STAT 761 (October 4, 2012)
Decision Trees for Multivariate Analysis

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

Office hours: 2:30–3:30pm W, 1217 MSC

Class email list: stat761-1-f12@lists.wisc.edu

Course objectives

1. Introduce a nonparametric approach to data analysis using decision trees
2. Emphasize prediction accuracy and interpretability instead of parametric inference
3. Contrast this with the traditional model-based approach that relies on significance tests
4. Review and compare some popular decision tree algorithms
5. Demonstrate the capabilities of the GUIDE software

Grades based on

1. Regular class attendance
2. A small number of data analysis homework problems
3. One approved data analysis project or a review of one journal article from:
   - `www.stat.wisc.edu/~loh/bib.html` or
   - `www.stat.wisc.edu/~loh/apps.html`

Classification and regression tree algorithms

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 trees—GUIDE, M5
5. Quantile regression trees—GUIDE
6. Poisson regression trees—GUIDE
7. Proportional hazards regression trees—GUIDE
8. Regression trees for multivariate and longitudinal response data—GUIDE
9. Logistic regression trees—LOTUS

Books, theses and papers

- Breiman, Friedman, Olshen and Stone (1984). *Classification and Regression Trees*, CRC Press, (CART, RPART)
- Quinlan (1993). *C4.5: Programs for Machine Learning*, Morgan Kaufmann
- Witten and Frank (2011). *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed., Morgan Kaufmann, (WEKA)
- University of Wisconsin PhD theses: Vanichsetakul (1986), Huang (1989), Ahn (1992), Kademan (1993), Lo (1993), Shih (1993), Yang (1993), Yao (1994), Yan (1995), Potter (1998), Kim (1998), Chan (2000), Gai (2000), Cho (2002), Song (2005), Chang (2008), Chen (2008), Zheng (2009), Wu (2011), He (2012)
- G. V. Kass (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, **29**, 119–127, (CHAID)

Statistical techniques used in the course

1. R-graphics (boxplots, barplots, scatterplots, stripcharts, contour plots)
2. Contingency tables and chi-squared test
3. ANOVA, ANCOVA, and linear mixed models
4. Weighted least squares, least median of squares, quantile, logistic, Poisson, and proportional hazards regression
5. Linear and quadratic discriminant analysis
6. Principal component analysis
7. Density estimation (kernel, nearest-neighbor) and K-means clustering
8. Bootstrap and cross-validation
9. Box-Cox transformations
10. Satterthwaite and Wilson-Hilferty approximations of chi-squared distributions

