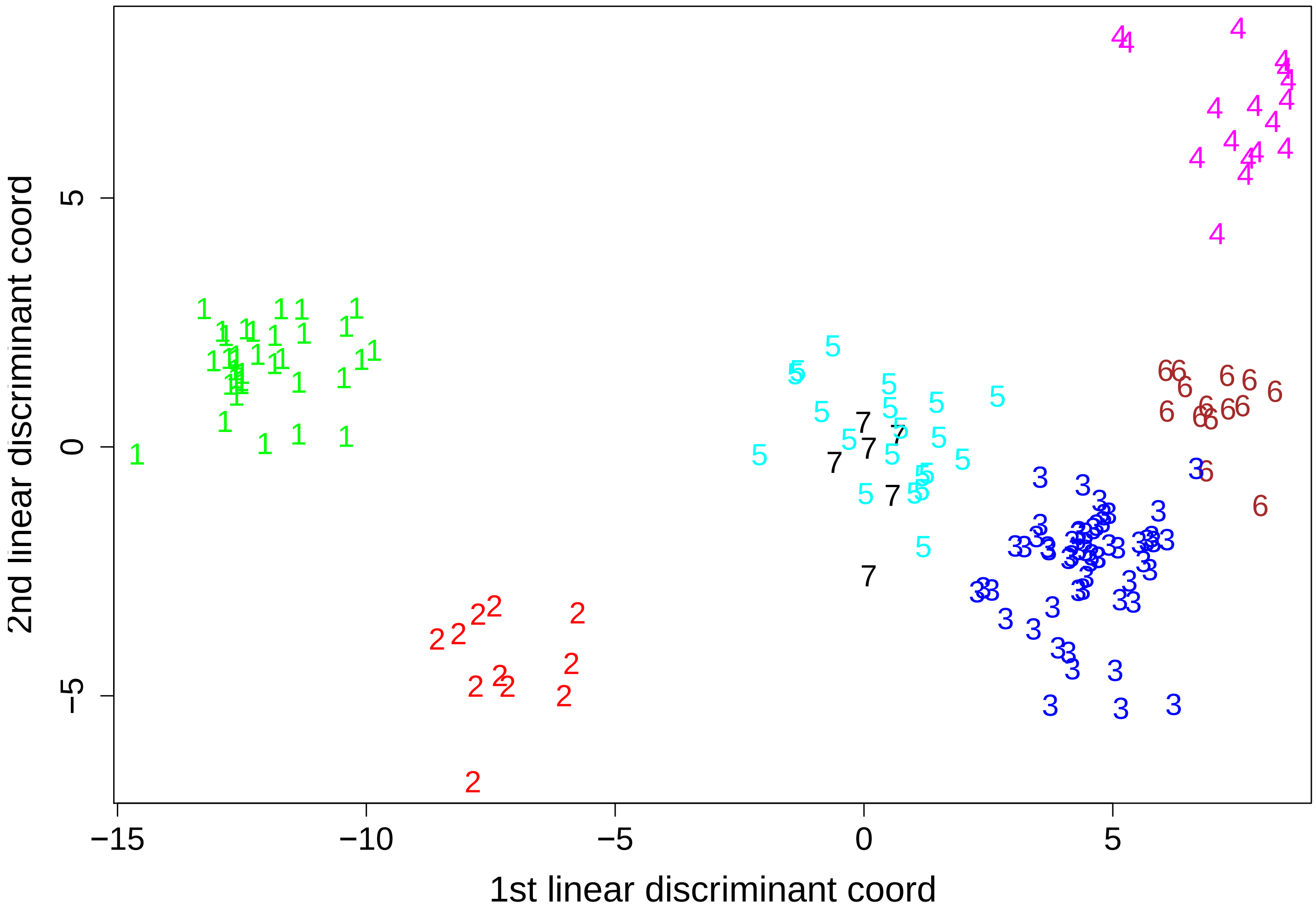
Boxplots of continuous variables



Sex by species

	Species							
Sex	Bream	Parkki	Perch	Pike	Roach	Smelt	White	Total
female	3	4	25	5	8	9	1	55
male	6	3	2	1	0	5	0	17
unknown	26	4	29	11	12	0	5	87
Total	35	11	56	17	20	14	6	159

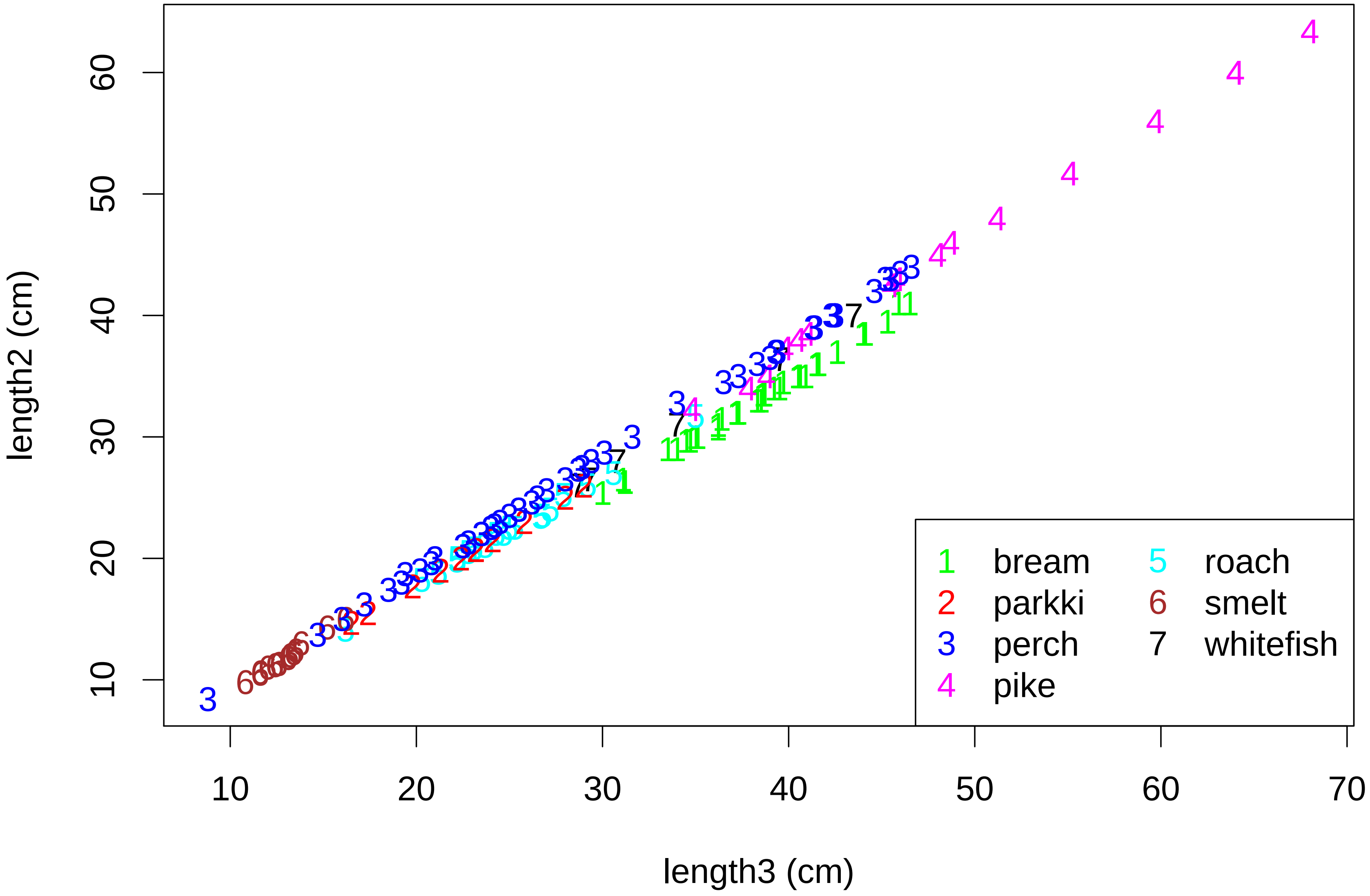
Linear discriminant analysis



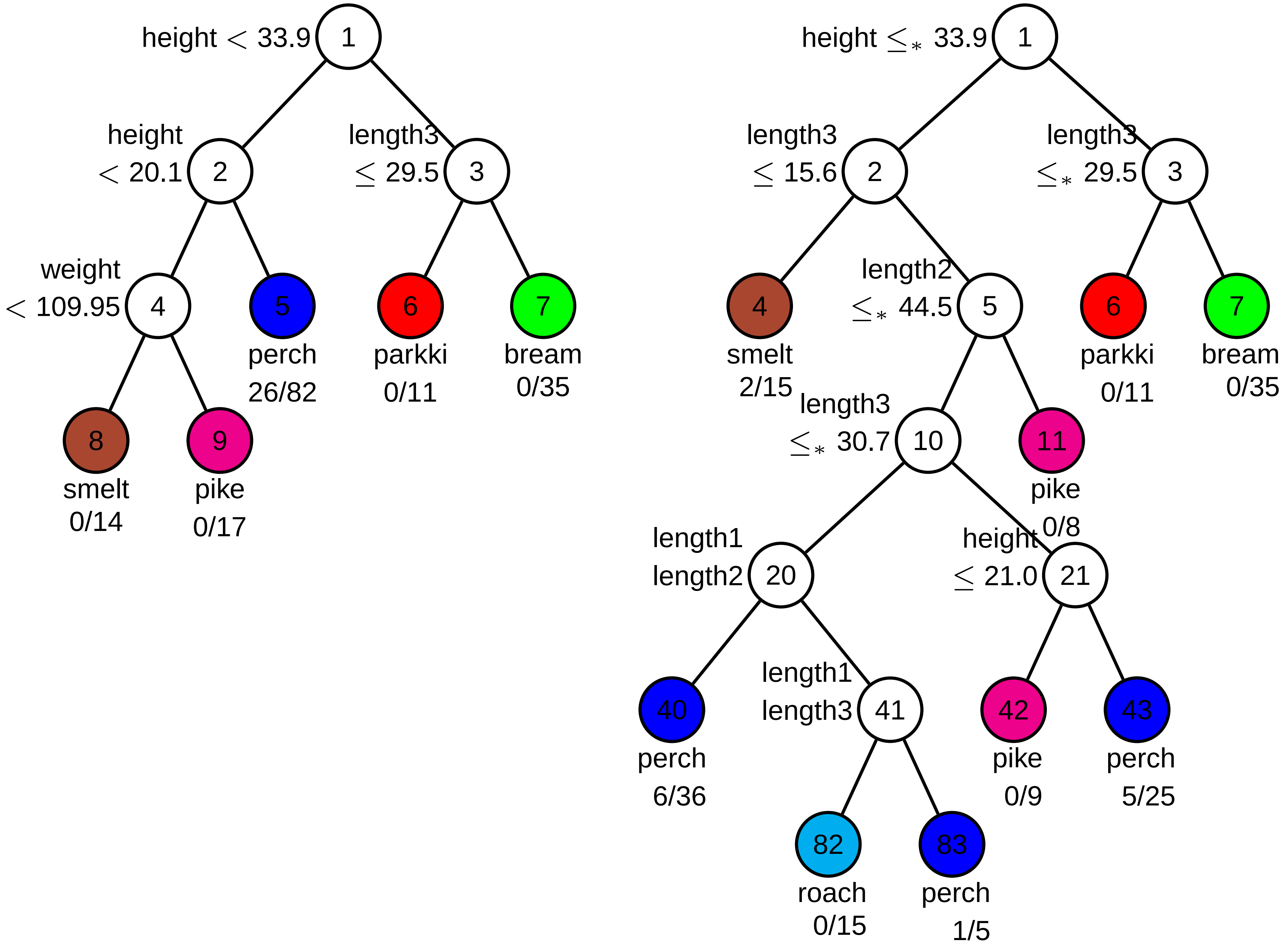
1 = Bream, 2 = Parkki, 3 = Perch, 4 = Pike, 5 = Roach, 6 = Smelt, 7 = Whitefish

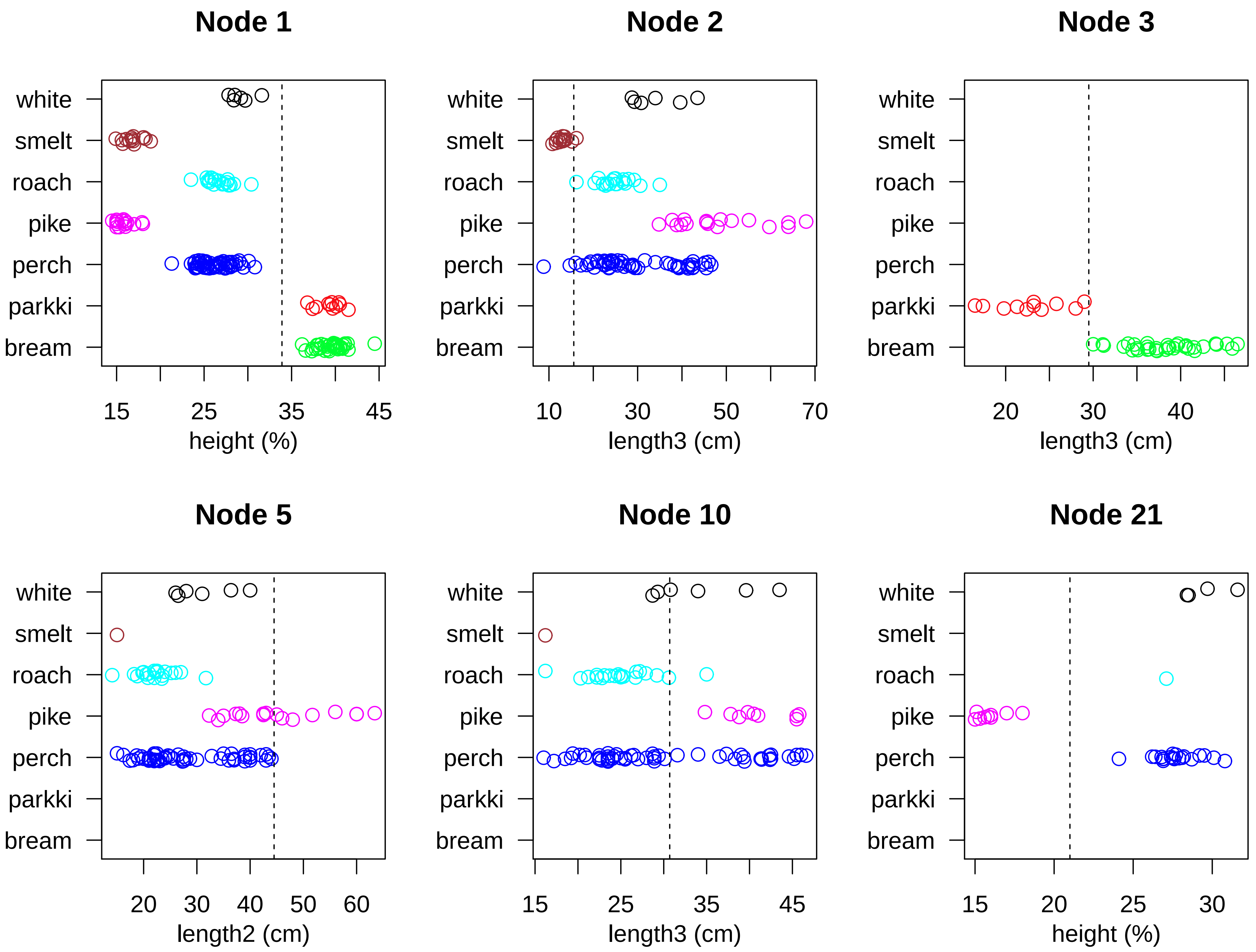
With Sex: 0/71 errors. Without Sex: 1/158 errors

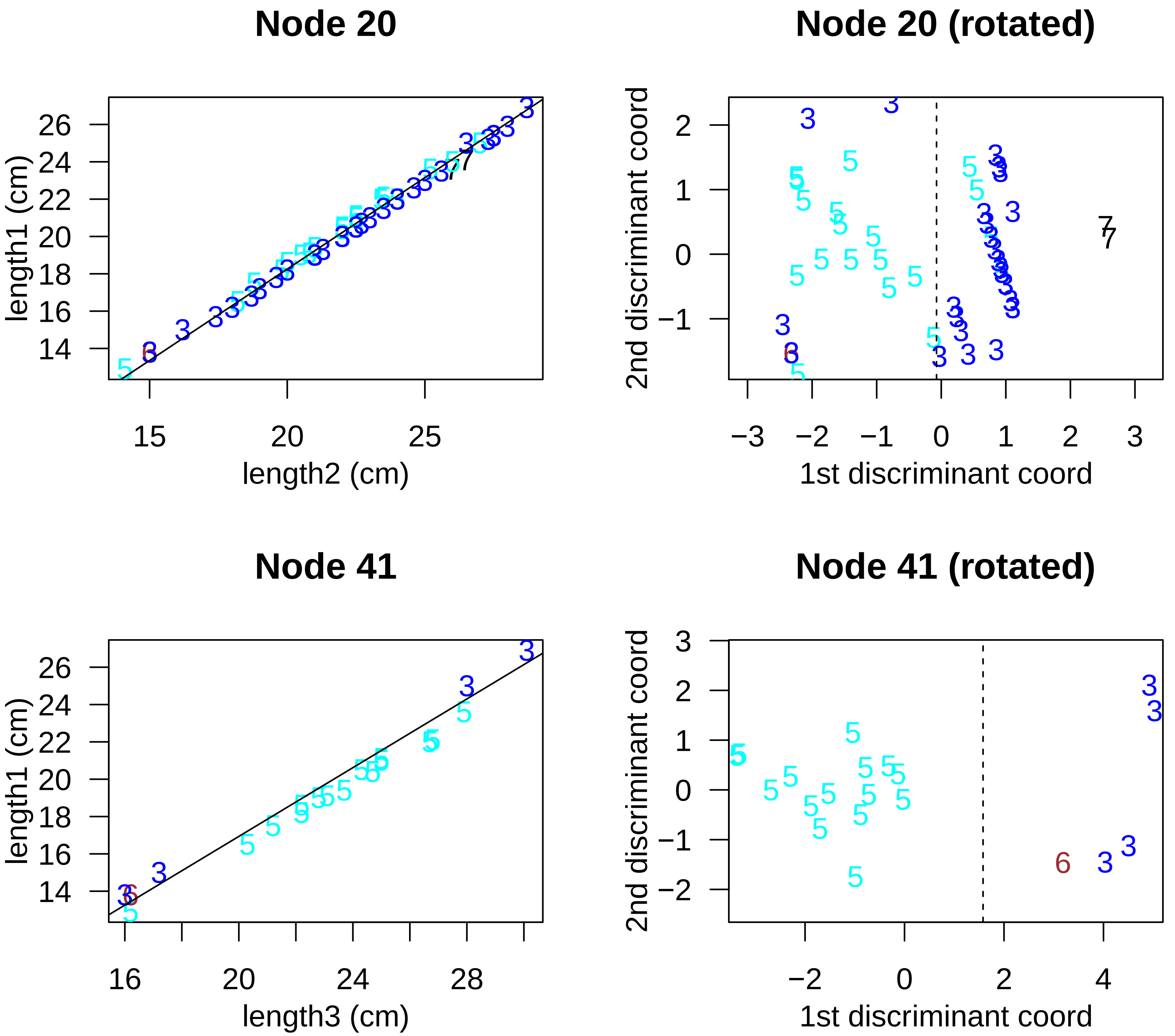
Plot of Length2 vs. Length3



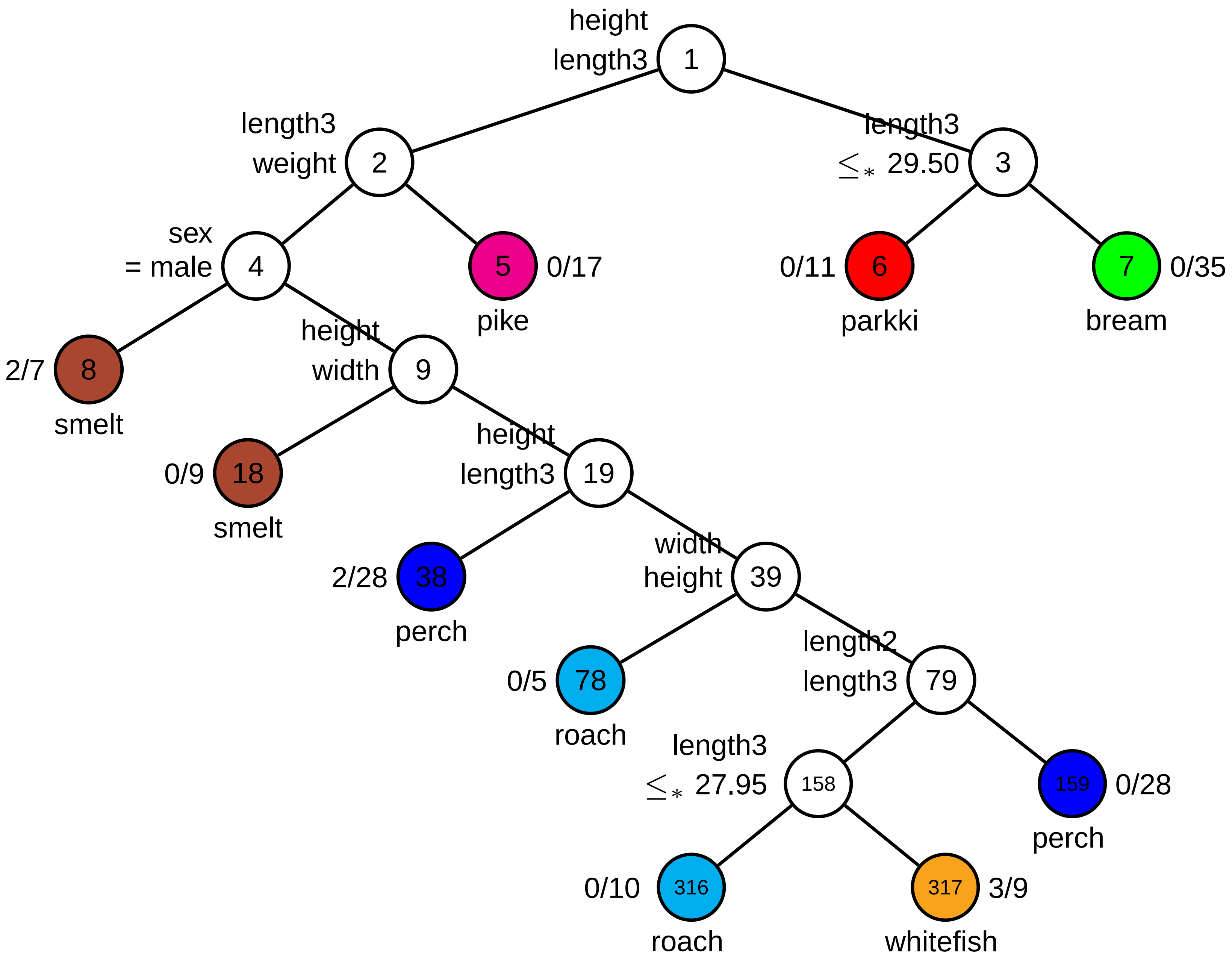
RPART (26 errors) and GUIDE (14 errors)