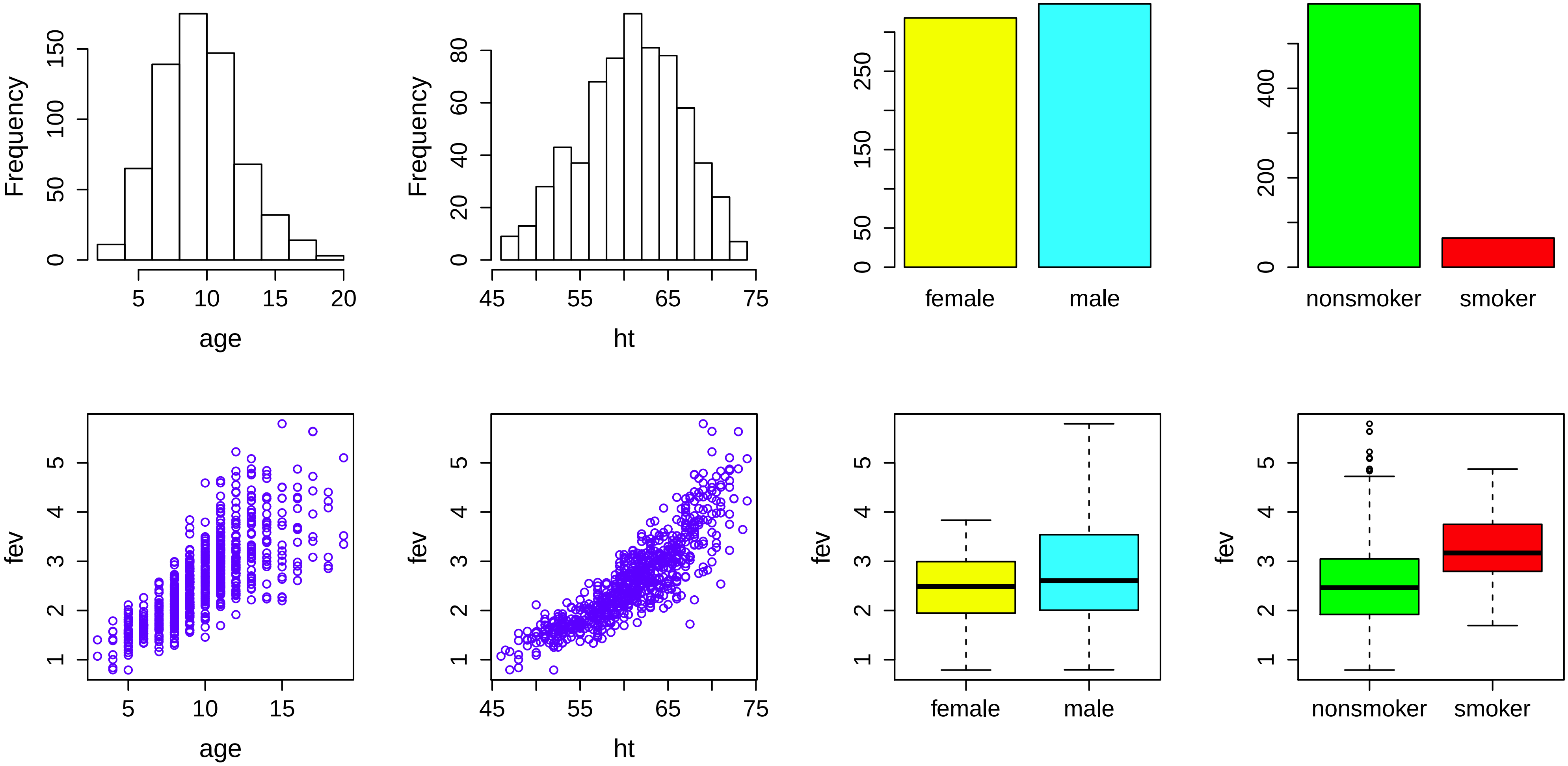
Free software

- CRUISE, GUIDE, LOTUS, QUEST—`www.stat.wisc.edu/~loh/`
- C4.5—`www.rulequest.com/Personal/c4.5r8.tar.gz` and `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/`
- RPART, RandomForest, PARTY—`cran.us.r-project.org/`
- LaTeX (text processing package)—CRUISE, GUIDE, LOTUS, QUEST produce tree diagrams in LaTeX format. PC version from `www.miktex.org/`

Difficulties of linear regression: Smoking and pulmonary function in children (Kahn, 2005)

- Forced expiratory volume (*FEV*, in liters) from 654 children aged 3–19 years
- Predictor variables are *age* (years), *ht* (height in inches), *sex* (0=female, 1=male), and *smoke* (0=nonsmoker, 1=smoker)

Distributions of predictor variables



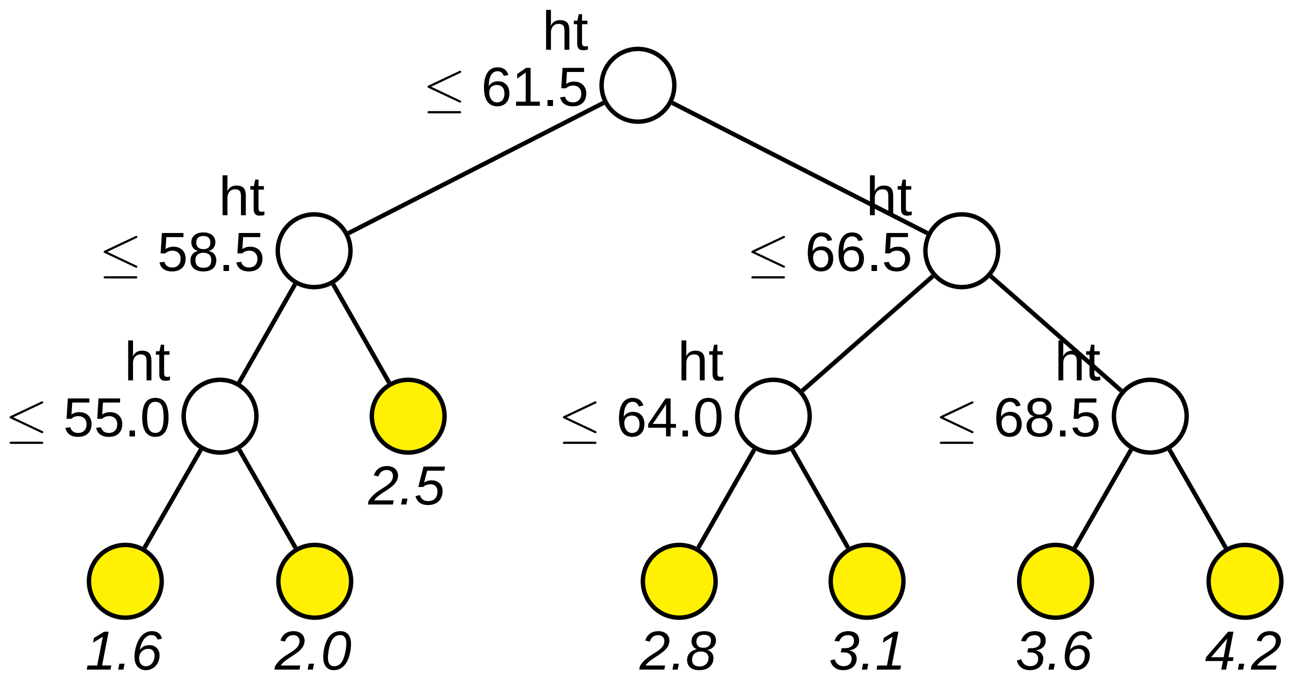
Some linear regression models

Model		Estimate	Std. error	t-value	p-value	Adj. R^2
1	smoke	0.7107	0.1099	6.46	1.99e-10	0.059
2	age	0.2306	0.0082	28.18	$<$ 2e-16	0.575
	smoke	-0.2090	0.0807	-2.59	0.00986	
3	age	0.0655	0.0095	6.90	1.21e-11	0.774
	ht	0.1042	0.0048	21.90	$<$ 2e-16	
	sex	0.1571	0.0332	4.73	2.74e-06	
	smoke	-0.0872	0.0593	-1.47	0.141	
4	age	0.0695	0.0091	7.63	8.66e-14	0.792
	ht	-0.2742	0.0497	-5.52	4.92e-08	
	ht^2	0.0031	0.0004	7.65	7.35e-14	
	sex	0.0945	0.0329	2.88	0.00415	
	smoke	-0.1332	0.0571	-2.33	0.01997	
5	age	0.0745	0.0099	7.51	1.95e-13	0.793
	ht	-0.2795	0.0498	-5.61	3.01e-08	
	ht^2	0.0032	0.0004	7.71	4.72e-14	
	sex	0.0979	0.0330	2.97	0.00308	
	smoke	0.2555	0.3089	0.83	0.40839	
	age:smoke	-0.0295	0.0230	-1.28	0.20080	

Correlations

	age	height	sex	smoke
age	1.00	0.79	0.03	0.40
ht	0.79	1.00	0.16	0.28
sex	0.03	0.16	1.00	-0.08
smoke	0.40	0.28	-0.08	1.00