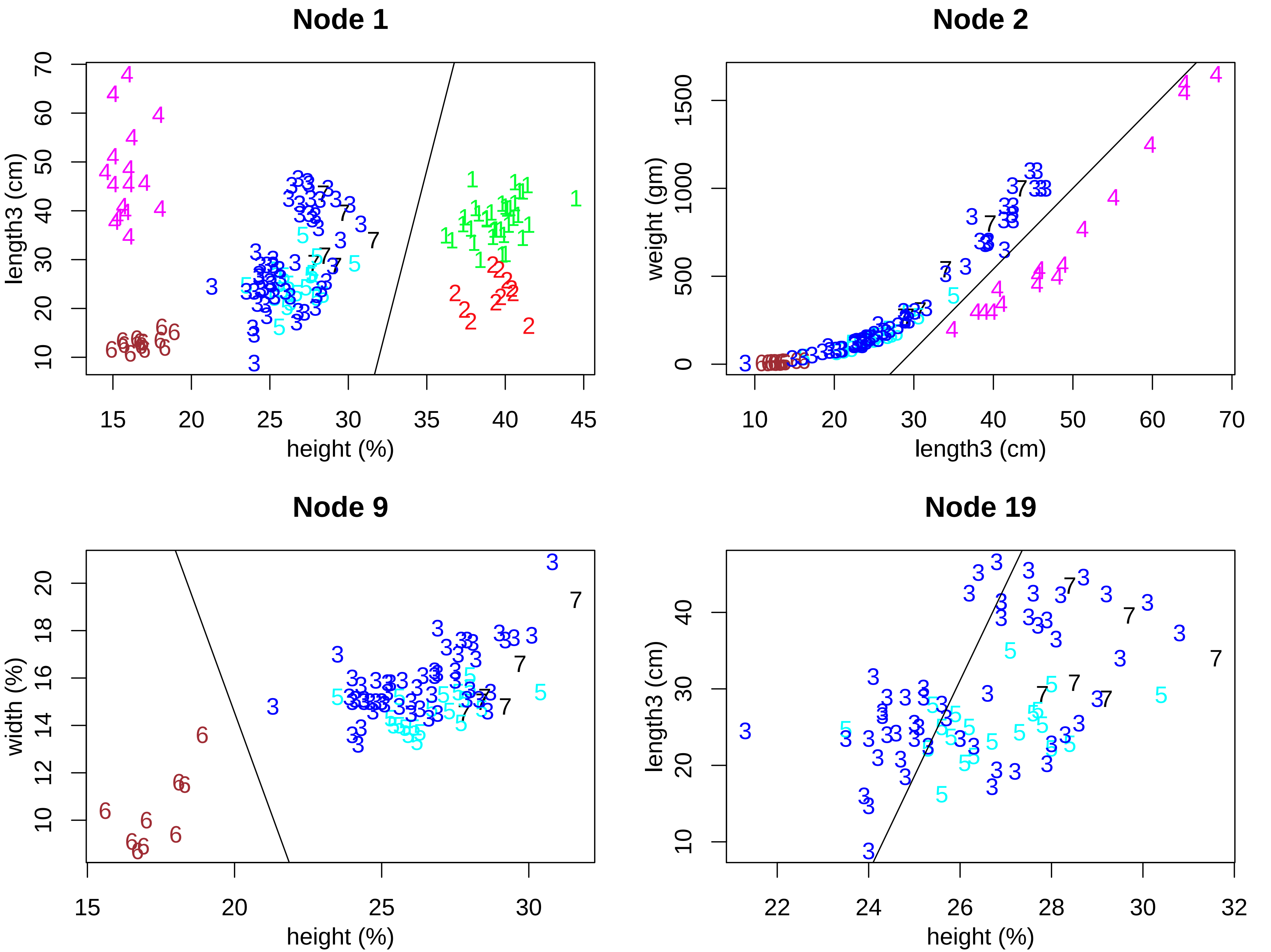


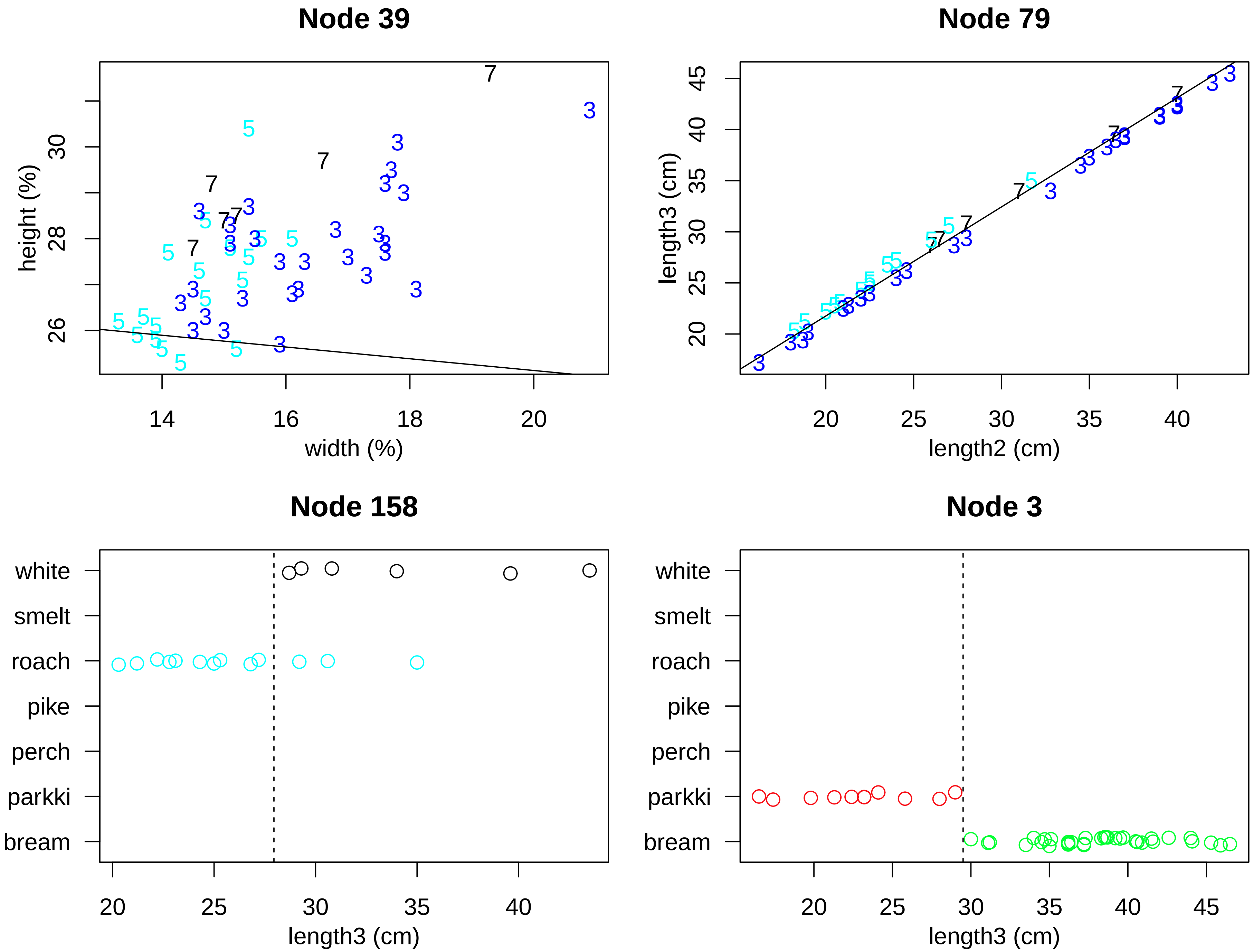

Node 1
white
smelt
roach
pike
perch
parkki
bream
15
25
35
45
height (%)
Node 2
white
smelt
roach
pike
perch
parkki
bream
10
30
50
70
length3 (cm)
Node 3
white
smelt
roach
pike
perch
parkki
bream
20
30
40
length3 (cm)
Node 5
white
smelt
roach
pike
perch
parkki
bream
20
30
40
50
60
length2 (cm)
Node 10
white
smelt
roach
pike
perch
parkki
bream
15
25
35
45
length3 (cm)
Node 21
white
smelt
roach
pike
perch
parkki
bream
15
20
25
30
height (%)


Node 20
length1 (cm)
14
16
18
20
22
24
26
15
20
25
length2 (cm)
Node 20 (rotated)
2nd discriminant coord
−1
0
1
2
−3
−2
−1
0
1
2
3
1st discriminant coord
Node 41
length1 (cm)
14
16
18
20
22
24
26
16
20
24
28
length3 (cm)
Node 41 (rotated)
2nd discriminant coord
−2
−1
0
1
2
3
−2
0
2
4
1st discriminant coord

Fish data with linear priority splits (7 errors)




Node 1
length3 (cm)
10
20
30
40
50
60
70
height (%)
15
20
25
30
35
40
45
Node 2
weight (gm)
0
500
1000
1500
length3 (cm)
10
20
30
40
50
60
70
Node 9
width (%)
10
12
14
16
18
20
height (%)
15
20
25
30
Node 19
length3 (cm)
10
20
30
40
height (%)
22
24
26
28
30
32


Node 39
height (%)
26
28
30
14
16
18
20
width (%)
Node 79
length3 (cm)
20
25
30
35
40
45
20
25
30
35
40
length2 (cm)
Node 158
white
smelt
roach
pike
perch
parkki
bream
20
25
30
35
40
length3 (cm)
Node 3
white
smelt
roach
pike
perch
parkki
bream
20
25
30
35
40
45
length3 (cm)

Importance ranking of variables

Importance score of X_i is

$$\text{IMP}(i) = \sum_t n(t) W_M(t,i)$$

- $W_M(t,i)$ is the Wilson-Hilferty marginal chi-squared value of X_i at t
- $n(t)$ is the training sample size at node t
- sum is over all intermediate nodes t

If X_i is constant at t, set $W_M(t,i) = 1$

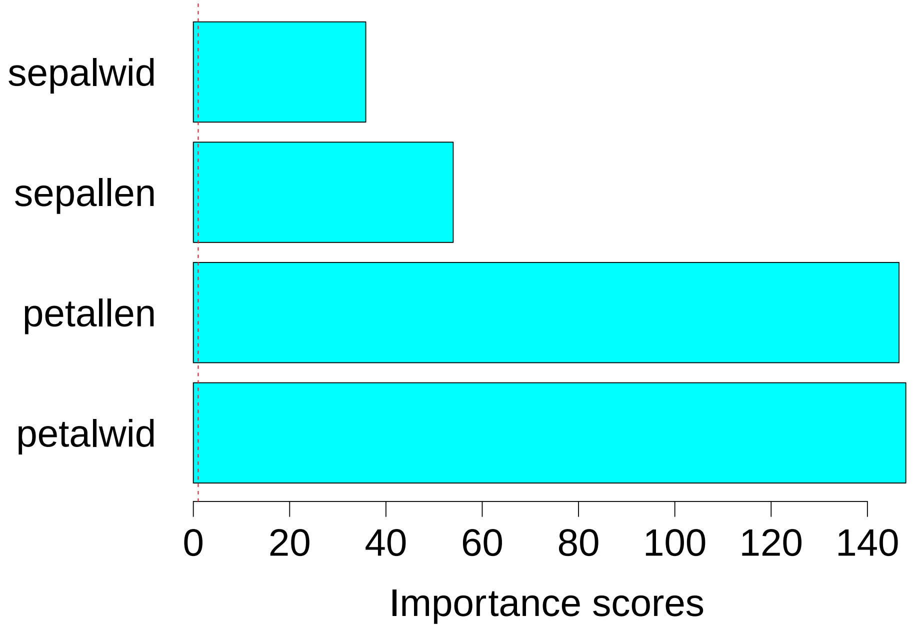
Null distribution of importance scores

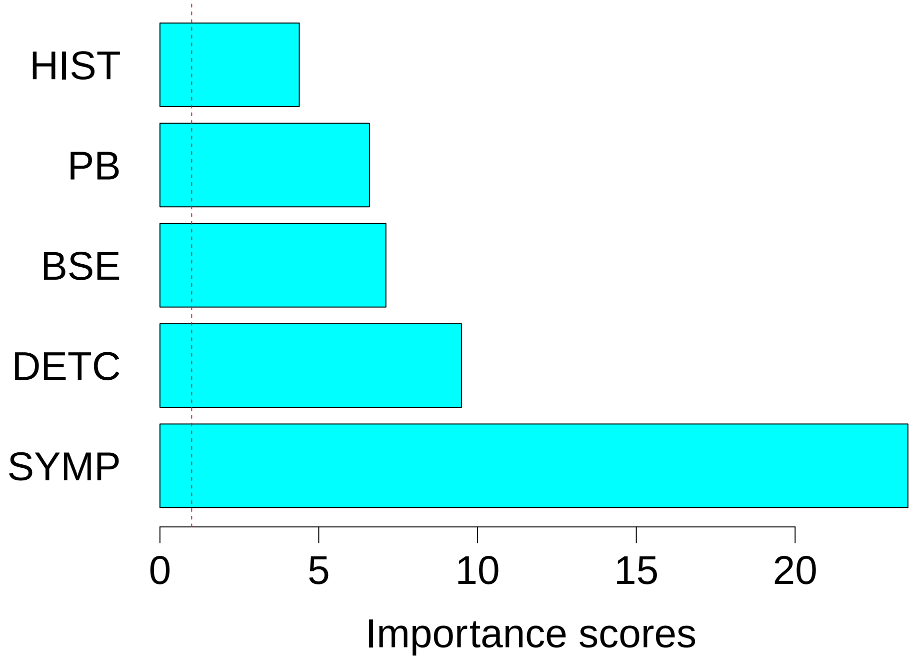
- If X_i is independent of Y, then
  - IMP(i) is a linear combination of independent chi-squared variables
  - Use Satterthwaite (1946) method to approximate distribution of IMP(i)
- Cut-off score for separating important from unimportant variables is the upper-α quantile of the corresponding chi-squared distribution, where

  $$\alpha = k_0/K$$

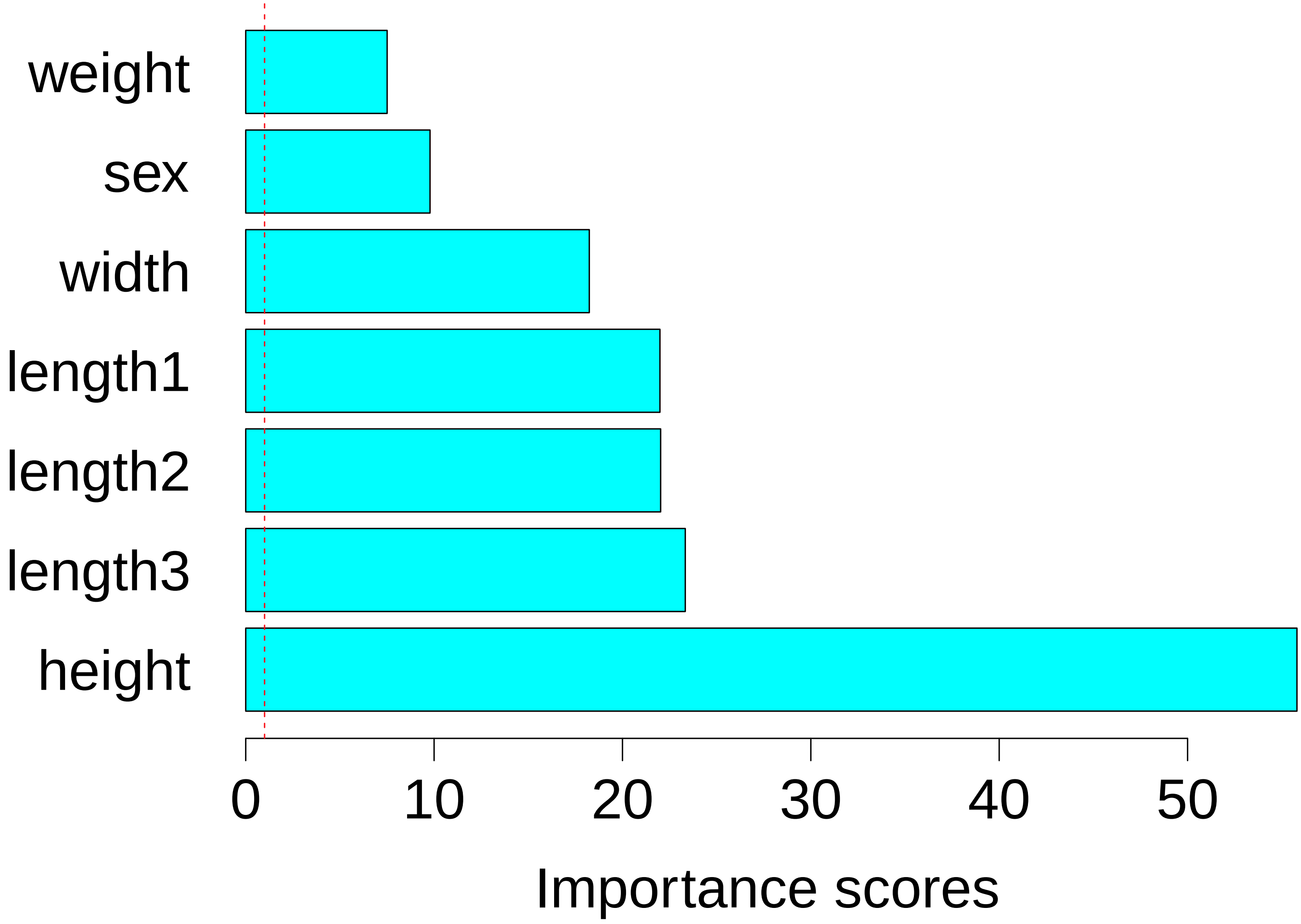
  and k_0 is a user-specified expected number of unimportant variables erroneously identified as important (default value of k_0 is 2 for classification and 1 for regression)

Importance scores for iris (left) and mammography (right) data

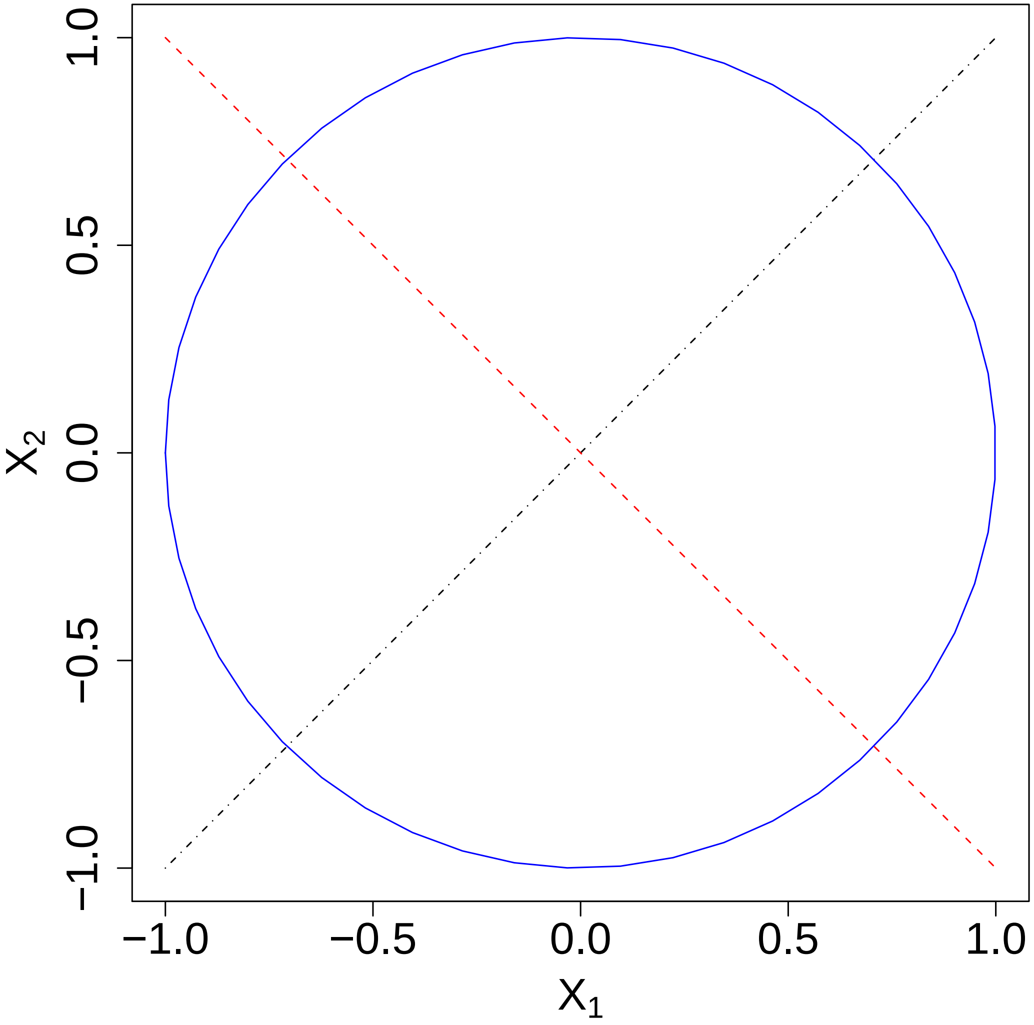




Importance scores for fish data



A hard three-class problem with 8 predictors



Class 1 dist. on circle, class 2 on one diagonal, class 3 on other diagonal

Kernel density estimation

1. Let s and r be the SD and inter-quartile range of $x_1, x_2, \ldots, x_n$

2. The kernel density estimate is

$$\hat{f}(x) = (nh)^{-1} \sum_{i=1}^{n} \phi\{(x - x_i)/h\}$$

where ϕ is the standard normal density function and h is the bandwidth

$$h = \begin{cases} 2.5 \min(s, 0.7413r) n^{-1/5}, & \text{if } r > 0 \\ 2.5 s n^{-1/5}, & \text{otherwise} \end{cases}$$

Kernel node models

Let Y denote the class variable

1. If the split is due to a marginal chi-squared, let X be the selected variable and fit a kernel density estimate to X for each class in t
2. If the split is due to an interaction chi-squared, let X_1 and X_2 be the selected variables. Fit a bivariate density estimate to (X_1, X_2) for each class in t:
   (a) If X_1 and X_2 are categorical, use their sample class joint density
   (b) If X_1 is categorical and X_2 is ordinal, for each combination of (X_1, Y) values in t, let $h(Y, X_1)$ be the bandwidth and $\bar{h}(Y)$ their average. For each value of X_1 and Y, find a kernel density estimate for X_2 using $\bar{h}(Y)$ as bandwidth.
   (c) If X_1, X_2 are ordinal, fit a bivariate Gaussian kernel density to each class with correlation equal to the class sample correlation

The predicted class is the one with the largest estimated density

Nearest-neighbor node models

Given n, define $k = \max(3, \lceil \log n \rceil)$

1. If the split is due to a marginal chi-squared, let X be the selected variable
   (a) If X is categorical, $\hat{Y}$ is the highest probability class among the observations in t with the same X value as the one to be classified
   (b) If X is ordinal, use k-NN classifier based on X with $n = N(t)$
2. If the split is due to an interaction chi-squared, let X_1 and X_2 be selected
   (a) If both are categorical, $\hat{Y}$ is the highest probability class among the cases in t with the same (X_1, X_2) values as the one to be classified
   (b) If X_1 is categorical and X_2 is ordinal, $\hat{Y}$ is given by the k-NN classifier based on X_2 applied to the set S of observations in t that have the same X_1 value as the one to be classified, with n being the size of S
   (c) If both variables are ordinal, use the bivariate k-NN classifier based on (X_1, X_2) with the Mahalanobis distance and $n = N(t)$

GUIDE treatment of missing values

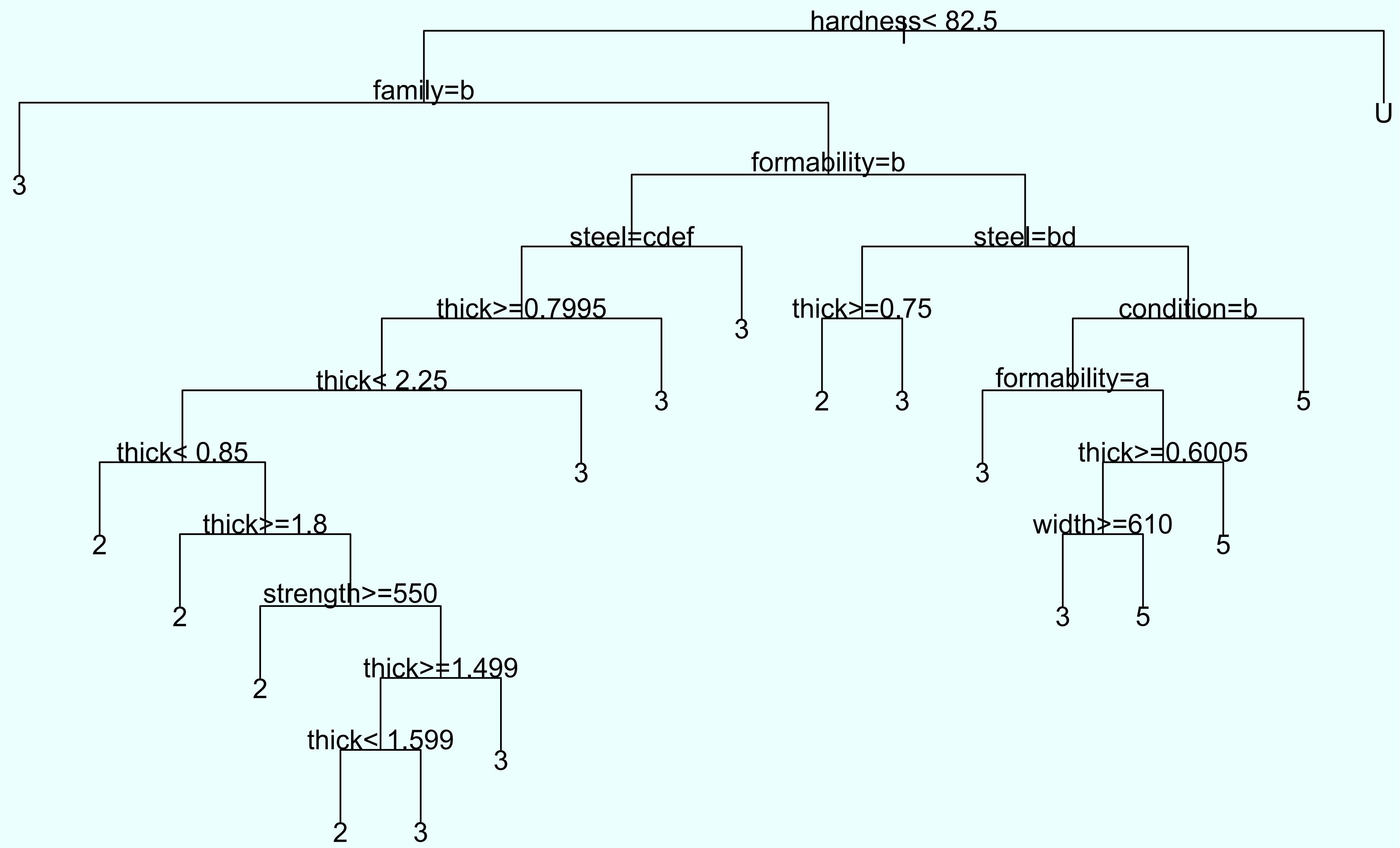
1. Cases with missing Y-values are not used for tree construction
2. For categorical X, missing values are assigned a separate “missing” category
3. For ordinal X:
   (a) Cases with missing values are assigned to a “missing” interval for selection of split variables
   (b) A split on missingness is always considered for split point selection
   (c) Two splits are evaluated for each split on a non-missing value: one for each way of sending the missing values

Annealing data: lots of missing values

Variable	C	M	Variable	C	M	Variable	C	M
class	5		surfacequality	4	217	exptl	1	796
family	2	687	enamelability	2	785	ferro	1	772
steel	7	70	bc	1	797	bright	3	793
carbon	o		bf	1	680	lustre	1	753
hardness	o		bt	1	736	shape	2	
temperrolling	1	675	bwme	2	609	width	o	
condition	2	271	bl	1	662	len	o	
formability	4	283	chrom	1	775	oil	2	740
strength	o		phos	1	791	bore	3	
nonageing	1	703	cbond	1	730	packing	2	789
surfacefinish	1	790	thick	o				

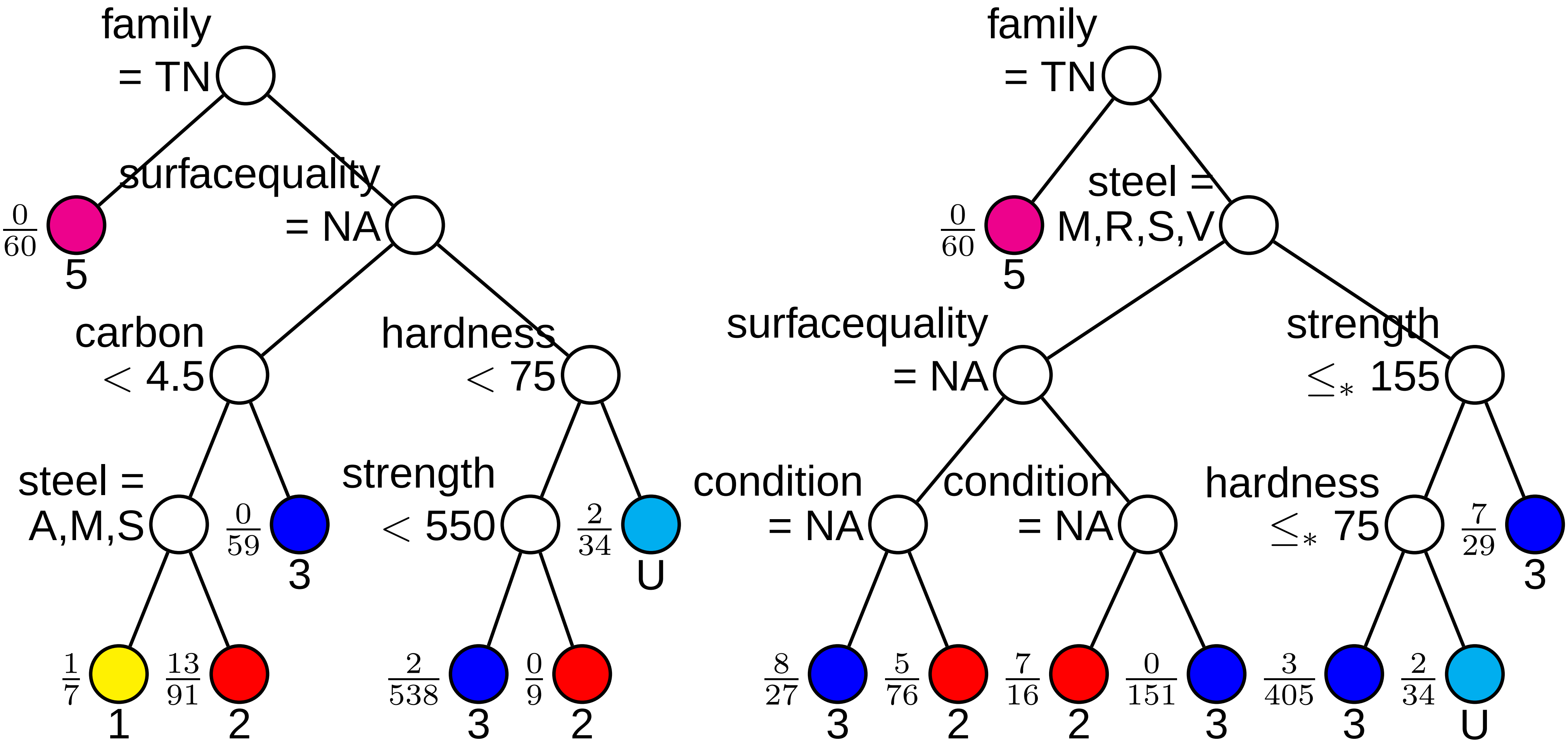
- 798 observations; 6 ordinal and 25 categorical variables
- Cols. C and M give numbers of categories and missing values (o = ordinal)

RPART tree for annealing data



RPART does not split on missing values and on 1-level categorical variables

RPART (left, with missing as separate category) and GUIDE (right) trees for annealing data



#Misclassified/sample size beside each node

C4.5 (Quinlan, 1993)

- Univariate splits only
- Binary splits on ordered predictors via exhaustive search; splits at data values
- Multiway splits on categorical predictors
  — one subnode for each categorical value (with option to merge categories)
- Pruning based on statistical heuristics; no cross-validation
- Missing values handled by case weights
- Priors and misclassification costs cannot be specified
- Cross-validation error estimate available

C4.5: Gain ratio split criterion

- Define the "info" at node t as the entropy

$$\text{info}(t) = -\sum_j p(j|t) \log_2\{p(j|t)\}$$

- Suppose t is split into subnodes $t_1, \ldots, t_n$ by predictor X. Define

$$\begin{aligned}
\text{info}_X(t) &= \sum_i \text{info}(t_i) \frac{N(t_i)}{N(t)} \\
\text{gain}(X) &= \text{info}(t) - \text{info}_X(t) \\
\text{split info}(X) &= -\sum_i \frac{N(t_i)}{N(t)} \log_2 \frac{N(t_i)}{N(t)} \\
\text{gain ratio}(X) &= \frac{\text{gain}(X)}{\text{split info}(X)}
\end{aligned}$$

- Split that yields the highest gain ratio is selected

C4.5: Case weights for missing values

- Initialize the weight for each case to be 1 at the root node
- Suppose t is split by X into subnodes $t_1, \ldots, t_n$
- Let $W(t_i)$ be the sum of the weights of cases with known X that land in t_i and let $W(t) = \sum_i W(t_i)$
- If a case in learning sample with weight w is missing X, send it down each subnode with weight in t_i equal to

$$w_i = \frac{W(t_i)}{W(t)} w$$

- Do the same for each test case. If a test case ends up in more than 1 terminal node, assign it the class with largest total weight

Generalization when there are missing values

- Let $p_w(j|t) = \dfrac{\text{sum of class } j \text{ weights in } t}{\text{total weight in } t}$ and define:

$$\begin{aligned}\text{info}(t) &= -\sum_j p_w(j|t)\log_2\{p_w(j|t)\} \\ \text{info}_X(t) &= \sum_i \text{info}(t_i)\frac{W(t_i)}{W(t)}\end{aligned}$$

- Let f be the fraction of learning cases in t that are nonmissing X and define

$$\begin{aligned}\text{gain}(X) &= f \times \{\text{info}(t) - \text{info}_X(t)\} \\ \text{split info}(X) &= -\sum_i \frac{W(t_i)}{W(t)}\log_2\frac{W(t_i)}{W(t)} - (1-f)\log_2(1-f) \\ \text{gain ratio}(X) &= \frac{\text{gain}(X)}{\text{split info}(X)}\end{aligned}$$

C4.5: Pruning

- Suppose $N_E(t)$ learning cases are misclassified in node t
- C4.5 estimates the true misclassification probability with the upper 75% confidence bound p where

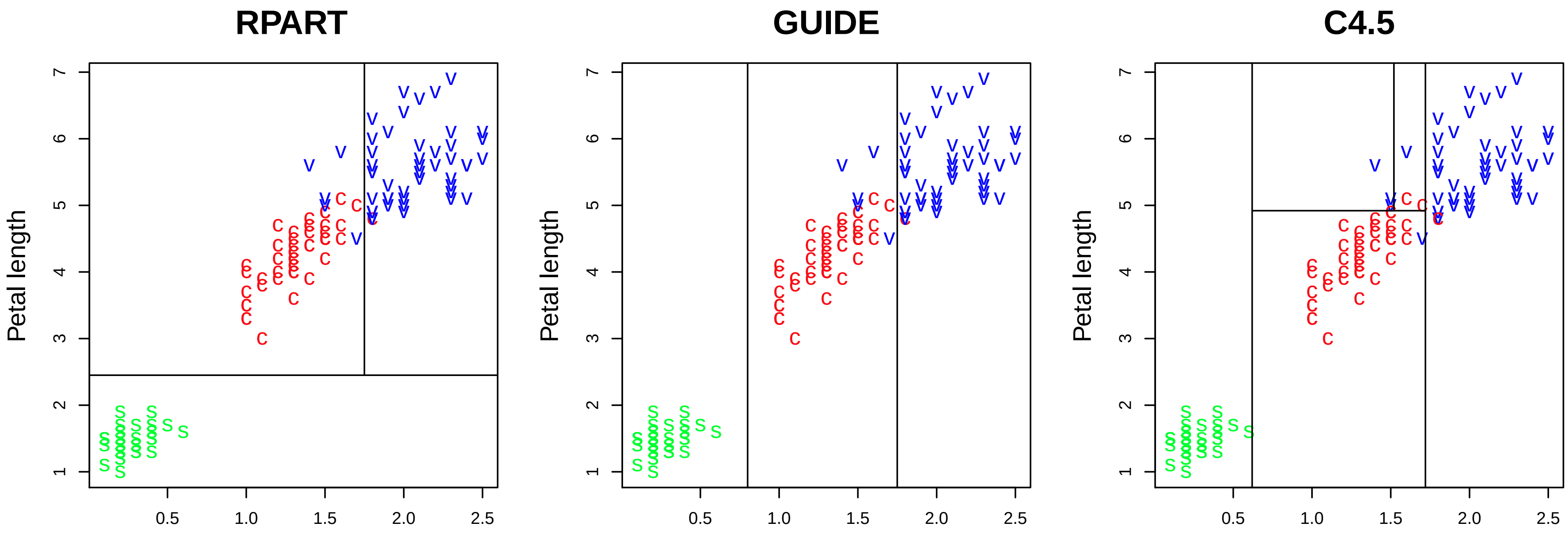
$$\sum_{i=0}^{N_E(t)} \frac{N(t)!}{i!\,(N(t)-i)!} p^i (1-p)^{N(t)-i} = 0.25$$

- Let $\nu_1 = 2(N(t) - N_E(t) + 1)$, $\nu_2 = 2N_E(t)$ and $F_{\nu_1,\nu_2;0.75}$ be the 75% percentile of the F_{ν_1,ν_2} dist. Then (Owen 1962, p. 273)

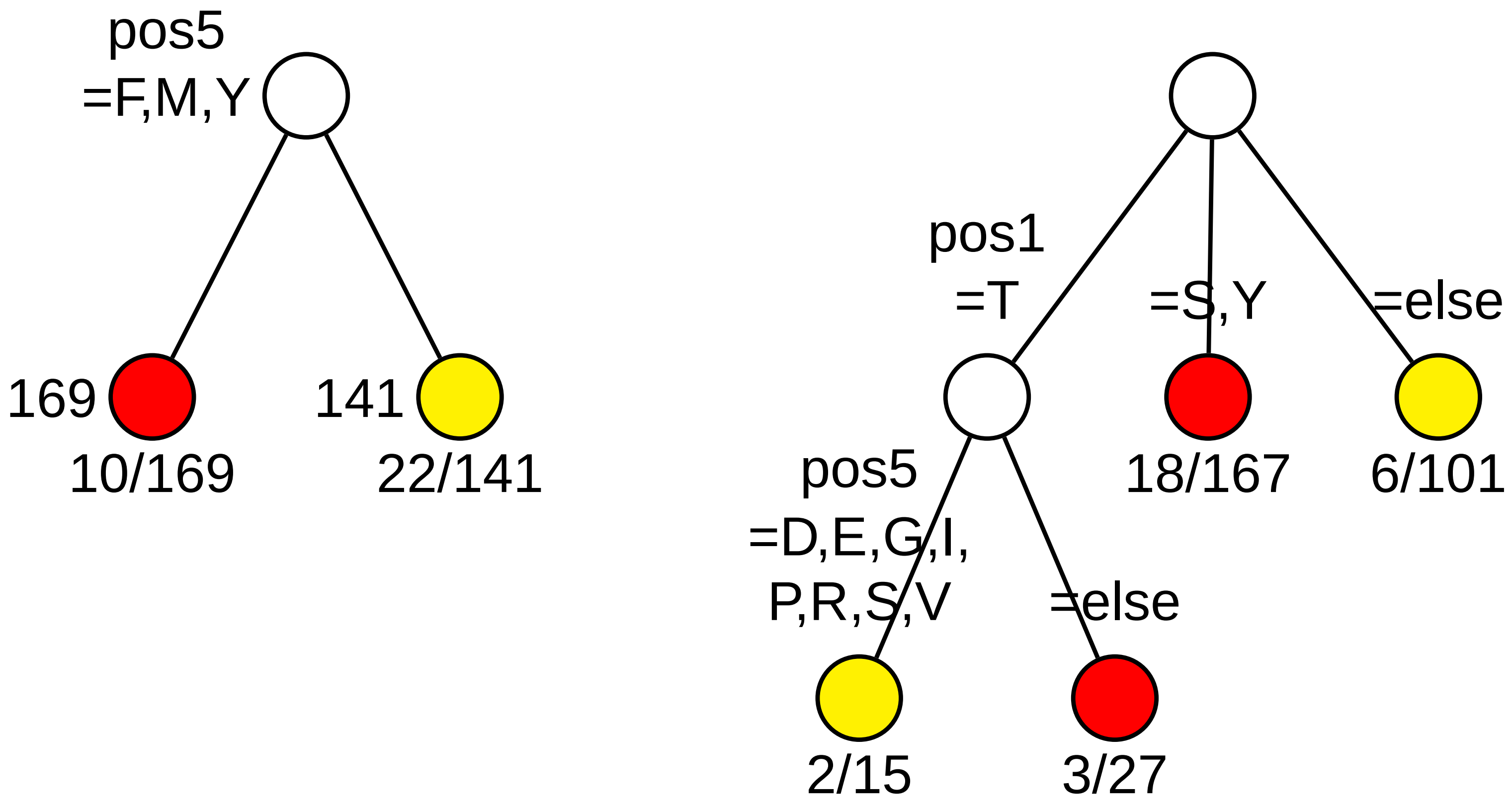
$$p = 1 - \frac{N_E(t)}{N_E(t) + (N(t) - N_E(t) + 1) F_{\nu_1,\nu_2;0.75}}$$

- The misclassification cost at t is estimated by $N(t)p$
- A branch is pruned if its estimated cost is larger than its root node

RPART, GUIDE and C4.5 trees for iris data



RPART (left) and J48 (right) trees for peptide data

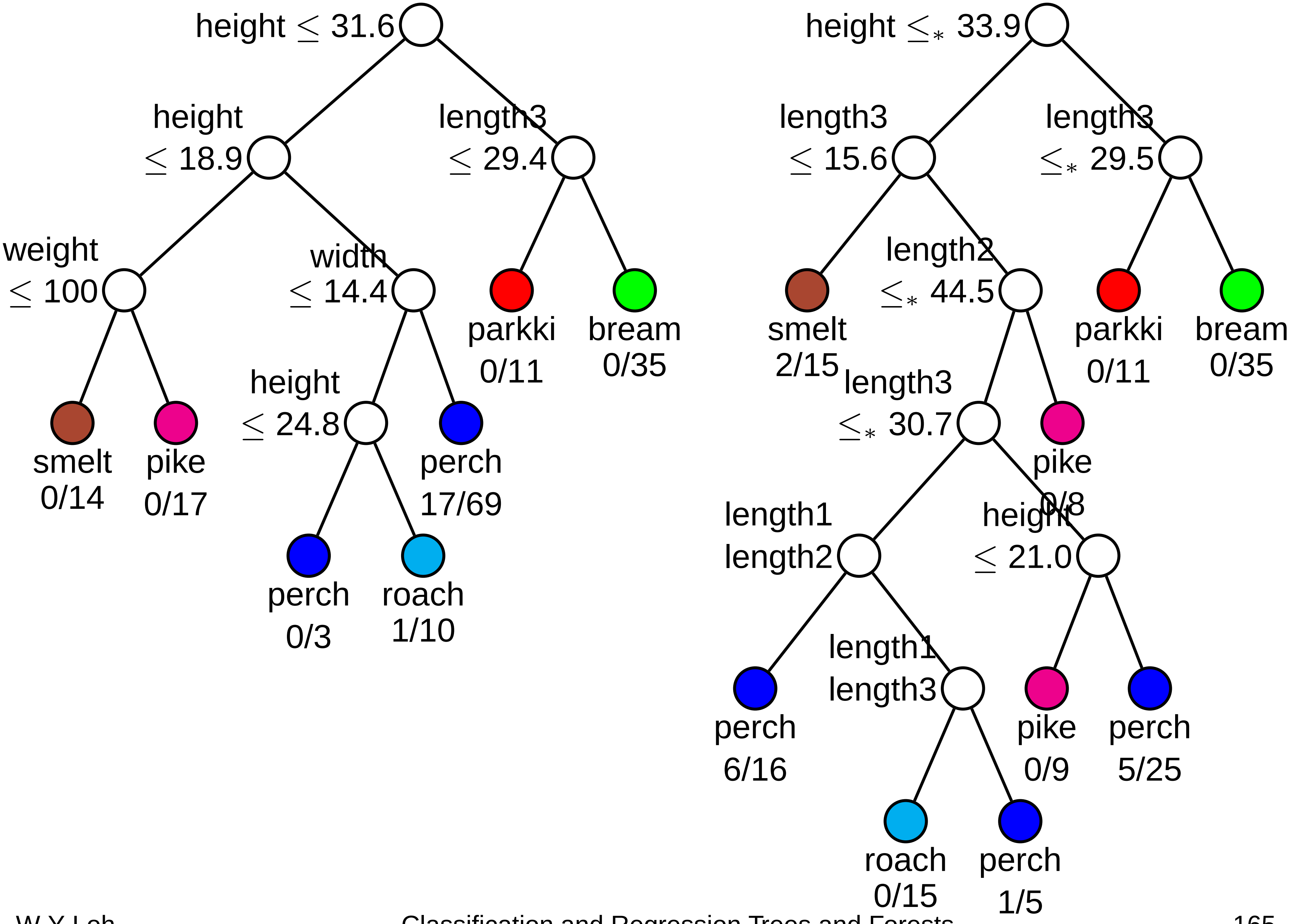


Red denotes binder, yellow denotes non-binder

Numbers beneath nodes are misclassified/sample size

RPART and J48 misclassify 32 and 29 cases, respectively

J48 (left) and GUIDE (right) trees for fish data



CHAID (Kass, 1980)

- Extends AID to categorical and ordered dependent variables
- Uses a direct stopping rule; no pruning
- Uses significance tests to select split variables and split points
- Uses Bonferroni method to control for multiple testing
- Can split each node into more than two subnodes

CHAID predictor types

Monotonic: Ordered or ordinal categorical

Free: Nominal categorical

Floating: Ordinal categorical with exception of a single category that either does not belong to the rest or whose position on the ordinal scale is unknown, e.g., “missing” category

Note: A variable is treated as floating only if it has some missing values in the learning sample. Otherwise it is treated as either monotonic or free. Therefore if a learning sample has no missing values, the tree may not be able to classify future cases that have missing values.

CHAID algorithm

Let $\alpha_1 > \alpha_2$ and α_3 be three given significance levels.

Prepare predictors. Discretize values of each ordinal X into 10 interval groups. For categorical X, the groups are the categories.

Merge categories. Do for each predictor variable:

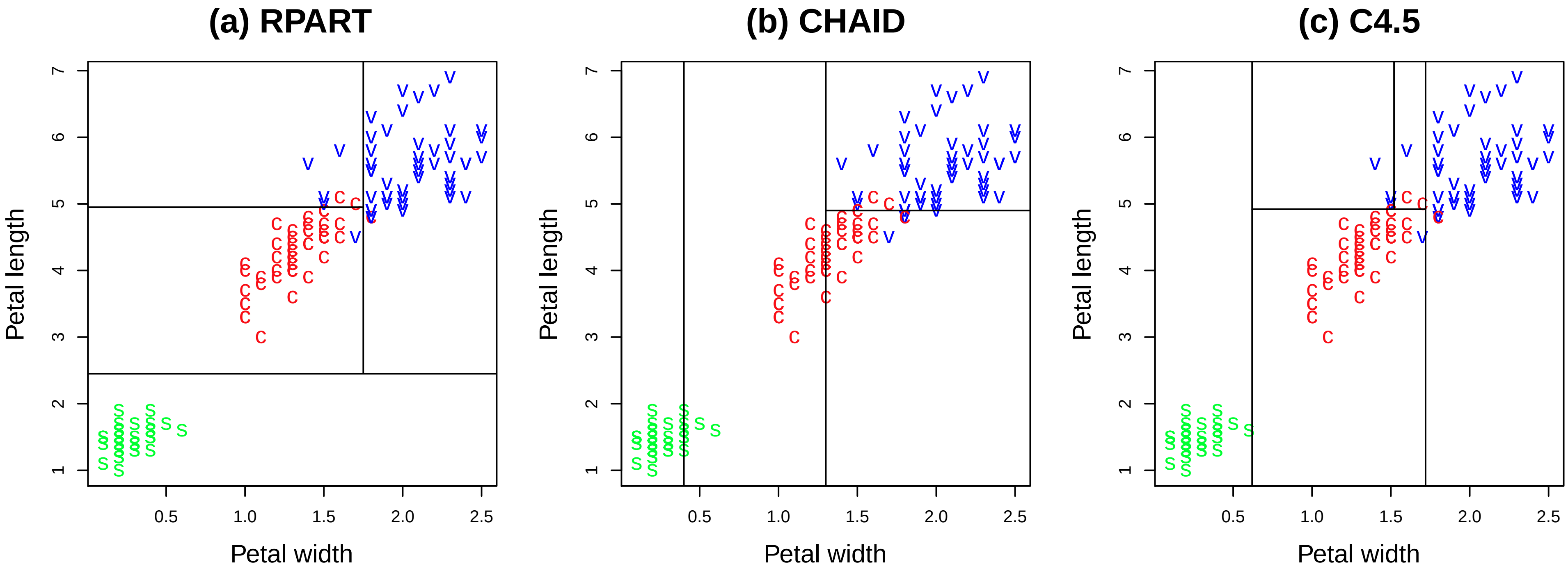
1. For classification, take each pair of categories and compute the p-value of the chi-squared test of independence between categories and class
2. For regression, take each pair of categories and compute the p-value of the two-sample two-sided t-test, using the categories as groups
3. Find least significant pair of categories. If $p > \alpha_1$, merge the two categories and repeat this step.
4. For each compound category containing three or more of the original categories, find the most significant binary split.
   If $p < \alpha_2$, split the compound category and return to Step 3.

Select split. Compute Bonferroni-adjusted p-value for each predictor. If smallest $p < \alpha_3$, split the node; otherwise stop.