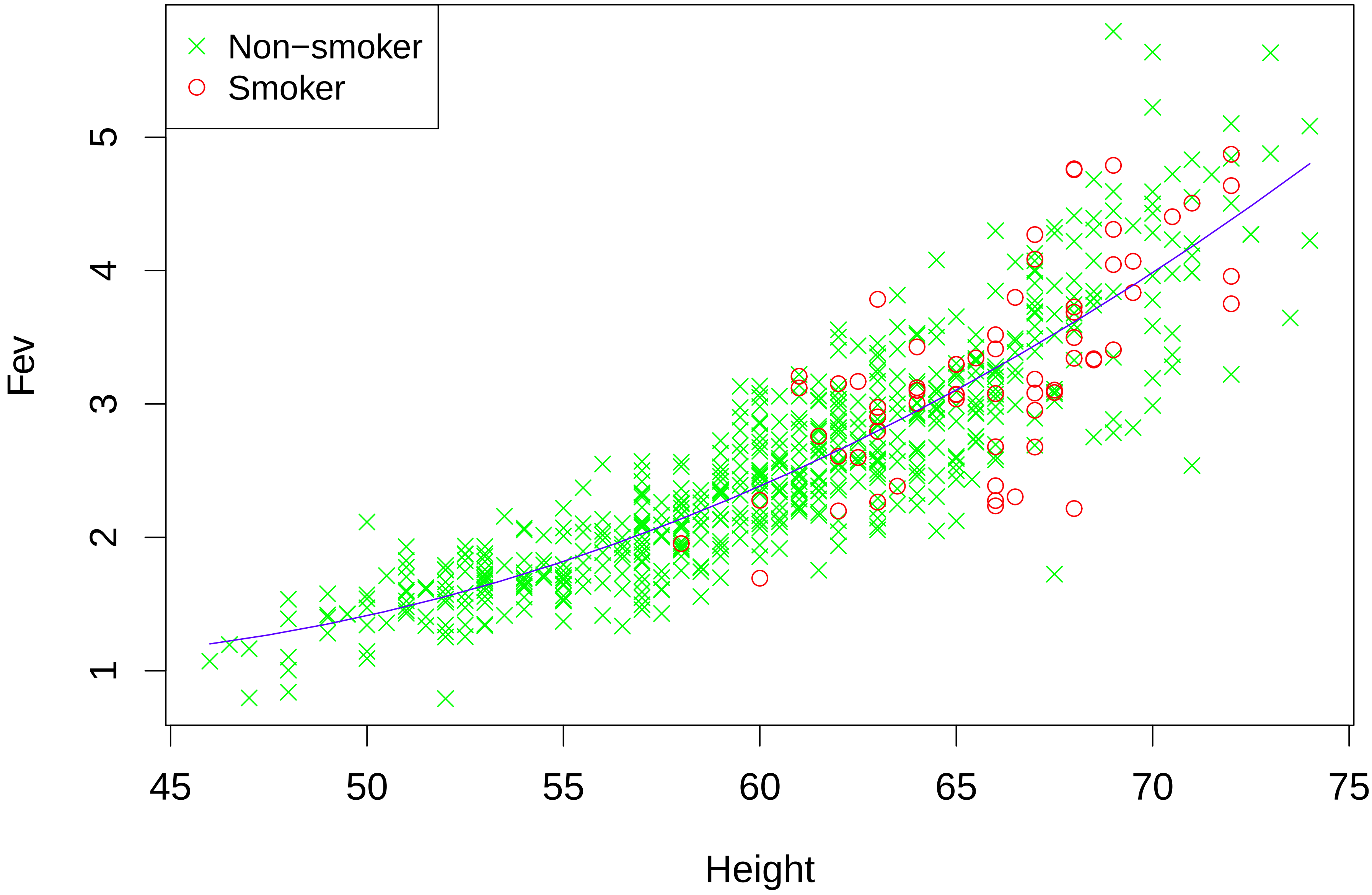
GUIDE piecewise constant model



Fev mean below terminal nodes

Smoke not significant in piecewise linear (in smoke) model

Fev vs. Height with quadratic fit (smoker in red)