CHAID Bonferroni multipliers

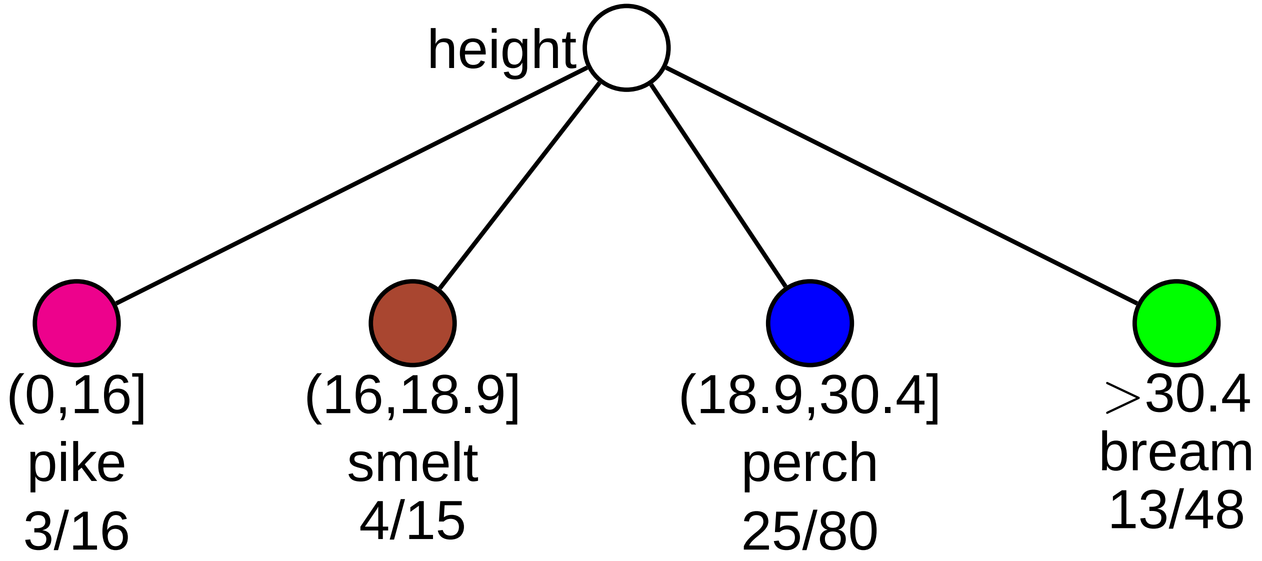
Suppose a predictor with c original categories is merged into r categories. The Bonferroni adjustments to the p-values are:

$$\begin{aligned}
\text{Monotonic:} \quad B &= \binom{c-1}{r-1} \\
\text{Free:} \quad B &= \sum_{i=0}^{r-1} (-1)^i \frac{(r-i)^c}{i!\,(r-i)!} \\
\text{Floating:} \quad B &= \binom{c-2}{r-2} + r\binom{c-2}{r-1}
\end{aligned}$$



s = Setosa, c = Versicolour, v = Virginica

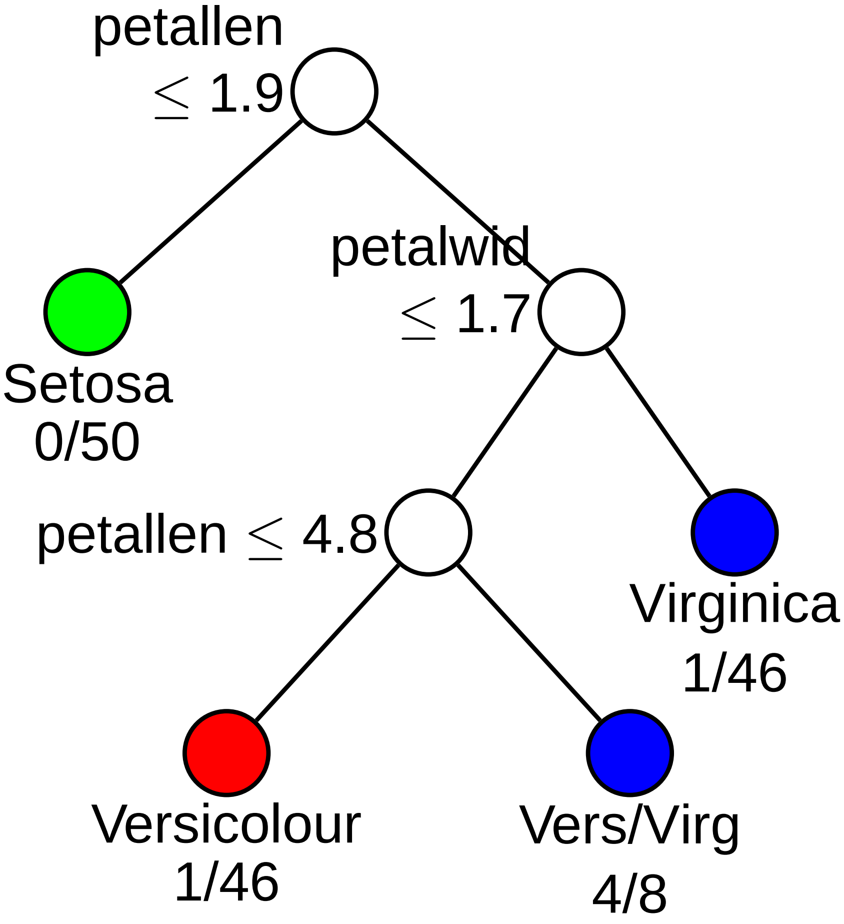
CHAID tree for fish data (45 misclassified)



CTREE (Hothorn et al., 2006)

1. Use conditional permutation tests to select variables
   - Requires computation of p-values, either by exact calculation, Monte Carlo simulation, or asymptotic approximation
2. Use stopping rules controlled by Bonferroni adjustments; no pruning
3. Surrogate splits are used to deal with missing values
4. Permutation tests (with subnode label as response variable) are used to find the surrogate variables

CTREE tree for iris data



petallen ≤ 1.9

Setosa 0/50

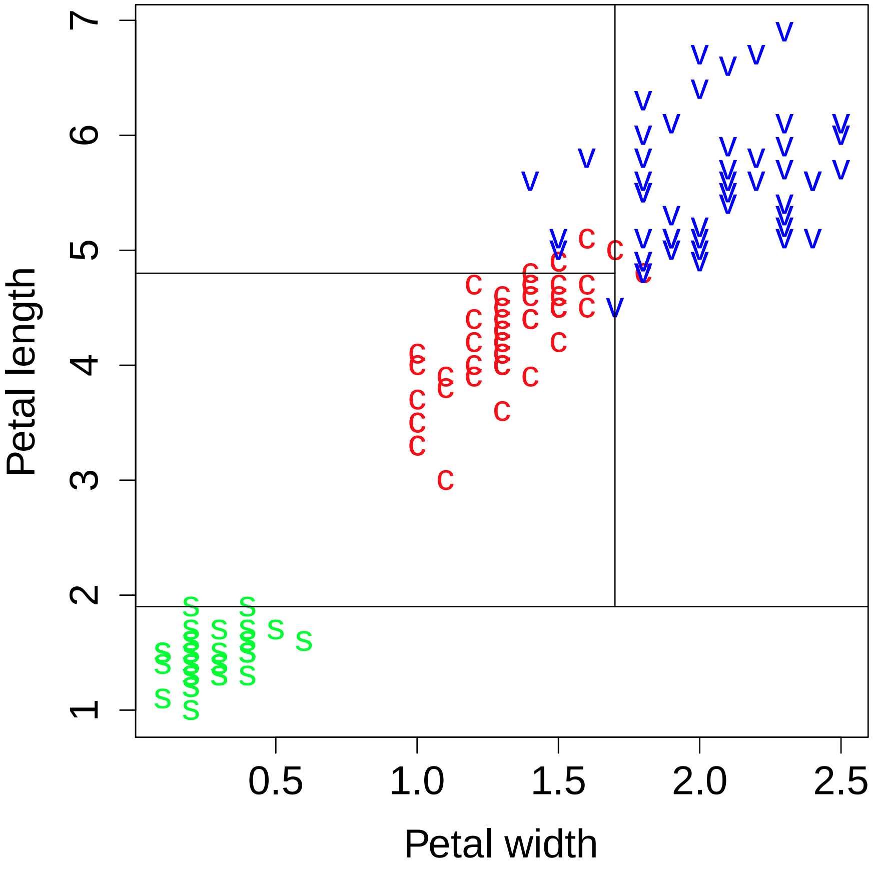
petalwid ≤ 1.7

petallen ≤ 4.8

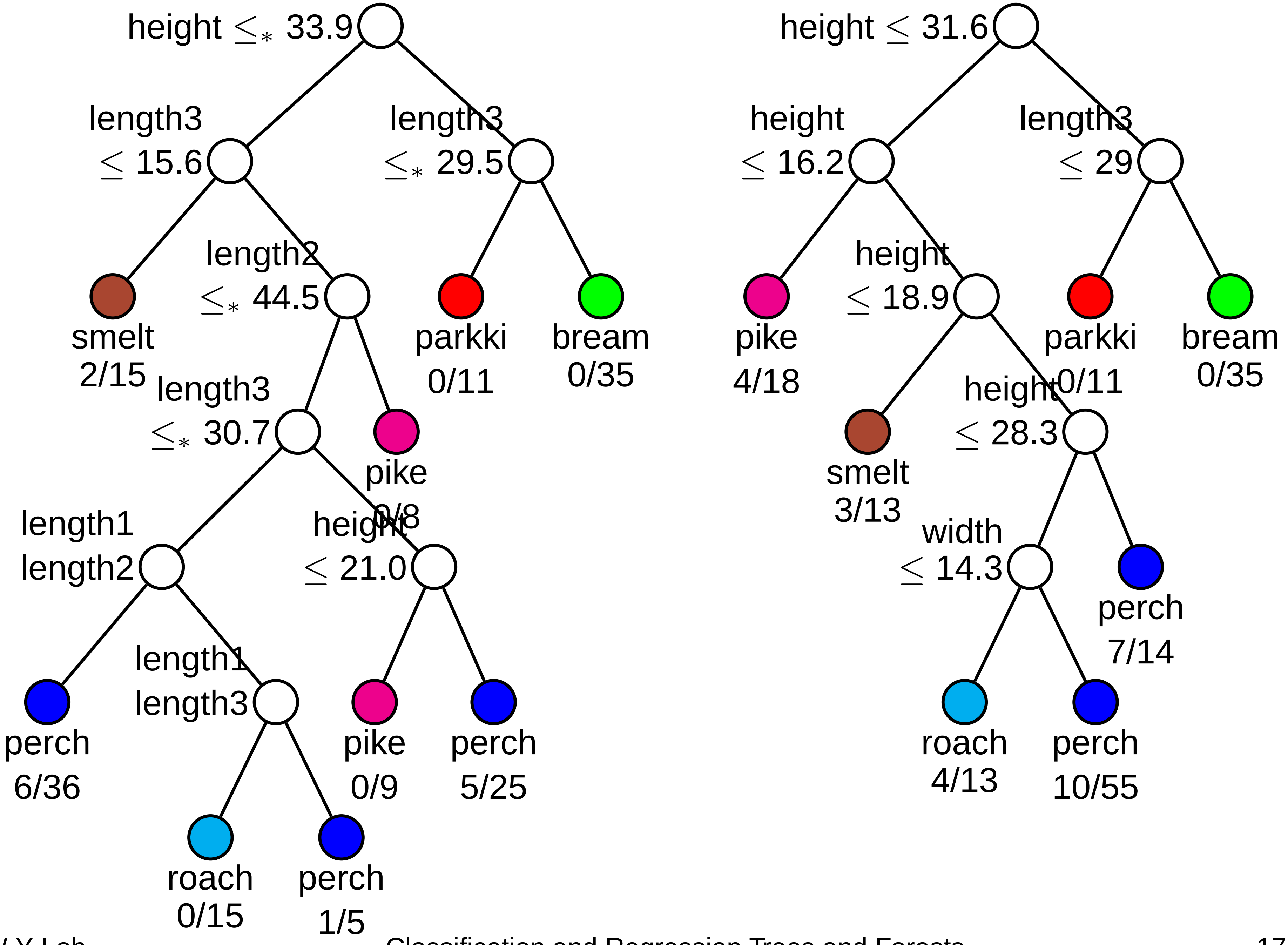
Virginica 1/46

Versicolour 1/46

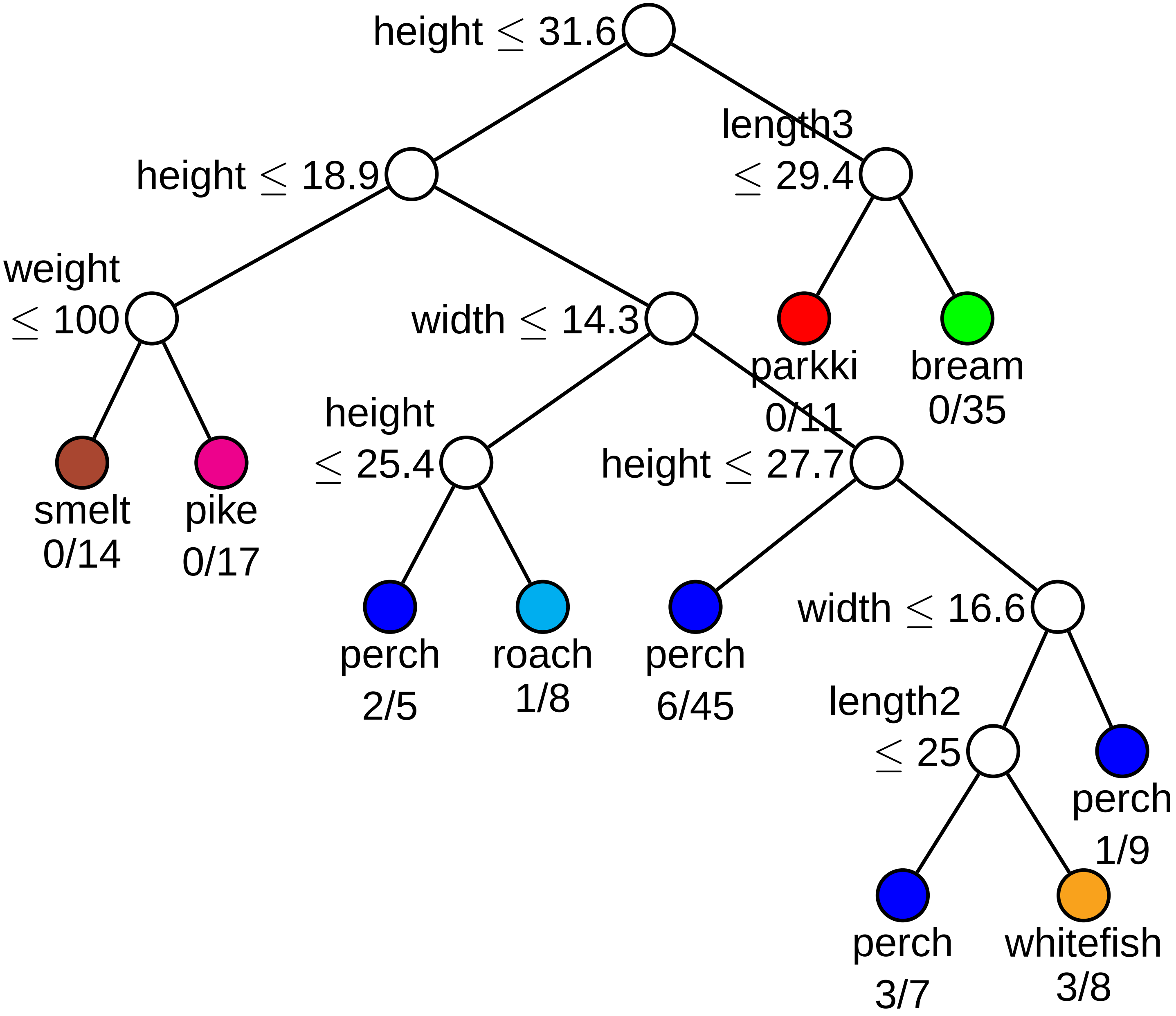
Vers/Virg 4/8



GUIDE (14 errors) and CTREE (28 errors) for fish data



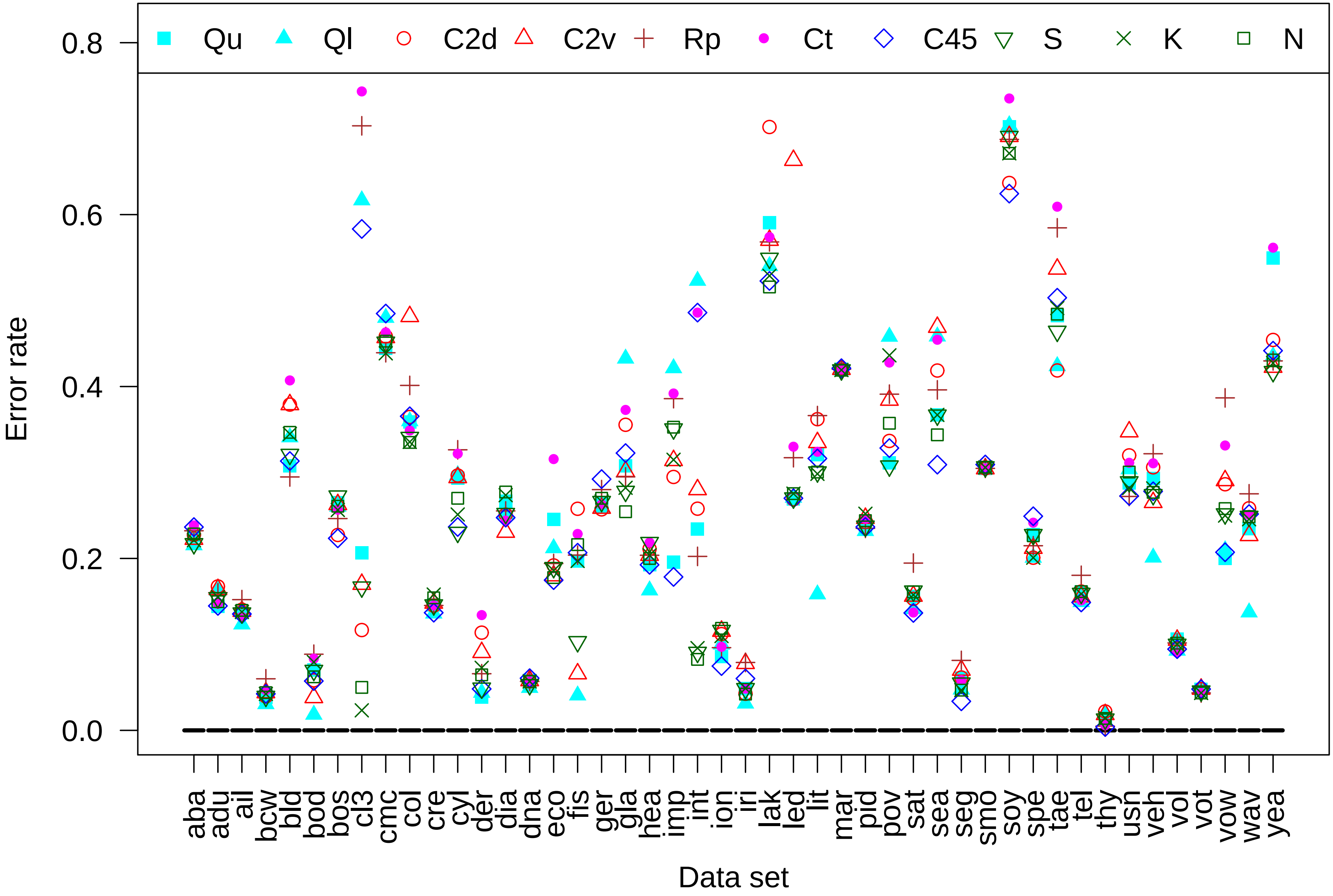
J48 (16 errors) for fish data



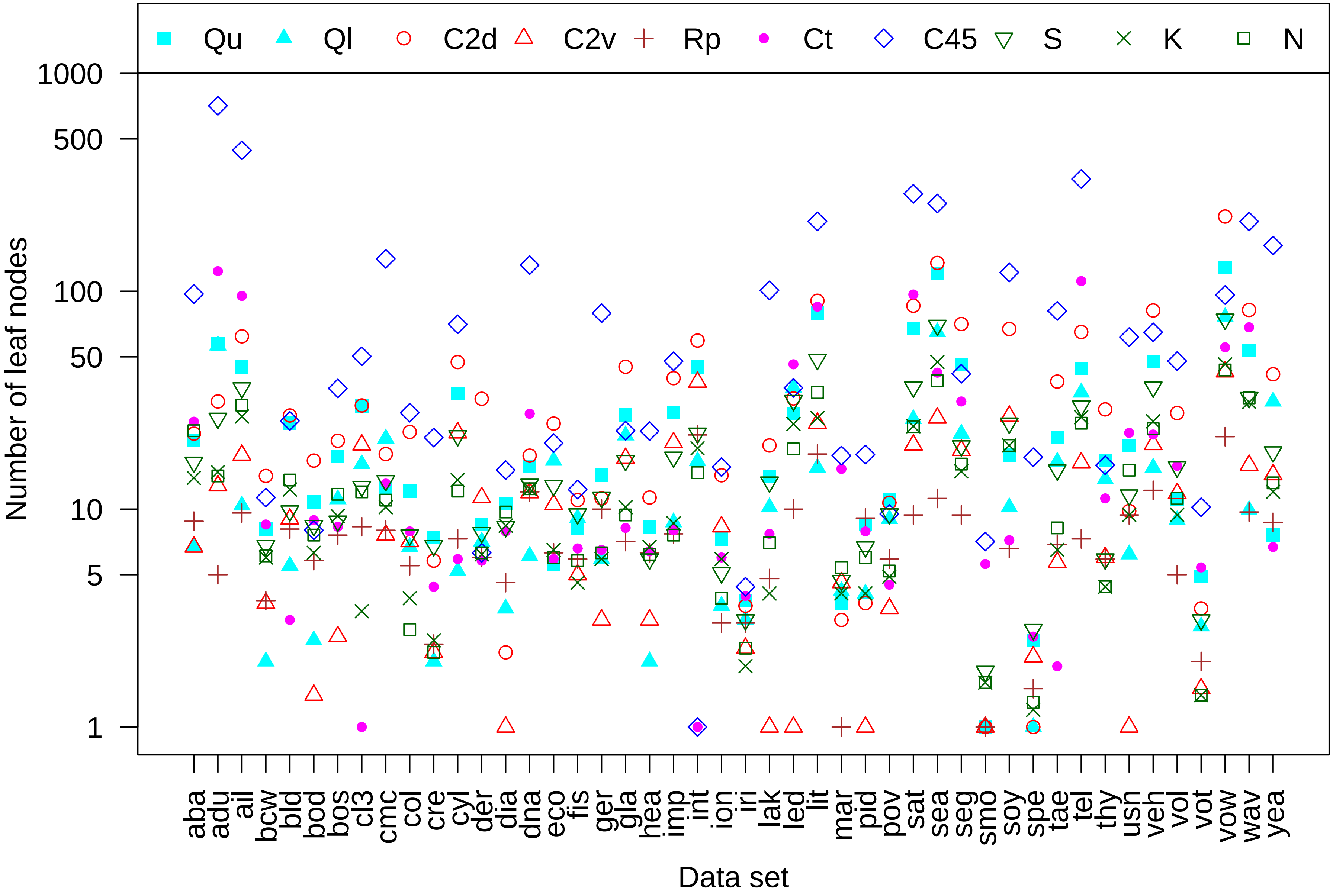
Comparisons on 46 datasets (Loh, 2009)

C45	C4.5
C2d	CRUISE with interaction detection and simple node models
C2v	CRUISE with interaction detection and linear discriminant node models
Qu	QUEST with univariate splits
Ql	QUEST with linear splits
Rp	RPART
Ct	CTree
S	GUIDE with simple node models
K	GUIDE with kernel node models
N	GUIDE with nearest-neighbor node models

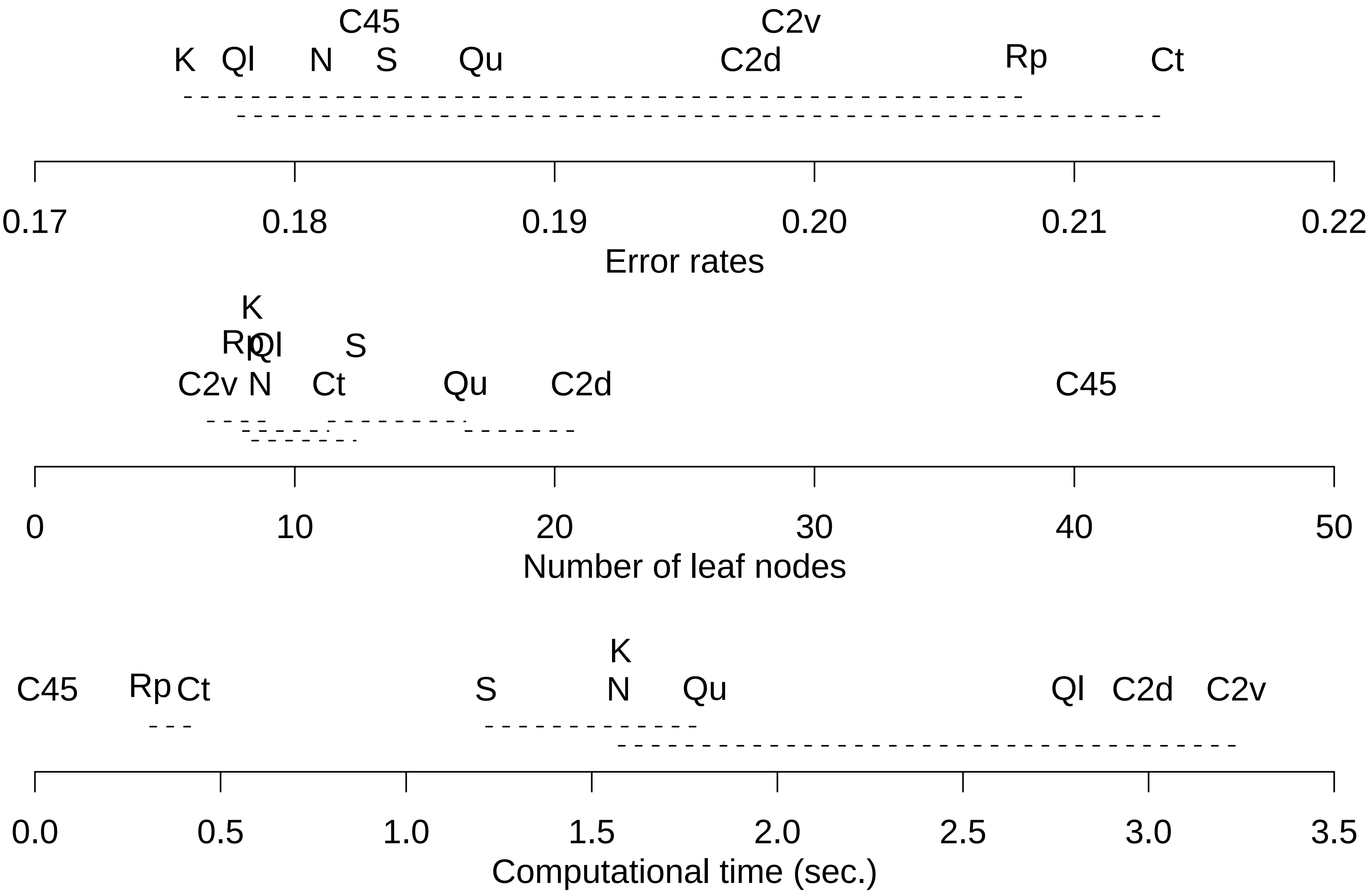
Error rates by dataset



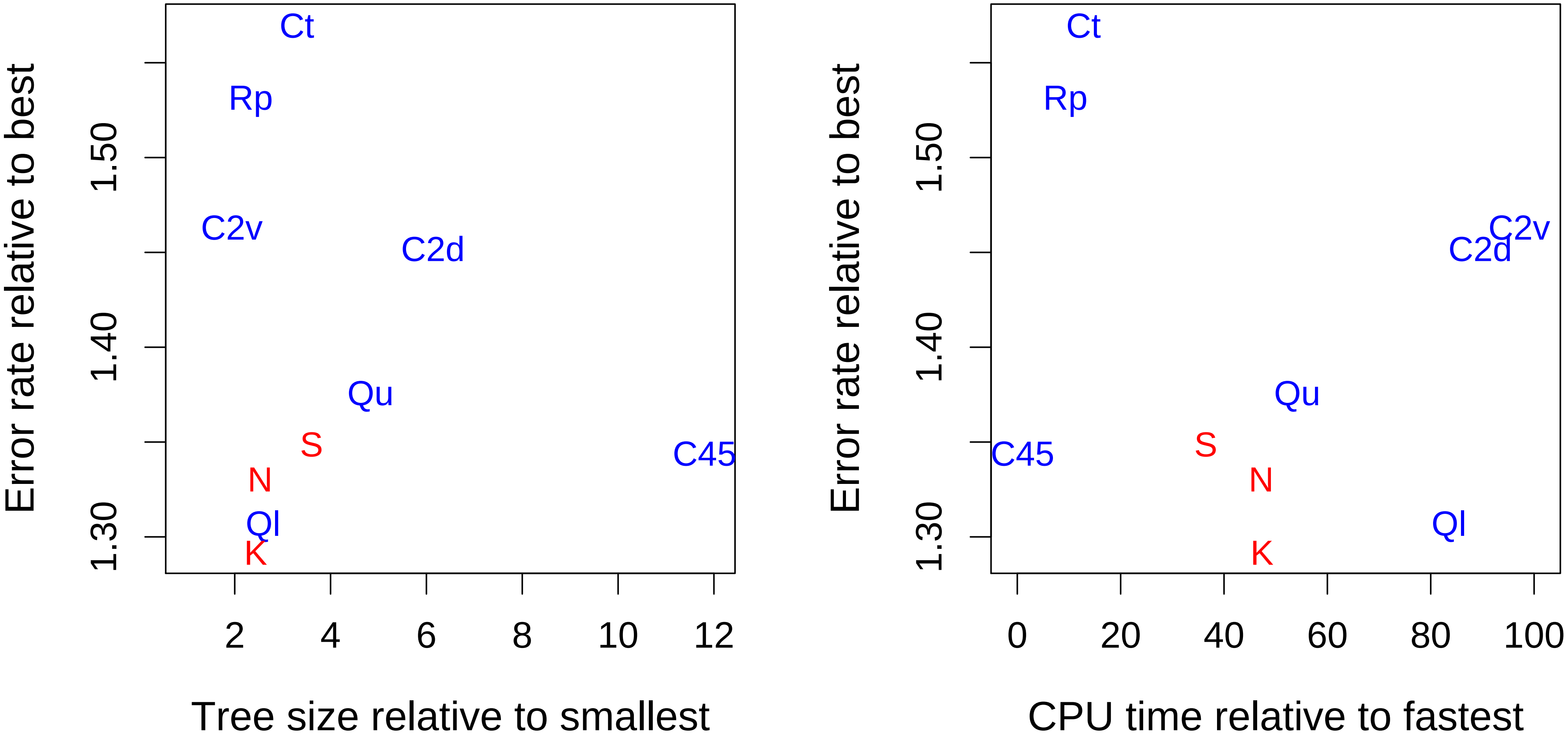
Number of leaf nodes by dataset



Geometric means over 46 datasets



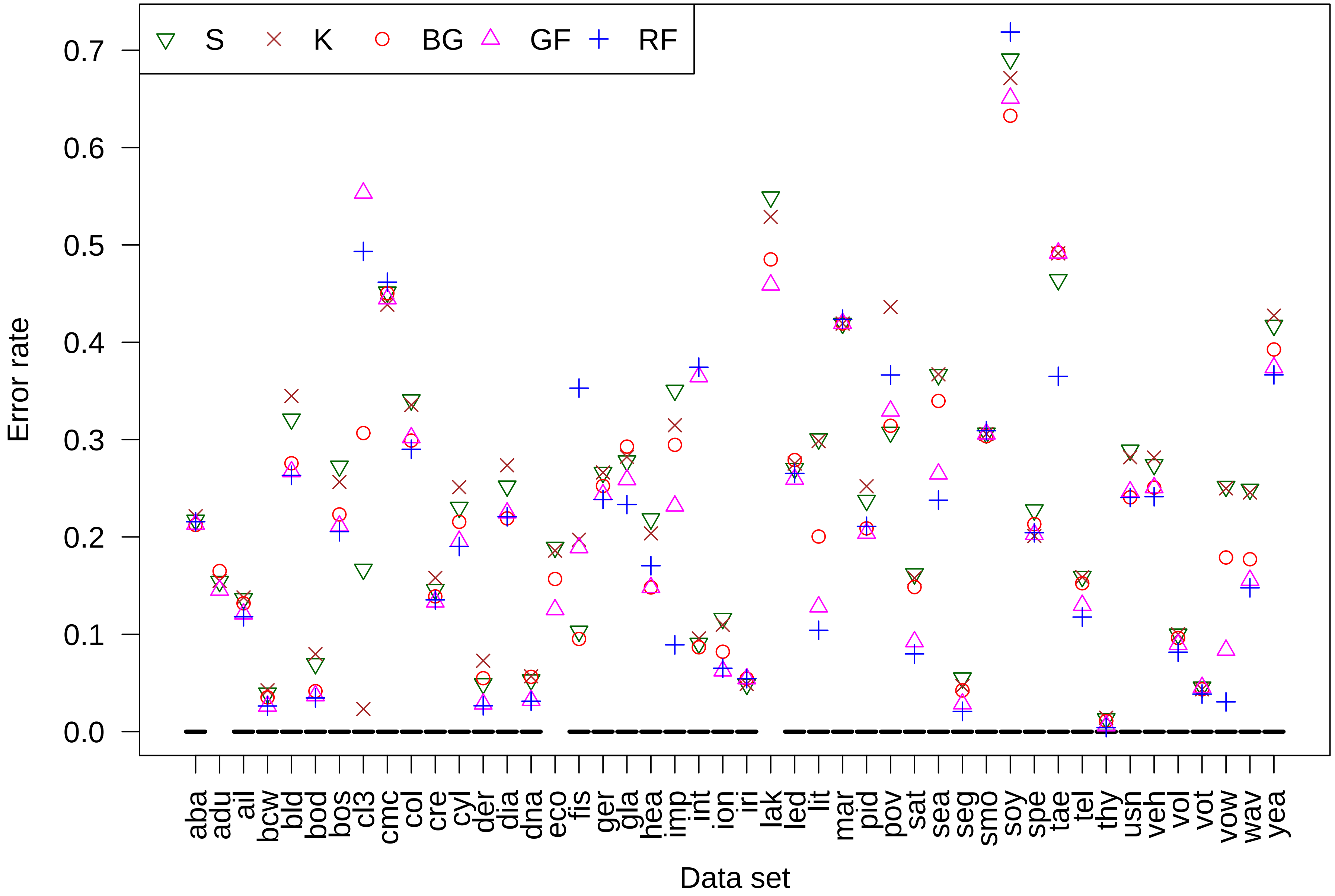
Geometric means relative to best for dataset



Tree ensembles

A tree ensemble uses the majority vote from a collection of tree models to predict the class of an observation

- *Bagging* (Breiman 1996) creates the ensemble by using bootstrap samples of the training data to construct the trees
- *Random Forest* (RF) employs 500 CART trees, but chooses a random subset of $\sqrt{K}$ variables to split each node
- *Bagged GUIDE* (BG) is an ensemble of 100 pruned GUIDE trees, each constructed using the S method from a bootstrap sample
- *GUIDE Forest* (GF) is an ensemble of 500 unpruned GUIDE trees constructed by the S method without interaction and linear splits. As in RF, GF uses a random subset of $\sqrt{K}$ variables to split each node


▽ S
× K
○ BG
△ GF
+ RF
Error rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
aba
adu
ail
bcw
bld
bod
bos
cl3
cmc
col
cre
cyl
der
dia
dna
eco
fis
ger
gla
hea
imp
int
ion
iri
lak
led
lit
mar
pid
pov
sat
sea
seg
smo
soy
spe
tae
tel
thy
usn
veh
vol
vot
vow
wav
yea
Data set

Mean error rates over 43 datasets (Loh, 2009)

Algorithm	S	K	BG	GF	RF
Error rate	0.228	0.231	0.212	0.212	0.206

Notes:

- Although the differences in mean error rates are not statistically significant, ensemble methods tend to have 10% or higher higher prediction accuracy than single-tree methods
- RF is inapplicable if categorical variables have more than 32 levels
  — datasets adu and lak have this characteristic
- RF gives an error if the test sample contains class values that do not appear in the training sample
  — dataset eco has this characteristic

Computational times (sec.) of GUIDE

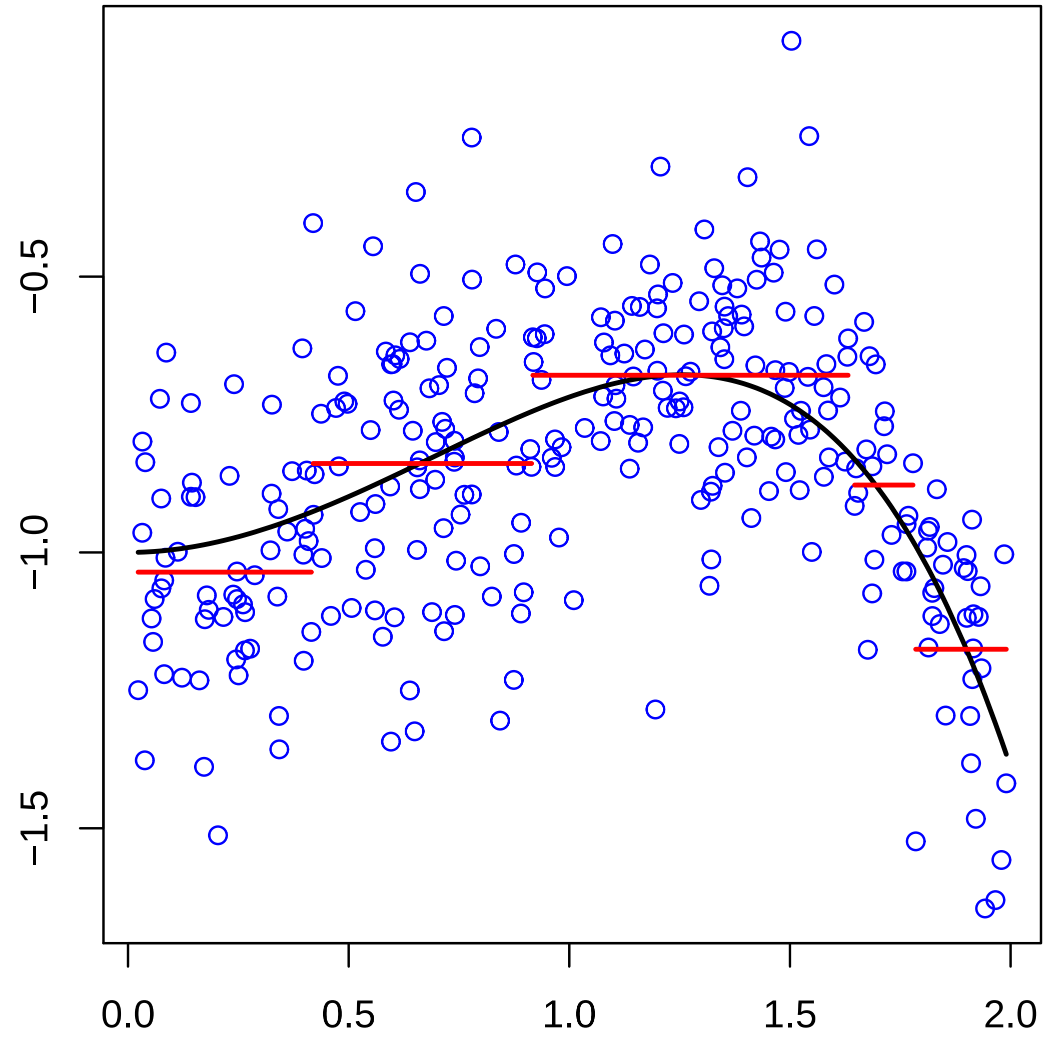
Data	#Obs	#Cat	#Ord	Linux	Win7
Golub	72	0	3,571	2.5	2.8
Adult	32,561	7	6	6.6	7.7
Coil	5,822	2	83	31	36
Arcene	200	0	10,000	71	83
Cover	495,141	2	10	92	106
Gene	1,504	288	17	289	307
Gisette	6,000	0	5,000	403	459
Birthwt	4,268,495	11	12	1933	2198

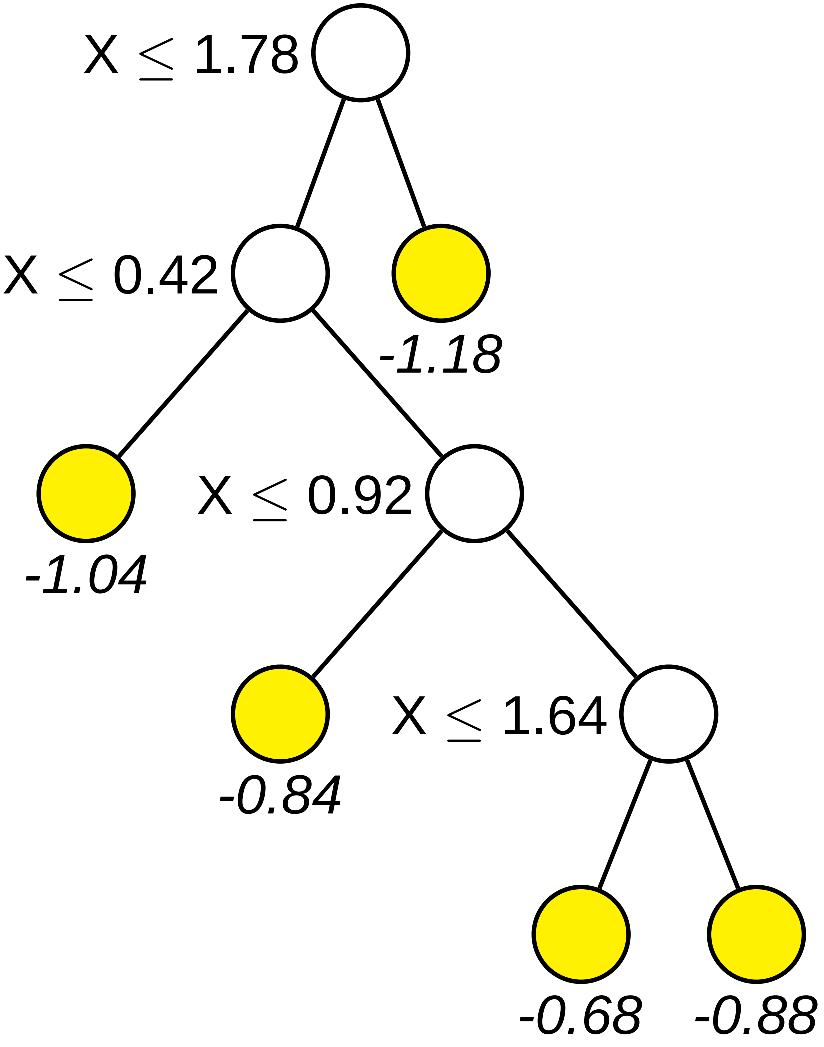
Computer: 16GB 3.3GHz i3-2120; Compiler: Intel Fortran

CART regression tree algorithm

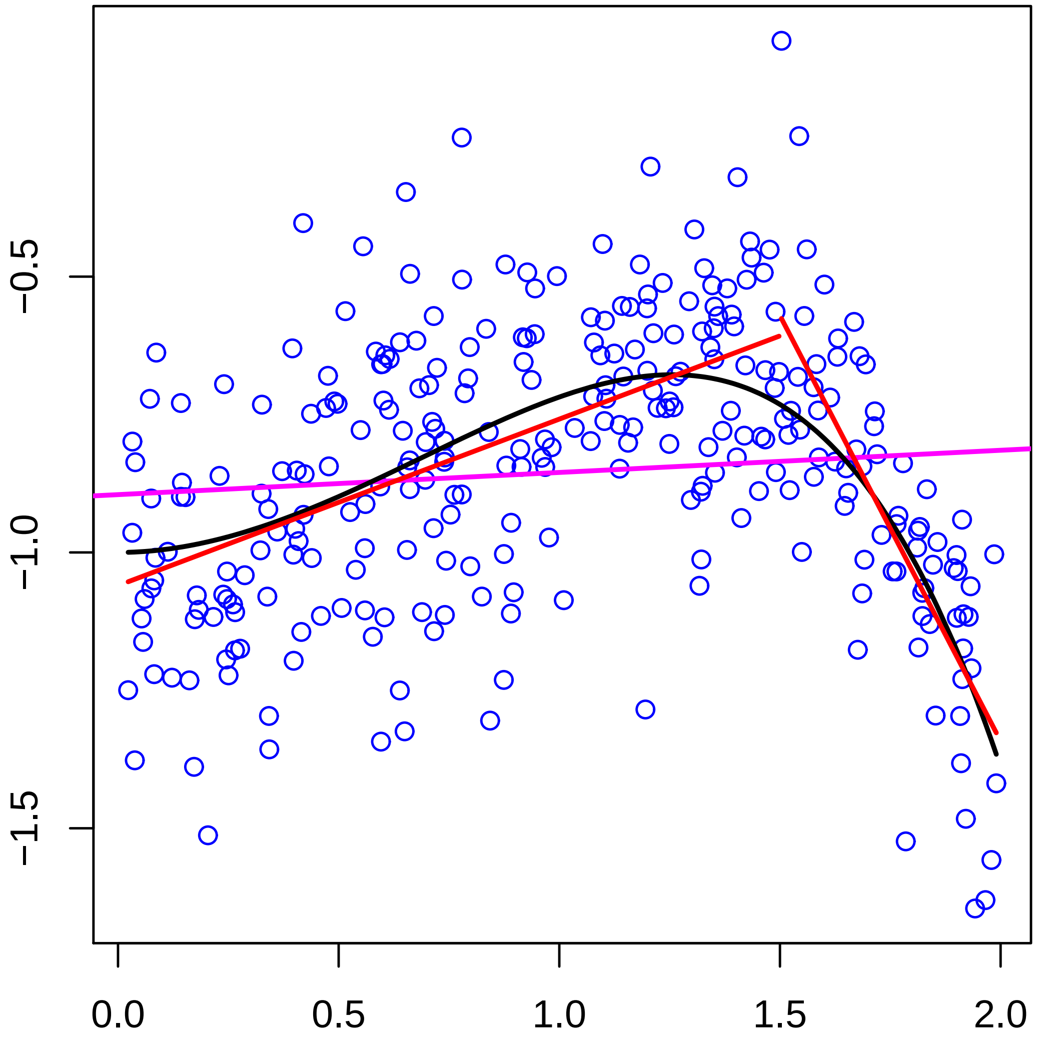
- Fit a constant, the node mean $\bar{y}$, at each node
- Use sum of squared deviations $\sum_i (y_i - \bar{y})^2$ as node impurity
- Keep rest of the CART algorithm unchanged

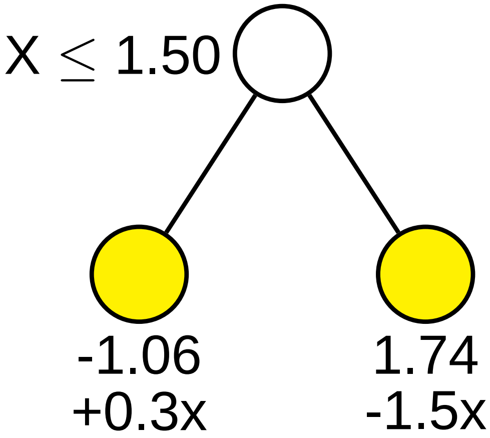
Piecewise-constant regression model





Piecewise-linear regression model





X ≤ 1.50

-1.06
+0.3x

1.74
-1.5x

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

Ref: Chaudhuri et al. (1994, 1995); Chaudhuri and Loh (2002); Loh (2002, 2006, 2008b); Loh and Zheng (2013)

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)

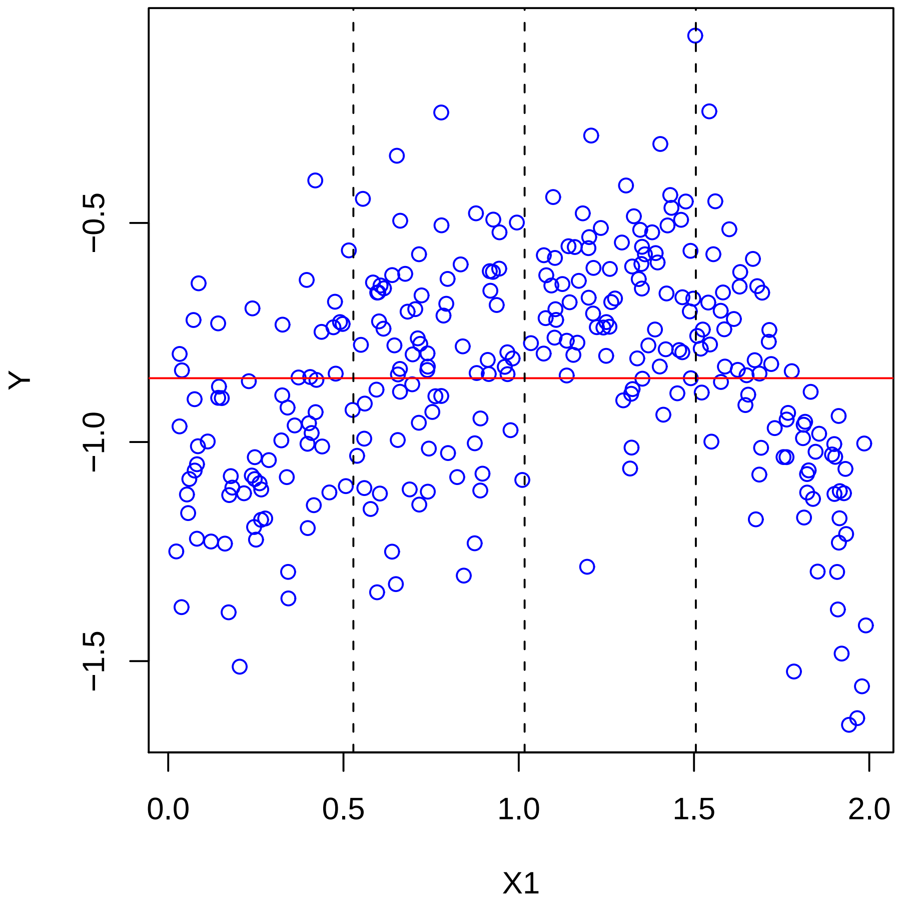
Z: Offset variable (Poisson regression)

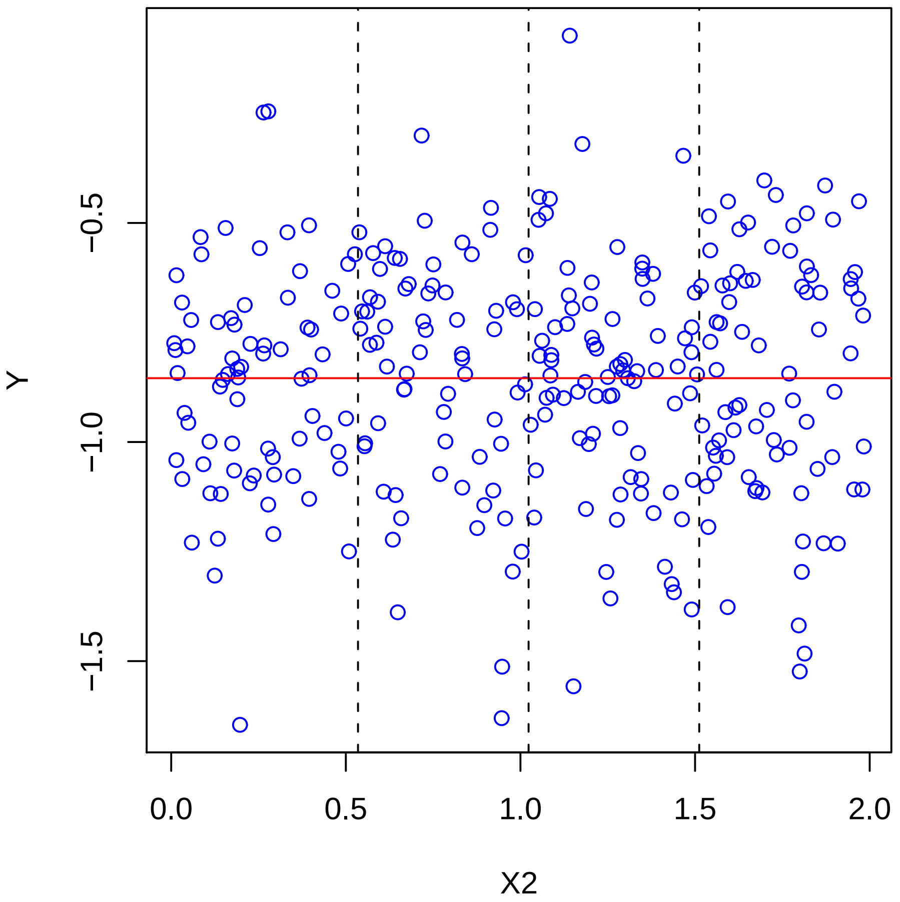
X: Excluded variable

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}

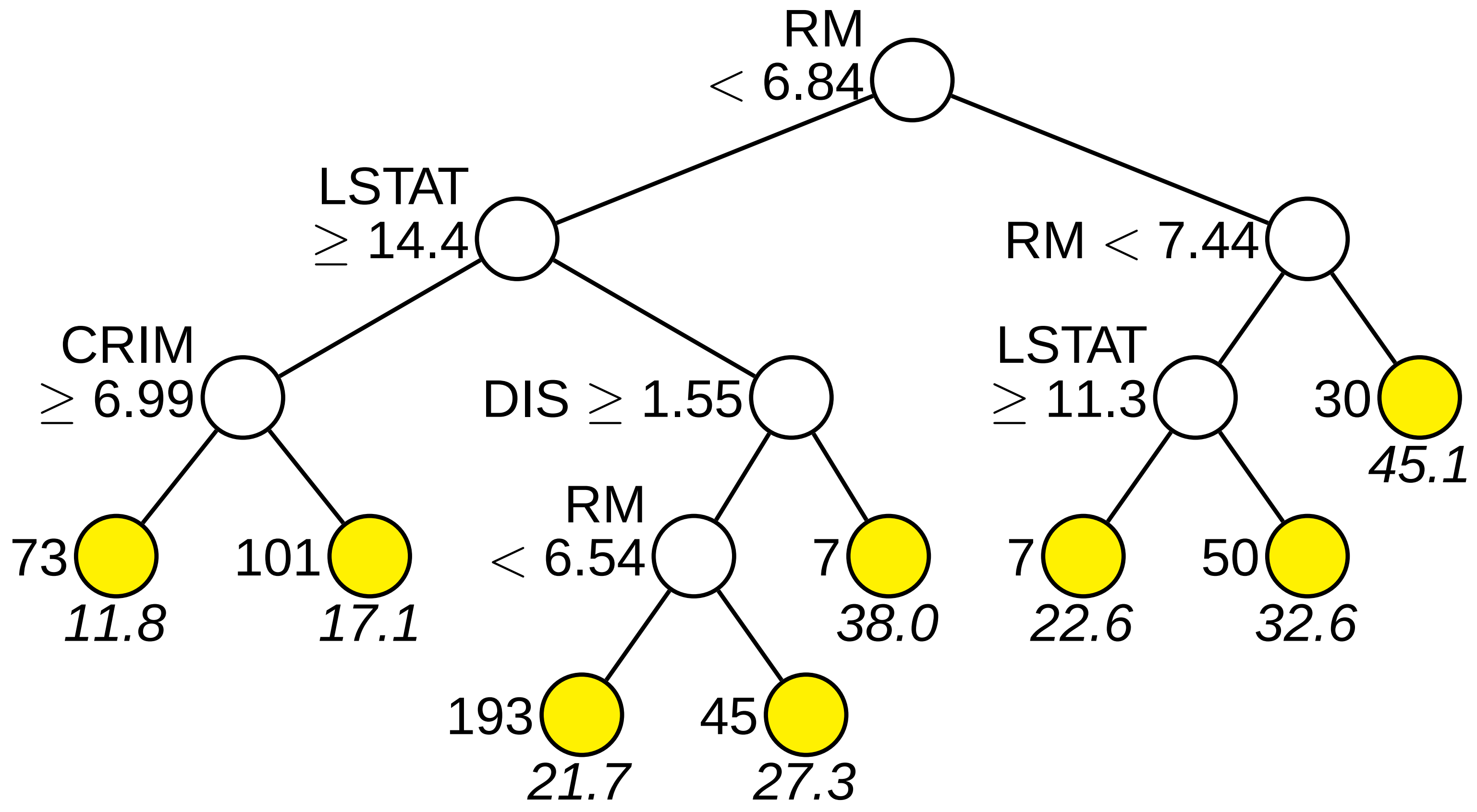
Pos. res.	37	41	45	39
Neg. res.	34	28	39	37

χ^2_3 = 1.14, p = 0.77

Selection bias: Boston housing data

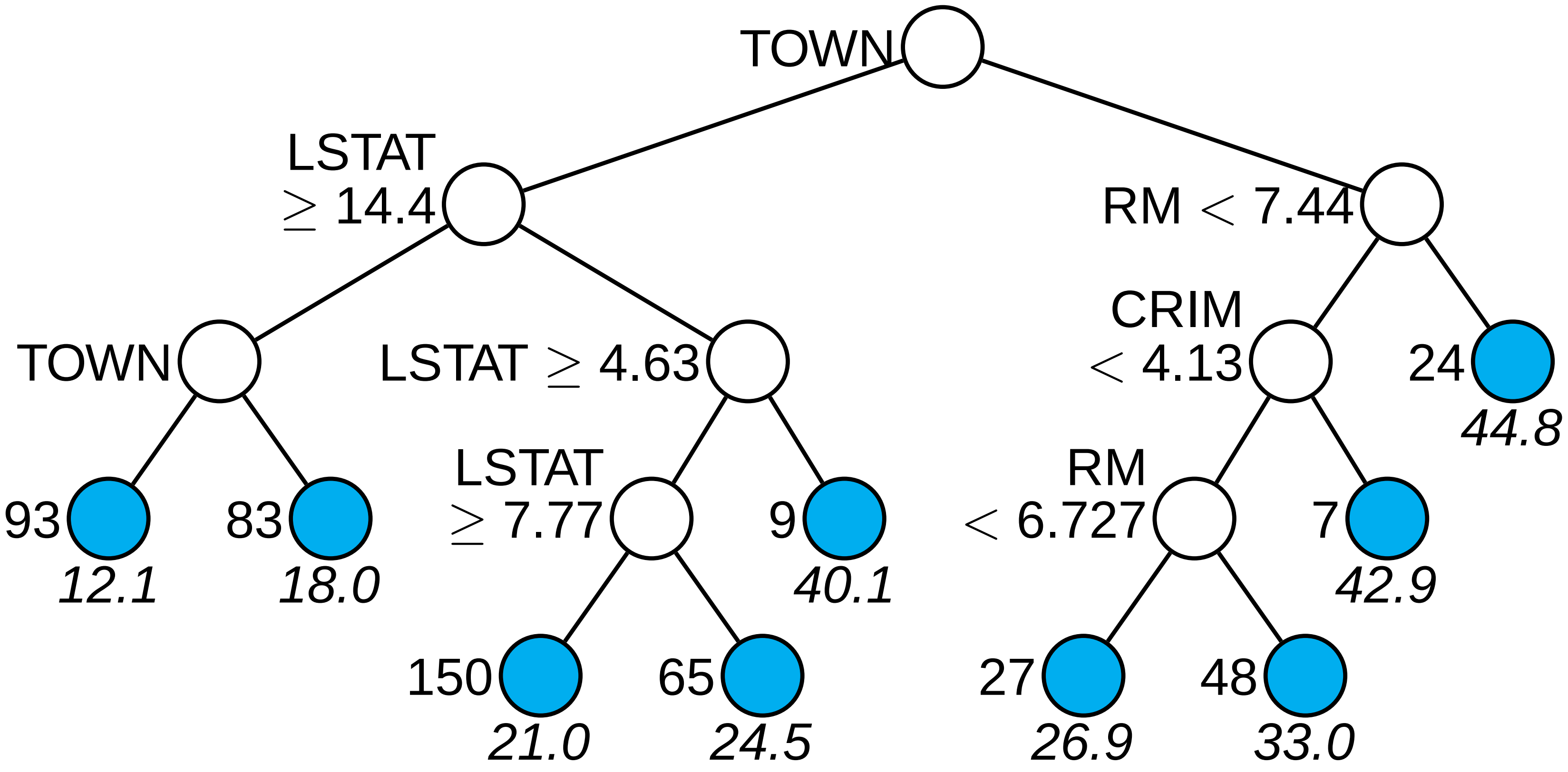
- Categorical variable TOWN has 92 values
- If TOWN is included, RPART has a high chance to select it
  - actually, RPART can search over at most 32 categorical values
  - it is unclear how it deals with TOWN
- GUIDE is much less influenced by the presence of TOWN

RPART tree for MEDV without TOWN



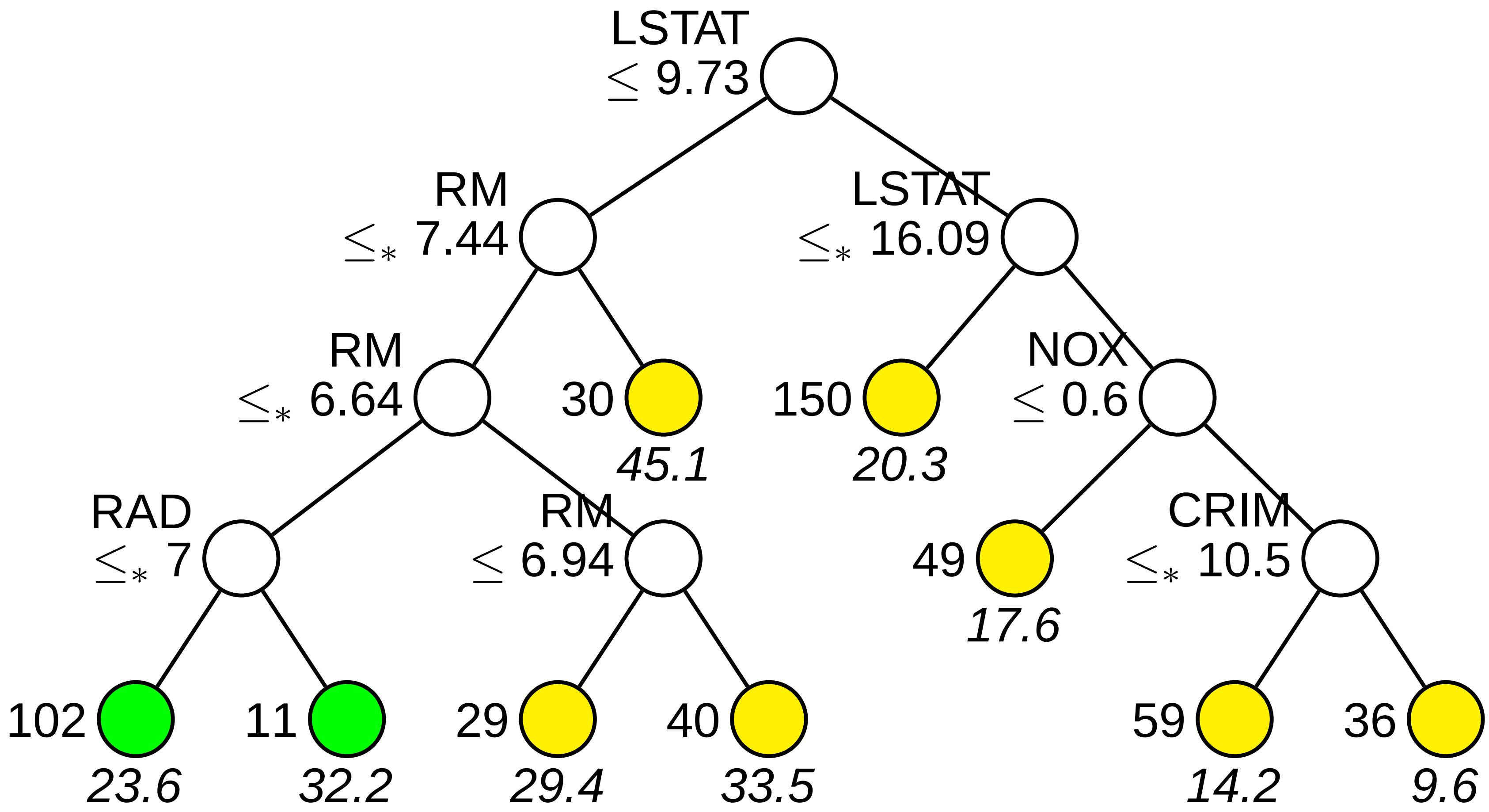
Predicted MEDV values beneath terminal nodes; sample sizes on left

RPART tree for MEDV with TOWN



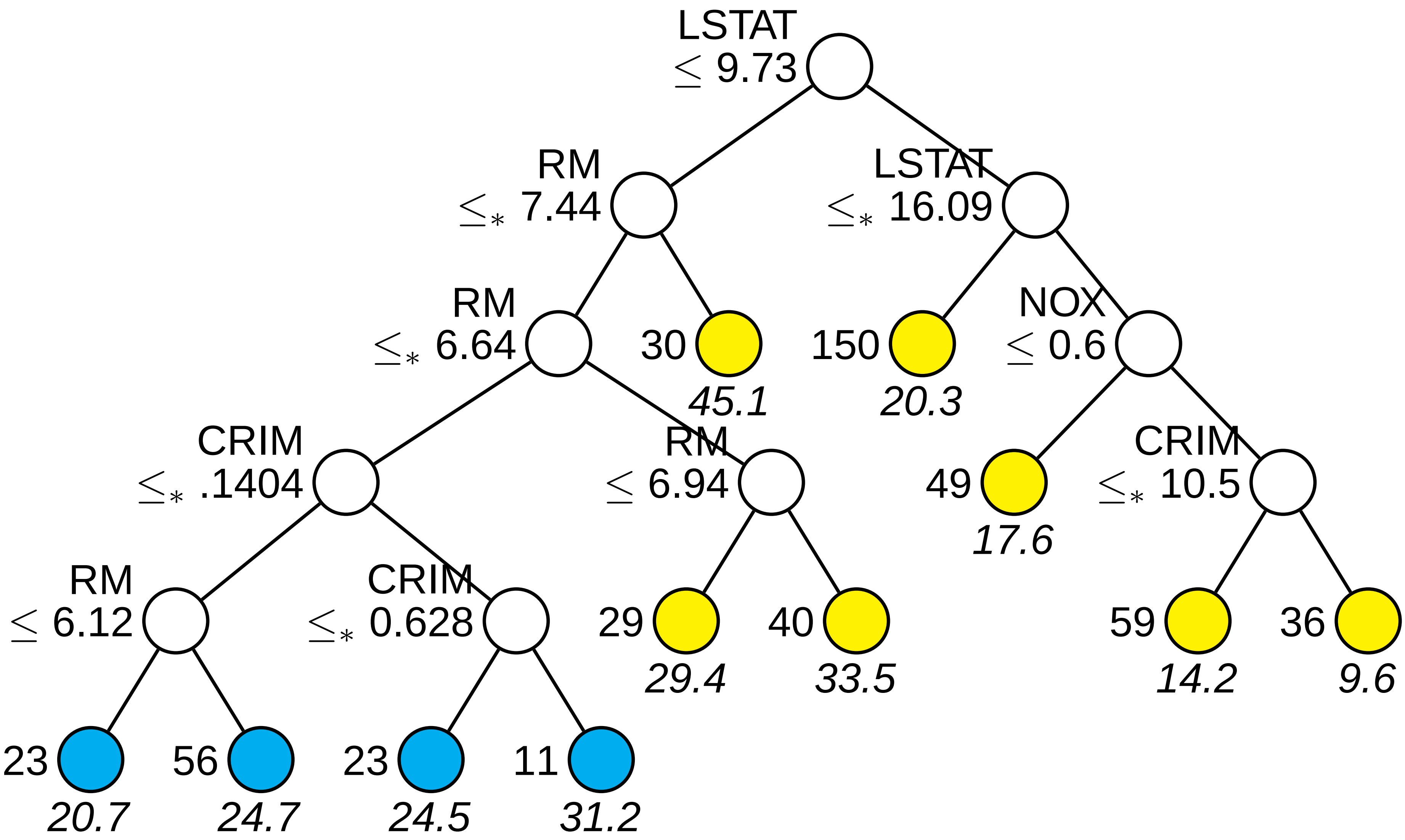
Predicted MEDV values beneath terminal nodes; sample sizes on left

GUIDE tree for MEDV without TOWN



Predicted MEDV values beneath terminal nodes; sample sizes on left

GUIDE tree for MEDV with TOWN



Predicted MEDV values beneath terminal nodes; sample sizes on left

Converting categorical variables to dummy variables is undesirable

1. Transform each X into a 0-1 dummy vector $(U_1, \ldots, U_c)$
2. Use $U_1, \ldots, U_c$ as predictors in model
3. Resulting ANCOVA model
   (a) uses up many degrees of freedom if c is large
   (b) forces all non-categorical predictors to have constant slope coefficients for all values of categorical predictors
4. Splits of the form $U_i = 0$ vs. $U_i = 1$ translates to unappealing “singleton” splits of the form $X = a$ vs. $X \neq a$

GUIDE treatment of categorical predictors

- Categorical predictors can be used for splitting and/or model fitting; ANCOVA models tend to yield shorter trees
- Splits are on subsets of categories

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_1^2-value
3. Do the same for each **c**-variable, using categories to form columns of table
4. Select the variable with the largest χ_1^2 value

Selection bias in linear fit

- Residuals uncorrelated 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

		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 bias correction: basic idea

- Since chi-squared values of **n**-variables are stochastically smaller, scale them with a multiple $\gamma > 1$
- Estimate γ with the bootstrap: randomly permute the Y values and find the γ that yields equal selection probabilities

GUIDE regression in a nutshell

1. Fit a model to the node and use residual signs to form two classes
2. Apply GUIDE classification to select a variable to split node
3. If selected variable is due to a marginal test:

   X **is n or s:** Search all splits of form $X \leq c$ to minimize sum of deviances

   X **is b or c:**

   (a) If 9 or fewer unique X values, search exhaustively

   (b) Otherwise apply GUIDE classification to the two-class problem

4. If selected variables are due to an interaction test, use GUIDE classification to select variable and split set

   See p.132 for Steps 3(b) and 4.

GUIDE approach to missing values for regression

1. A “missing” category is created for each categorical variable for split selection
2. For each split on an ordered variable, missing values are sent to the left or right node, depending on which one reduces node impurity more. The split that sends all missing values to one node and all nonmissing to the other is also considered.
3. For piecewise constant models, only cases complete in the **d**, **w**, **t**, and **z** variables are used for split selection and model fitting
4. For all other models, fitting is restricted to cases complete in the **n** and **f** variables; the node Y mean is fitted to the other cases
5. Bootstrap bias-correction is performed for multiple linear models only

Quantile regression example:
Which colleges are the most expensive?

- Data on 1134 U.S. colleges and universities for year 1995 from *U. S. News & World Report* (`http://lib.stat.cmu.edu/`)
- Response variable is out-of-state tuition
- Goal: Identify the top 10% most expensive colleges, after allowing for various explanatory variables

Explanatory variables for college data

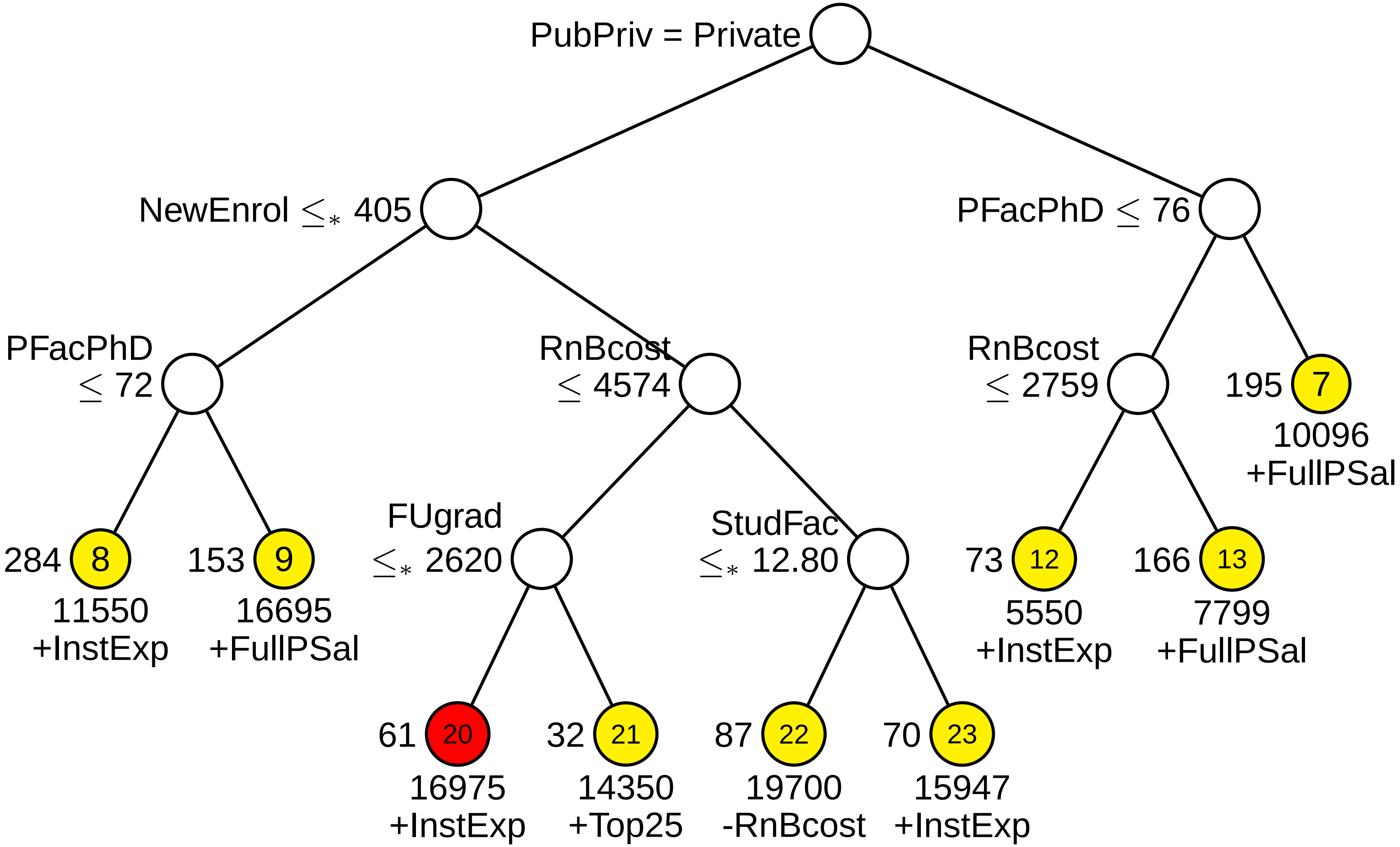
Name	Description	#Missing
PubPriv	Public or private college (binary)	0
CombSAT	Average Combined SAT score	471
AppsRec	Number of applications received	9
AppsAcc	Number of applicants accepted	9
NewEnrol	Number of new students enrolled	5
Top10	Percent new students from top 10% of H.S. class	183
Top25	Percent new students from top 25% of H.S. class	155
FUgrad	Number of fulltime undergraduates	3

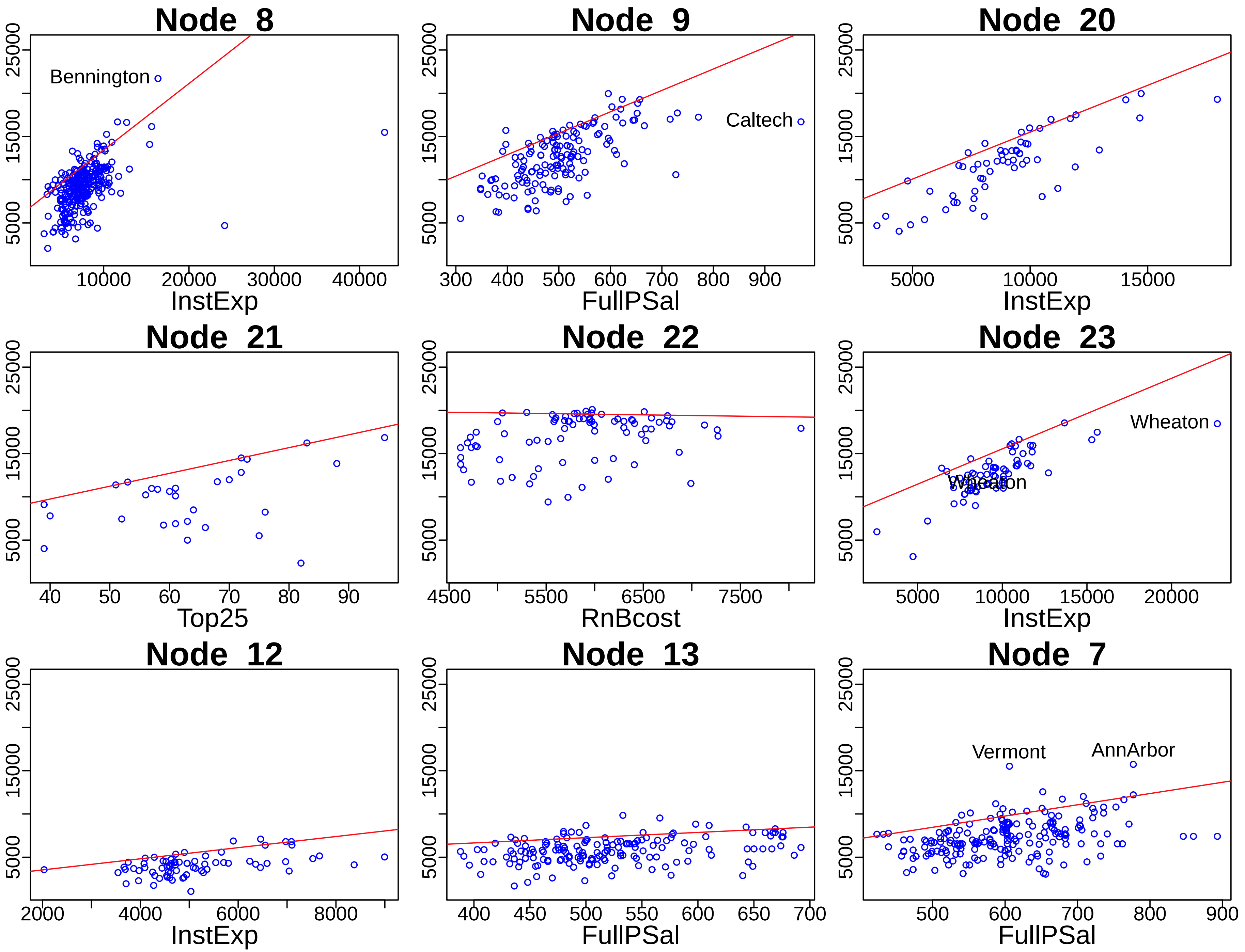
Explanatory variables for college data (cont'd)

Name	Description	#Missing
RnBcost	Room and board costs	57
PFacPhD	Percent of faculty with Ph.D.'s	29
StudFac	Student/faculty ratio	2
InstExp	Instructional expenditure per student	24
GradRate	Graduation rate	69
Type	College type (I: doctoral, IIA: master, or IIB: bachelor)	0
FullPSal	Average salary—full professors (in $100's)	61
NFullProf	Number of full professors	0

513 cases with complete observations

GUIDE simple linear 90th-percentile tree for out-of-state tuition




Node 8
Bennington
InstExp
10000
20000
30000
40000
Node 9
Caltech
FullPSal
300
400
500
600
700
800
900
Node 20
InstExp
5000
10000
15000
Node 21
Top25
40
50
60
70
80
90
Node 22
RnBcost
4500
5500
6500
7500
Node 23
Wheaton
Wheaton
InstExp
5000
10000
15000
20000
Node 12
InstExp
2000
4000
6000
8000
Node 13
FullPSal
400
450
500
550
600
650
700
Node 7
Vermont
AnnArbor
FullPSal
500
600
700
800
900
5000
15000
25000

Subgroup identification for differential treatment effects: an approach to personalized medicine

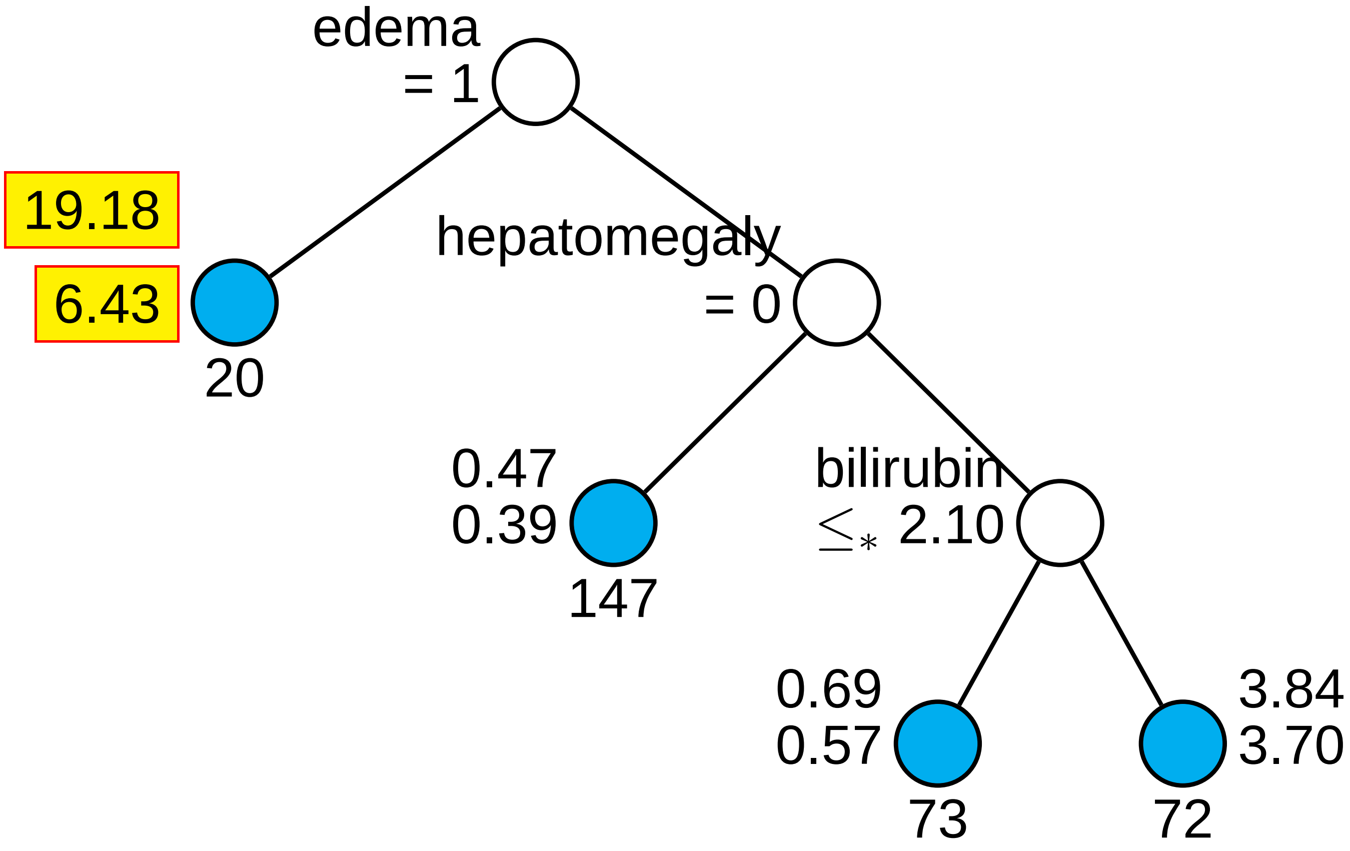
- A piecewise linear model is required for detection of treatment effects
- Piecewise constant models are ineffective because splitting on the treatment variable is useless
- Solution: use the treatment variable as the only linear predictor (after converting to dummy vector)
- Use all other variables for splitting
- Ref: Loh et al. (2013)

Example: primary biliary cirrhosis (PBC) of the liver (Fleming and Harrington, 2005)

- Randomized placebo controlled trial for the drug **D-penicillamine**
- 312 PBC patients, referred to Mayo Clinic during 1974–84
- Response variable is number of days between registration and the earlier of death, liver transplantation, or study analysis time in July, 1986

1	age	days
2	sex	0=male, 1=female
3	presence of ascites	0=no 1=yes
4	presence of hepatomegaly	0=no 1=yes
5	presence of spiders	0=no 1=yes
6	presence of edema	0=no edema and no diuretic therapy for edema
		0.5 = edema present w/o diuretics, or edema resolved by diuretics
		1 = edema despite diuretic therapy
7	serum bilirubin	mg/dl
8	serum cholesterol	mg/dl
9	albumin	gm/dl
10	urine copper	ug/day
11	alkaline phosphatase	U/liter
12	SGOT	U/ml
13	triglicerides	amg/dl
14	platelets	per cubic ml / 1000
15	prothrombin	time in seconds
16	histologic stage of disease	1, 2, 3, 4, 5

GUIDE model for differential treatment effects



- Relative risks of death (drug, upper; placebo, lower) on left of nodes
- Sample sizes beneath nodes

Longitudinal data (Loh and Zheng, 2013)

1. Treat each data point as a curve (trajectory)
2. Fit a mean curve (lowess or smoothing spline) to data in the node
3. Group trajectories into classes according to shapes relative to mean curve
4. For each X variable, find p-value of chi-squared test of class vs. X
5. Select X with smallest p-value to split node
6. For each split point, fit a mean curve to each child node
7. Select the split that minimizes sum of squared deviations (normalized if desired) of trajectories from mean curves in two child nodes
8. Stop splitting when sample size in node is too small
9. Prune the tree using cross-validation

Example: CD4 counts from an AIDS study

- Randomized, double-blind, study of 1309 AIDS patients with advanced immune suppression (Fitzmaurice et al., 2004)

- Four dual or triple combinations of HIV-1 reverse transcriptase inhibitors:

  1: 600mg *zidovudine* alternating monthly with 400mg *didanosine* (dual therapy)

  2: 600mg *zidovudine* + 2.25mg *zalcitabine* (dual therapy)

  3: 600mg *zidovudine* + 400mg *didanosine* (dual therapy)

  4: 600mg *zidovudine* + 400mg *didanosine* + 400mg *nevirapine* (triple therapy)

- CD4 counts at baseline and at 8-week intervals during 40-week follow-up

- Observations during follow-up varied from 1–9, with median of 4 due to: (i) mistiming and (ii) missingness from skipped visits and dropout

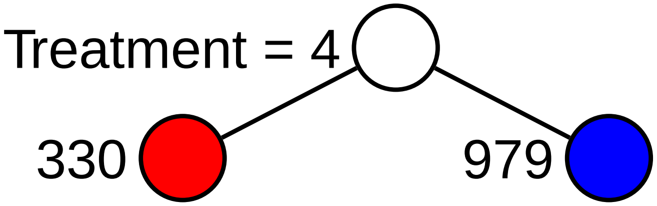
- Response variable is log(CD4 counts + 1); covariates are age and gender

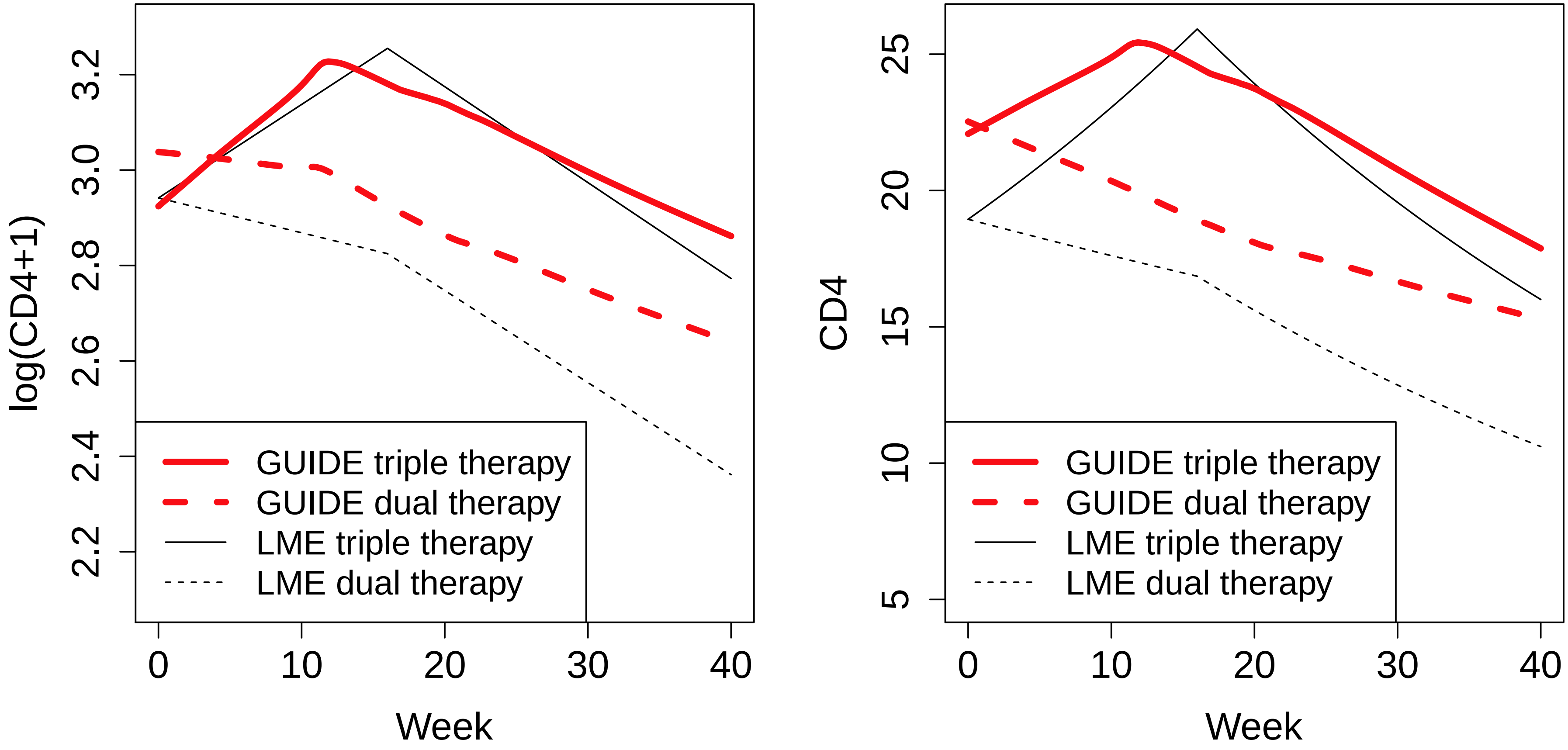
Fitzmaurice et al. (2004) linear mixed effects model

$$\begin{aligned} E(Y_{ij} \,|\, b_i) &= \beta_1 + \beta_2 t_{ij} + \beta_3 (t_{ij} - 16)_+ + \beta_4 I(\texttt{Trt} = 4) \times t_{ij} \\ &\quad + \beta_5 I(\texttt{Trt} = 4) \times (t_{ij} - 16)_+ + b_{1i} + b_{2i} t_{ij} + b_{3i} (t_{ij} - 16)_+ \end{aligned}$$

1. $Y_{ij} = \log(\texttt{CD4}_{ij} + 1)$ for subject i at time t_{ij}
2. All fixed effects significant (p $<$ 0.005)
3. Significant difference in rate of change from baseline to week 16 between dual and triple therapies
4. No sig. difference in rate of change from week 16 to 40 between groups
5. Substantial within and between-patient variability (large random effects)

GUIDE regression tree for AIDS data

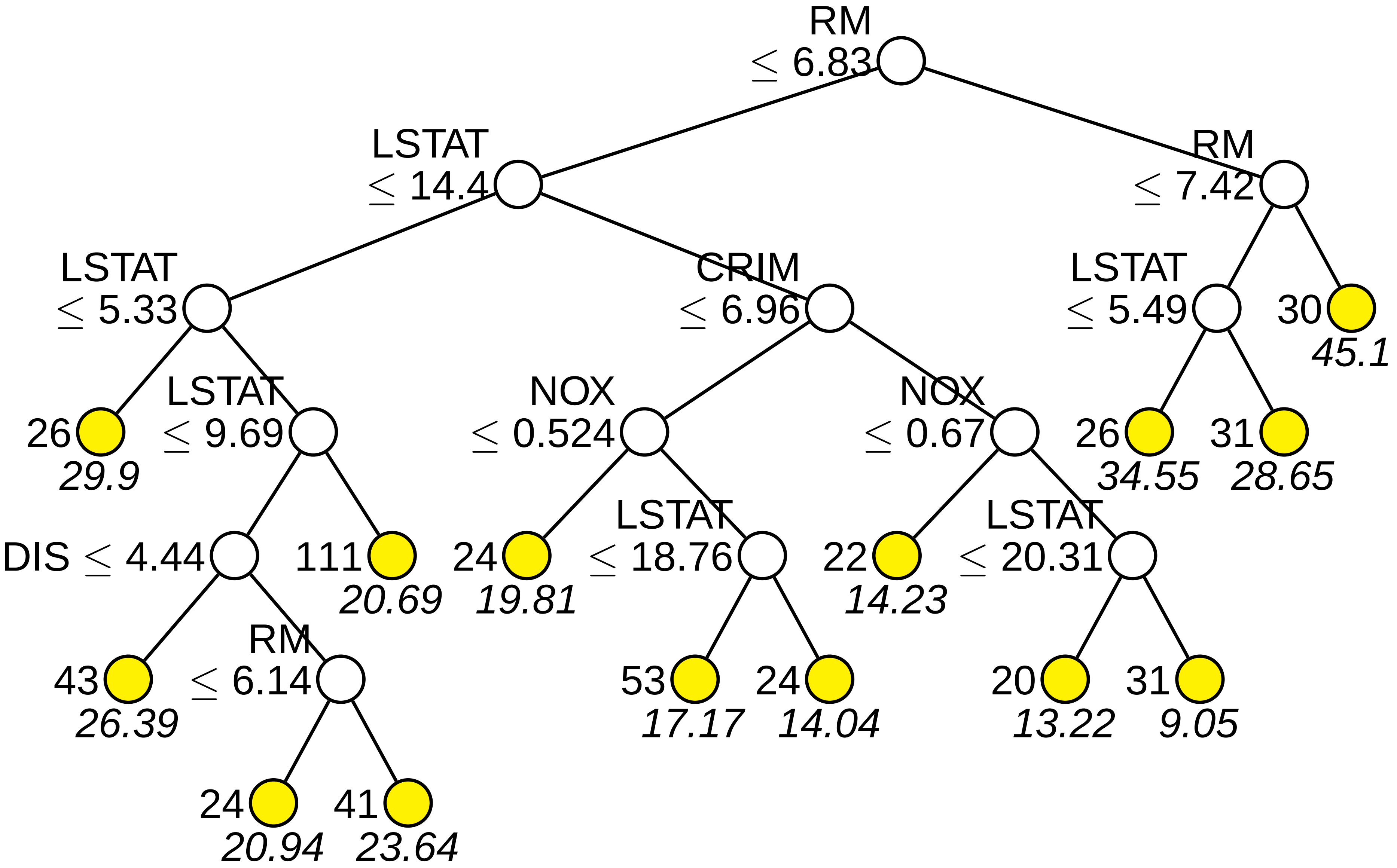




MOB: Model-based recursive partitioning (Zeileis et al., 2008)

1. Fit a model once to data in the current node.
2. Assess whether parameter estimates are stable with respect to each split variable, using Bonferroni-adjusted p-values.
3. If minimum p-value is sufficiently small, select the most unstable variable and split the node into two. Otherwise stop.

MOB for MEDV without TOWN (crashes with TOWN)



Predicted MEDV values beneath terminal nodes; sample sizes on left

M5 regression tree (Quinlan, 1992)

1. **Grow large tree:** Grow a piecewise-constant tree using as reduction in error
$$\frac{m}{n}\left\{SD(t) - \frac{n_L SD(t_L) + n_R SD(t_R)}{n}\right\}$$
where node t (with sample size n) is split into t_L and t_R (with sample sizes n_L, n_R), $SD(t)$ is the sample standard deviation of the cases in t, and m is the number of non-missing values in the split variable

2. **Fit linear models:** After the tree is grown, fit a multiple linear regression model to the cases in each node t, using as regressors only the variables that are selected for splitting in subtree T_t

3. **Estimate error:** Estimate the prediction error of each node t with
$$\text{Err}(t) = \frac{\sum_i |y_i - \hat{y}_i|}{n} \times \frac{n+\nu}{n-\nu}$$
where ν is the number of fitted parameters and n is the sample size in t

4. **Simplify linear models:** Use backward stepwise regression to reduce the number of regressors in each node

5. **Prune tree:** Starting from the bottom, remove branch T_t if Err$(T_t) \geq$ Err(t)

6. **Smooth predicted values:** Let t^* be the parent node of t. Given a case, let its predicted value at t and t^* be $\hat{y}$ and $\hat{y}^*$. The smoothed predicted value is

$$\hat{y}^{**} = (n\hat{y} + k\hat{y}^*)/(n + k)$$

where k is a constant (default value 15). Repeat all the way up to root node.

Categorical predictors in M5

- Each categorical variable is converted to a vector of binary variables
- Suppose categorical variable X takes values $X_1, X_2, \ldots, X_c$.
  1. Order the X values by their sample mean Y-values
  2. Denote the ordered values by $X'_1, X'_2, \ldots, X'_c$
  3. Create binary variables $U_1, U_2, \ldots, U_{c-1}$ such that

     $$U_i = \begin{cases} 0 & \text{if } X \in \{X'_1, \ldots, X'_i\} \\ 1 & \text{otherwise} \end{cases}$$

  4. Replace X by $(U_1, U_2, \ldots, U_{c-1})$
- The conversion is usually carried out only at the root node

Missing values in M5

Training data: Use the Y variable to form a surrogate split: Compare the Y-value of the observation with the mean of the Y-values in the two subnodes

Test data: Replace missing values with means from the training sample in the node

M5 is implemented in Witten et al. (2011) as M5'

Empirical comparison of regression algorithms (Loh et al., 2007)

15 algorithms

- 10 regression tree methods
- 3 ensemble (bagged) methods
- 2 spline methods

52 datasets

- Training sample size from 96 to 21,252
- Number of ordered predictor variables from 1 to 28
- Number of categorical variables from 0 to 6
- Number of variables in model fitting from 3 to 104

15 regression algorithms

GUIDE piecewise simple linear (G1)

GUIDE piecewise simple quadratic (G2)

GUIDE piecewise simple cubic (G3)

GUIDE piecewise multiple linear (Gm)

GUIDE piecewise stepwise linear (Gs)

GUIDE stepwise pairs (Gp)

GUIDE simple ancova (Ga)

Generalized additive model (gam)

Multivariate adaptive splines (mars)

M5 piecewise constant (mc)

M5 piecewise multiple linear (mm)

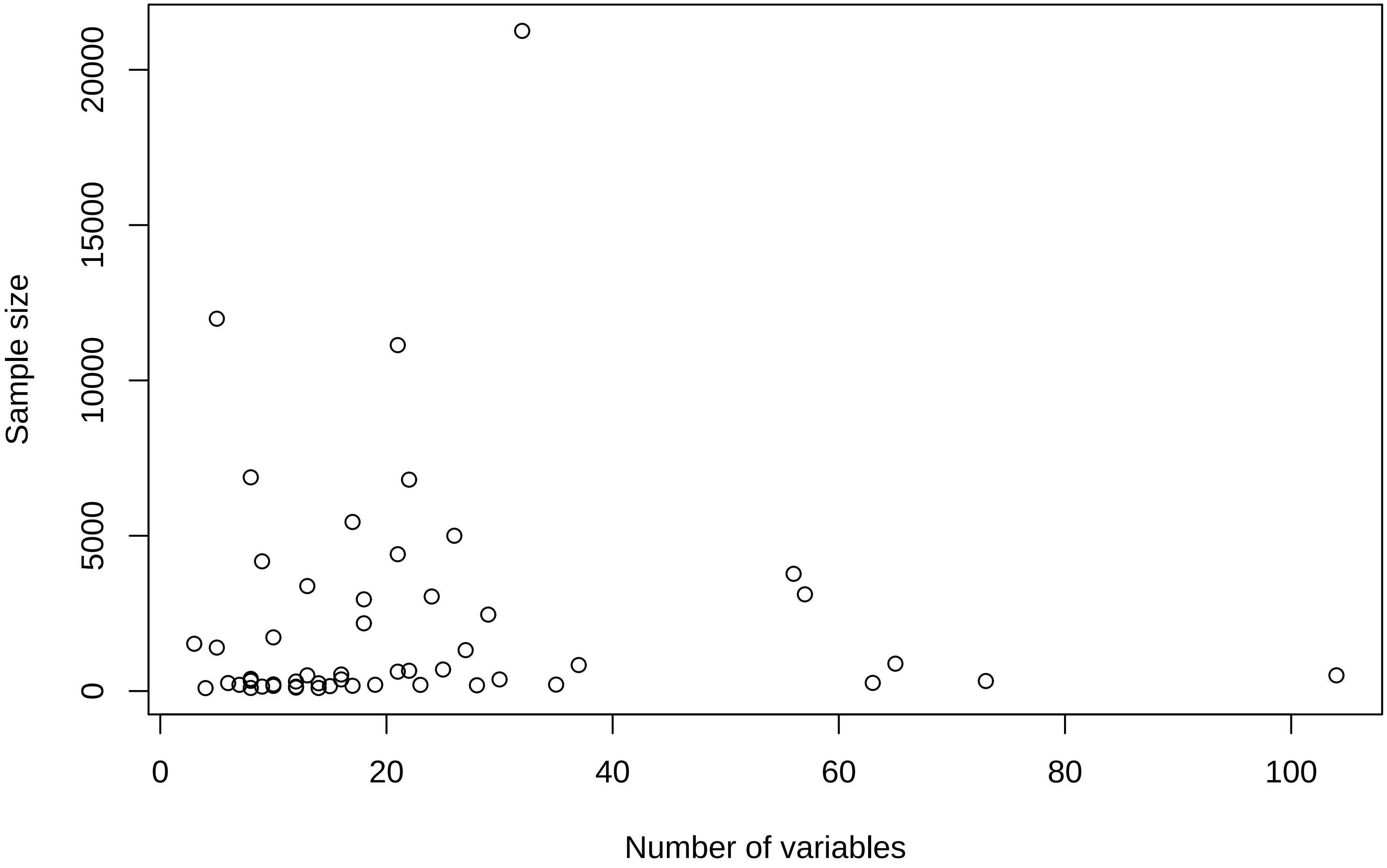
Bagged M5 constant (mcb)

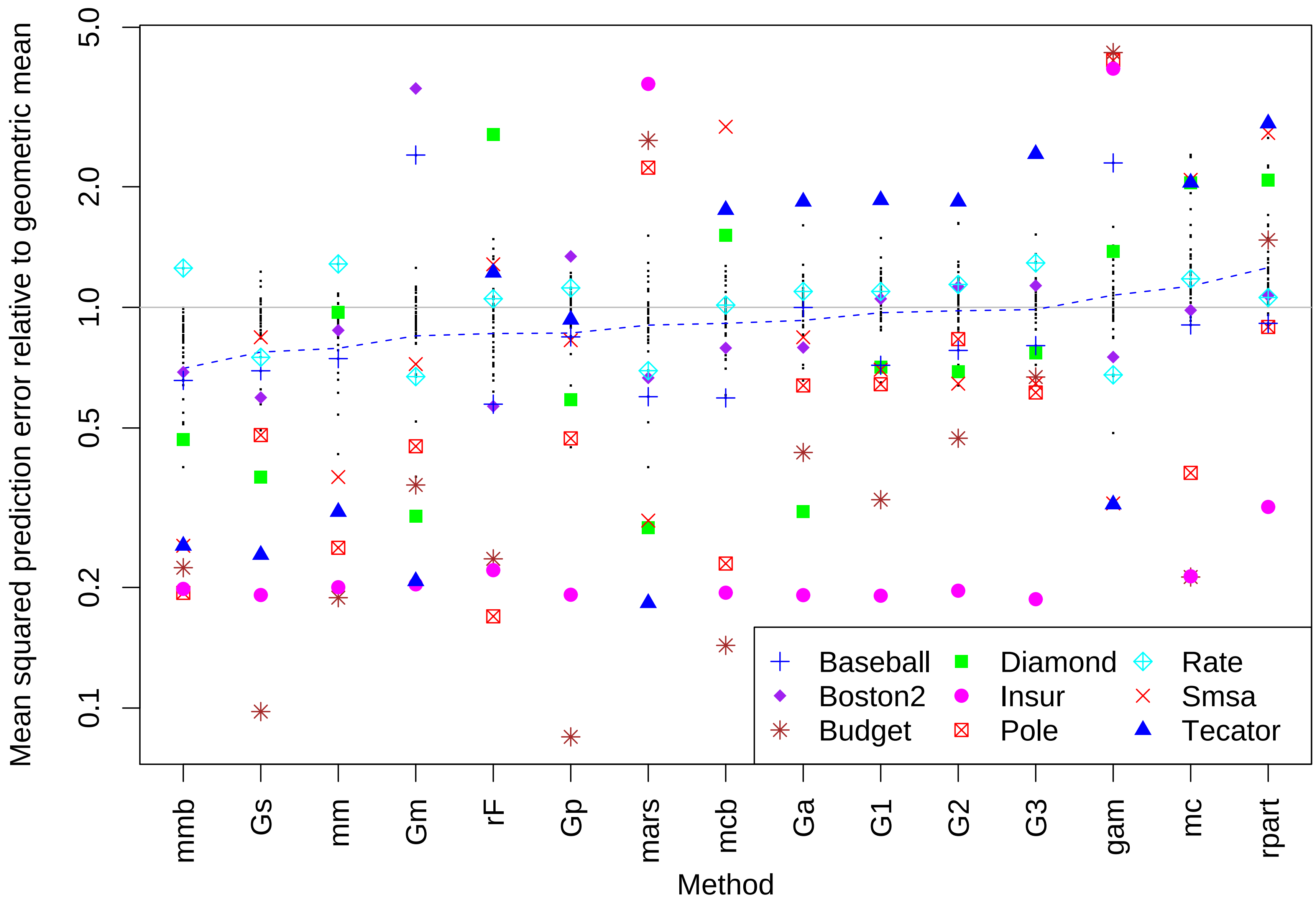
Bagged M5 multiple linear (mmb)

CART clone (rpart)

Random forest (rF)

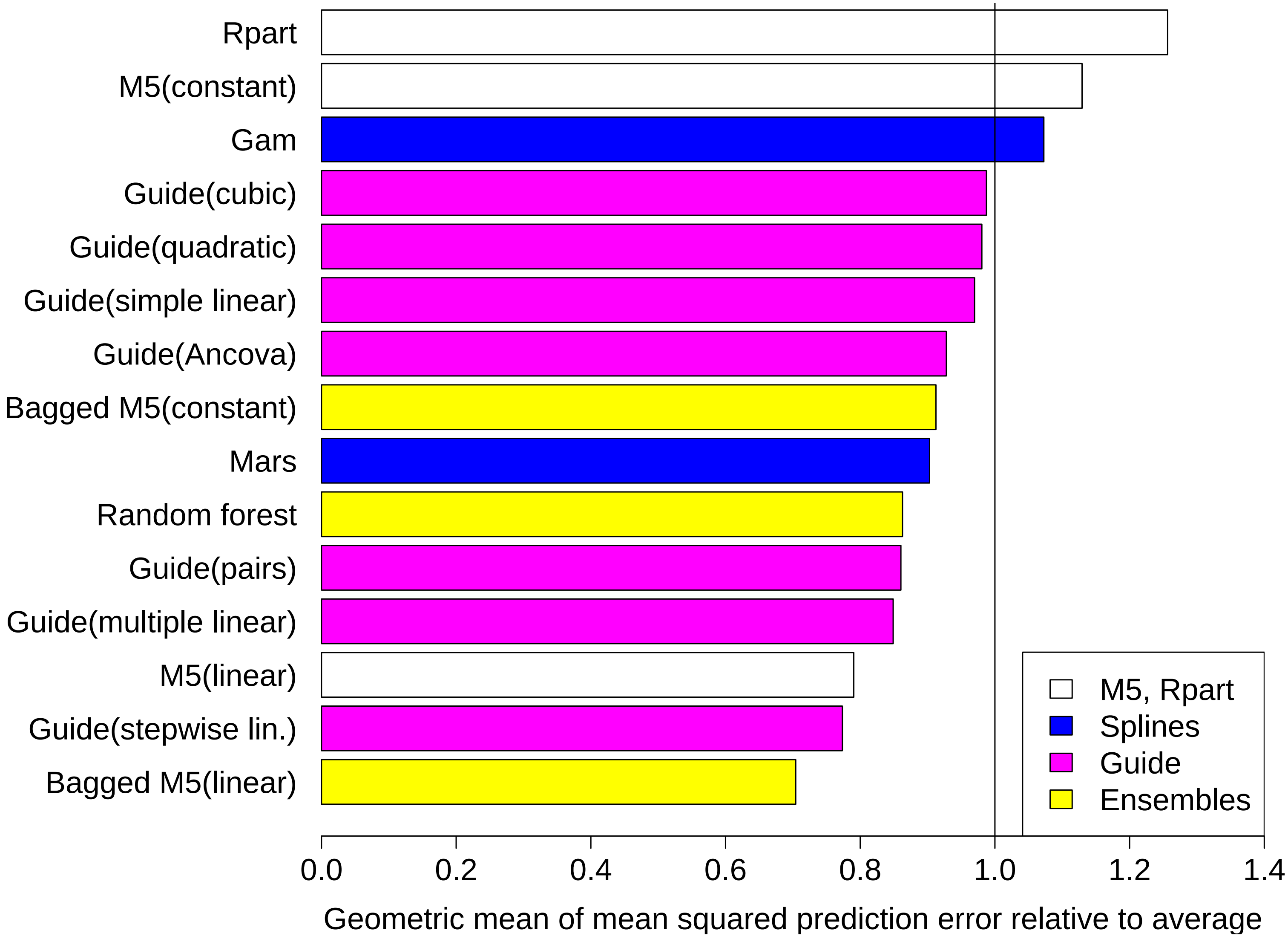
Characteristics of 52 datasets (no missing values)



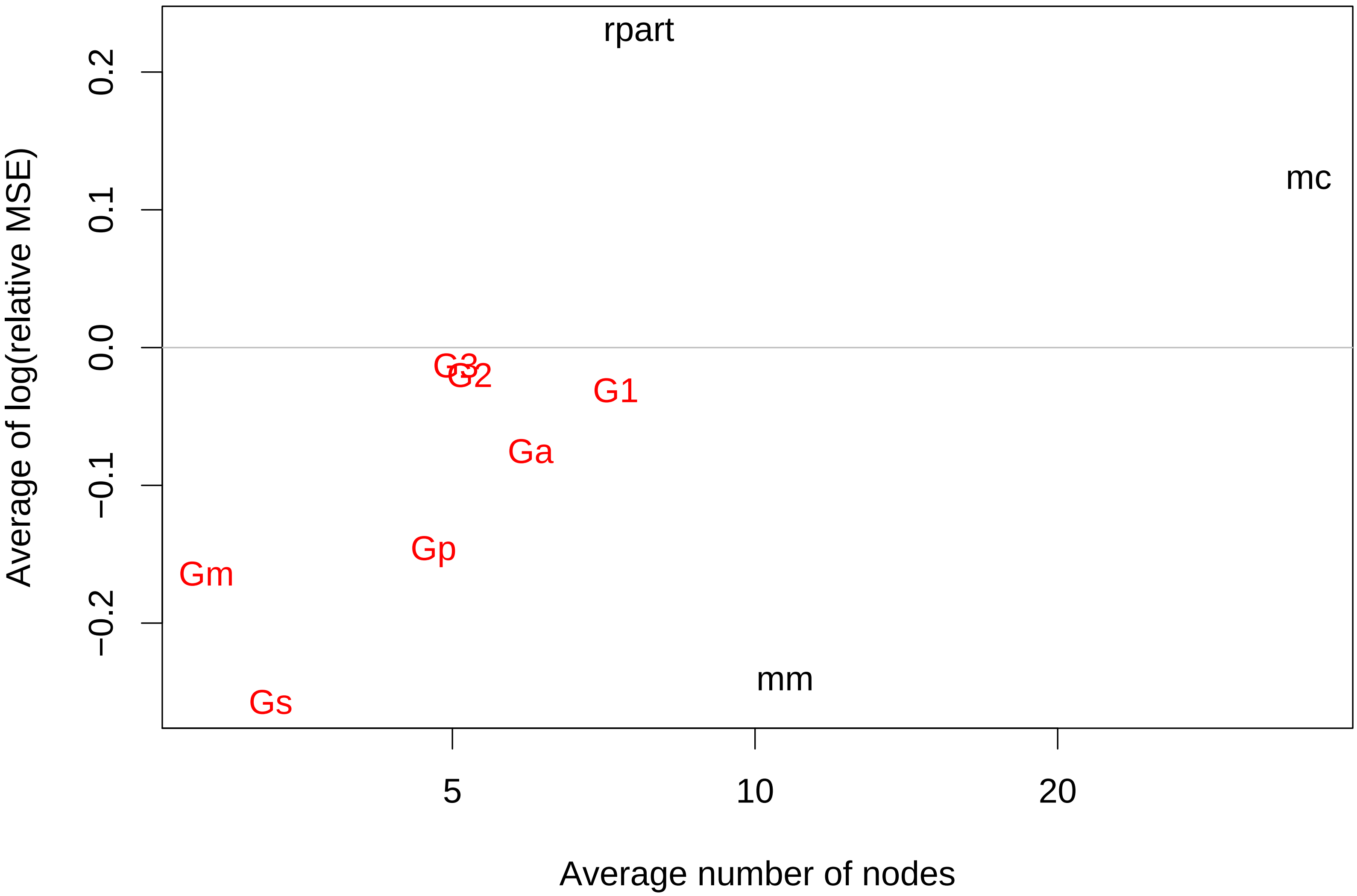


Means and medians joined by dashed and solid lines, respectively

Relative MSE for 52 datasets



Prediction error vs tree size over 52 datasets



Some notations and definitions for asymptotics

- Let $\mathcal{X}$ be M-dimensional Euclidean space.
- Given a fixed integer M_1, let $\mathcal{B}$ be the collection of all polyhedra in $\mathcal{X}$ having at most M_1 faces. These sets can be described as the solutions to at most M_1 inequalities, each inequality having the form $b_1 x_1 + \ldots + b_M x_M \leq c$ (or $< c$).
- If $M_1 \geq 2M$, $\mathcal{B}$ includes all boxes in $\mathcal{X}$ of the form

  $$B = \{(x_1, \ldots, x_M) \;:\; x_1 \in I_1, \ldots, x_M \in I_M\}$$

  where $I_1, \ldots, I_M$ are open, closed, half-open, or half-closed intervals.
- Let $X \in \mathcal{X}$ and $\{(X_i, Y_i) \;:\; i = 1, \ldots, N\}$ be a random sample with the same distribution as (X, Y).
- Given $N \geq 1$ and $t \in \mathcal{X}$, define $\eta_N(t) = \{i \;:\; X_i \in t, 1 \leq i \leq N\}$.
- Let $\tilde{T}_N$ be a partition of $\mathcal{X}$ into a finite number of disjoint sets, all of which are in $\mathcal{B}$, with $\tilde{T}_N$ indpendent of (X, Y).

- Let τ_N denote the partition function corresponding to $\tilde{T}_N$, so that $\tau_N(x)$ is the set $t \in \tilde{T}_N$ containing x.

- Let $\delta(t) = \sup_{x,x' \in t} |x - x'|$ be the diameter of t, where $|x|$ is Euclidean distance.

- Let $D_N(x) = \delta(\tau_N(x))$ be the diameter of the set $t \in \tilde{T}_N$ containing x.

- Let $d_N(x) = \bar{y}_N(\tau_N(x))$ be the estimate of the regression function d_B, where

$$\bar{y}_N(t) = \sum_{i \in \eta_N(t)} Y_i / |\eta_N(t)|.$$

- Let $p_N(t) = N^{-1}|\{i : X_i \in t,\, 1 \leq i \leq N\}|$ be the empirical distribution of X.

- Let k_N be nonnegative constants such that

$$p_N(t) \geq k_N \log(N)/N \text{ for } N \geq 1 \text{ and } t \in \tilde{T}_N.$$

Bayes risk consistency of piecewise-constant regression models (Breiman et al. 1984)

Theorem. Suppose that $E|Y|^q < \infty$ for some $1 \le q < \infty$ and that

$$k_N \to \infty \text{ and } D_N(X) \xrightarrow{P} 0 \text{ as } N \to \infty. \qquad (1)$$

Let $d_B(x) = E(Y \mid X = x)$. Then $E|d_N(X) - d_B(X)|^q \to 0$.

Given any function d on $\mathcal{X}$, let $R(d) = E[Y - d(X)]^2$ denote the mean squared error of $d(X)$.

Theorem. Suppose that $EY^2 < \infty$ and that condition (1) holds. Then $\{d_N\}$ is *risk consistent*, i.e., $ER(d_N) \to R(d_B)$ as $N \to \infty$.

Asymptotic uniform consistency (Kim et al., 2007)

Given $X = x$, let Y have mean $f(x)$. Suppose $f(x)$ is continuous in a compact rectangle C and there is $a > 0$ such that

$$\sup_{x \in C} E\{\exp(a|Y - f(x)|) \mid X = x\} < \infty$$

Let T_n be the regression tree based on training sample size n, m_n = minimum node sample size, and $\delta(t) = \sup_{x,z \in t} \|x - z\|$ be the diameter of node t

Assume that as $n \to \infty$,

1. $(\log n)/m_n \xrightarrow{P} 0$
2. $\sup_{t \in T_n} \delta(t) \xrightarrow{P} 0$
3. Minimum eigenvalue of node design matrices is bounded from 0 in probability

Let $\hat{f}(x)$ be the regression estimate at x. Then

$$\sup_{x \in C} |\hat{f}(x) - f(x)| \xrightarrow{P} 0$$

Conclusions

- Parametric models are often constrained by range restrictions, missing values, distributional assumptions, and number and variety of variables.
- Tree models do not have such constraints.
- When the assumptions hold, parametric models are often more accurate. But when the assumptions are wrong, the results can be very misleading.
- Parametric models depend on statistical inference for model selection. Statistical inference is treacherous when there are many variables.
- Statistical inference is irrelevant to tree models, for which model selection is automatic.
- Tree models can supplement parametric models by validating the assumptions and suggesting alternative functional forms.
- Tree models allow high-level visualization of multivariate data through the tree structures and low-level visualization through plots of terminal nodes.
- Tree models are not necessarily unique. If variable selection is unbiased, each model gives a truthful description of the data.

References

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, New York.

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.

Chambers, J. M. and Hastie, T. J. (1992). An appetizer. In Chambers, J. M. and Hastie, T. J., editors, *Statistical Models in S*, pages 1–12. Wadsworth & Brooks/Cole, Pacific Grove.

Chan, K.-Y. and Loh, W.-Y. (2004). LOTUS: An algorithm for building accurate and comprehensible logistic regression trees. *Journal of Computational and Graphical Statistics*, 13:826–852.

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.

Chaudhuri, P. and Loh, W.-Y. (2002). Nonparametric estimation of conditional quantiles using quantile regression trees. *Bernoulli*, 8:561–576.

Comizzoli, R. B., Landwehr, J. M., and Sinclair, J. D. (1990). Robust materials and processes: Key to reliability. *AT&T Technical Journal*, 69:113–128.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2004). *Applied Longitudinal Analysis*. Wiley, Hoboken, N.J.

Fleming, T. R. and Harrington, D. P. (2005). *Counting Processes and Survival Analysis*. Wiley, New York.

Harrison, D. and Rubinfeld, D. L. (1978). Hedonic prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5:81–102.

Hosmer, Jr., D. W. and Lemeshow, S. (2000). *Applied Logistic Regression*. Wiley, New York, 2nd edition.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: a conditional inference framework. *Journal of Computational and Graphical Statistics*, 15:651–674.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. (2006). Random survival forests. *Annals of Applied Statistics*, 2:841–860.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29:119–127.

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

Kim, H. and Loh, W.-Y. (2003). Classification trees with bivariate linear discriminant node models. *Journal of Computational and Graphical Statistics*, 12:512–530.

Kim, H., Loh, W.-Y., Shih, Y.-S., and Chaudhuri, P. (2007). Visualizable and interpretable regression models with good prediction power. *IIE Transactions*, 39:565–579.

Lemon, S. C., Roy, J., Clark, M. A., Friedman, P. D., and Rakowski, W. (2003). Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. *Annals of Behavioral Medicine*, 26:172–181.

Lock, R. H. (1993). 1993 new car data. *Journal of Statistics Education*, 1(1).

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

Loh, W.-Y. (2006). Regression tree models for designed experiments. In Rojo, J., editor, *The Second Erich L. Lehmann Symposium–Optimality*, volume 49, pages 210–228. Institute of Mathematical Statistics Lecture Notes-Monograph Series.

Loh, W.-Y. (2008a). Classification and regression tree methods. In Ruggeri, F., Kenett, R., and Faltin, F. W., editors, *Encyclopedia of Statistics in Quality and Reliability*, pages 315–323. Wiley, Chichester, UK.

Loh, W.-Y. (2008b). Regression by parts: Fitting visually interpretable models with GUIDE. In Chen, C., Härdle, W., and Unwin, A., editors, *Handbook of Data Visualization*, pages 447–469. Springer.

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

Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1:14–23.

Loh, W.-Y. (2013). Fifty years of classification and regression trees (with discussion). *International Statistical Review*. In press.

Loh, W.-Y., Chen, C.-W., and Zheng, W. (2007). Extrapolation errors in linear model trees. *ACM Trans. Knowl. Discov. Data*, 1(2):6.

Loh, W.-Y., He, X., and Man, M. (2013). A regression tree approach to subgroup identification. Manuscript.

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.

Merkle, E. C. and Shaffer, V. A. (2011). Binary recursive partitioning: background, methods, and application to psychology. *British Journal of Mathematical and Statistical Psychology*, 64:161–181.

Messenger, R. and Mandell, L. (1972). A modal search technique for predictive nominal scale multivariate analysis. *Journal of the American Statistical Association*, 67:768–772.

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.

Morgan, J. N. and Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58:415–434.

Murnane, R. J., Boudett, K. P., and Willett, J. B. (1999). Do male dropouts benefit from obtaining a GED, postsecondary education, and training? *Evaluation Reviews*, 23:475–502.

Quinlan, J. R. (1992). Learning with continuous classes. In *5th Australian Joint Conference on Artificial Intelligence*, pages 343–348.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

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

Segal, M. R., Cummings, M. P., and Hubbard, A. E. (2001). Relating amino acid sequence to phenotype: Analysis of peptide-binding data. *Biometrics*, 57(2):632–643.

Singer, J. D. and Willett, J. B. (2003). *Applied Longitudinal Data Analysis*. Oxford University Press, New York, NY.

Strobl, C., Malley, J., and Tutz, G. (2009). An introduction to recursive partitioning: rationale, application and characteristics of classification and regression trees, bagging and random forests. *Psychological Methods*, 14(4):323–348.

Therneau, T. and Atkinson, E. (2013). An introduction to recursive partitioning using the RPART routines. Technical report, Mayo Clinic, Division of Biostatistics. http://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf.

Therneau, T. M. and Atkinson, B. (2012). *RPART: Recursive partitioning*. R package version 3.1-51.

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.

Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, CA, third edition.

Yeh, I.-C. (2007). Modeling slump flow of concrete using second-order regressions and artificial neural networks. *Cement and Concrete Composites*, 29:474–480.

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