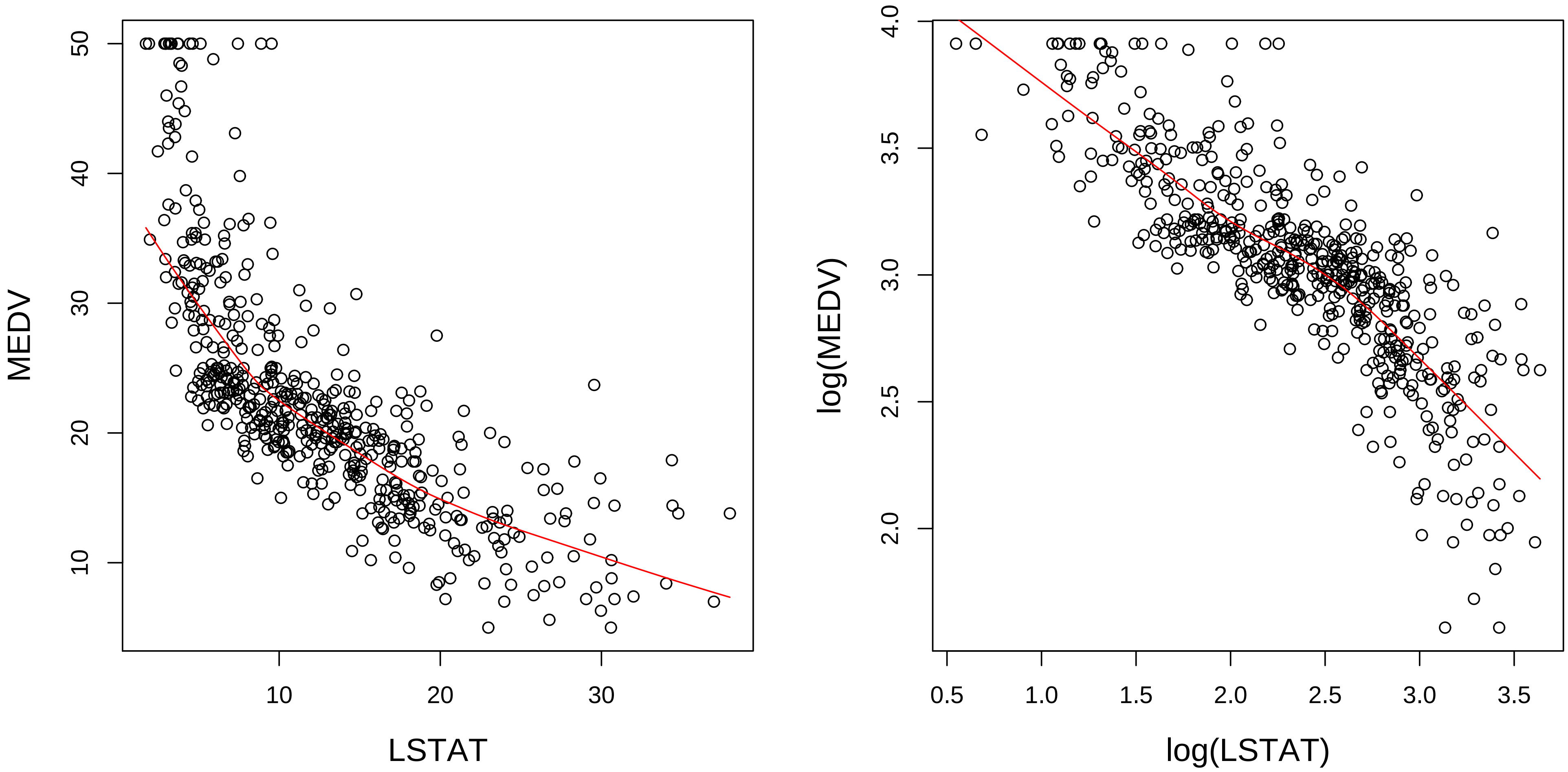
Linear regression: 1970 Boston housing data (Harrison and Rubinfeld, 1978; Belsley et al., 1980)

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

MEDV vs. LSTAT and log(MEDV) vs. log(LSTAT)



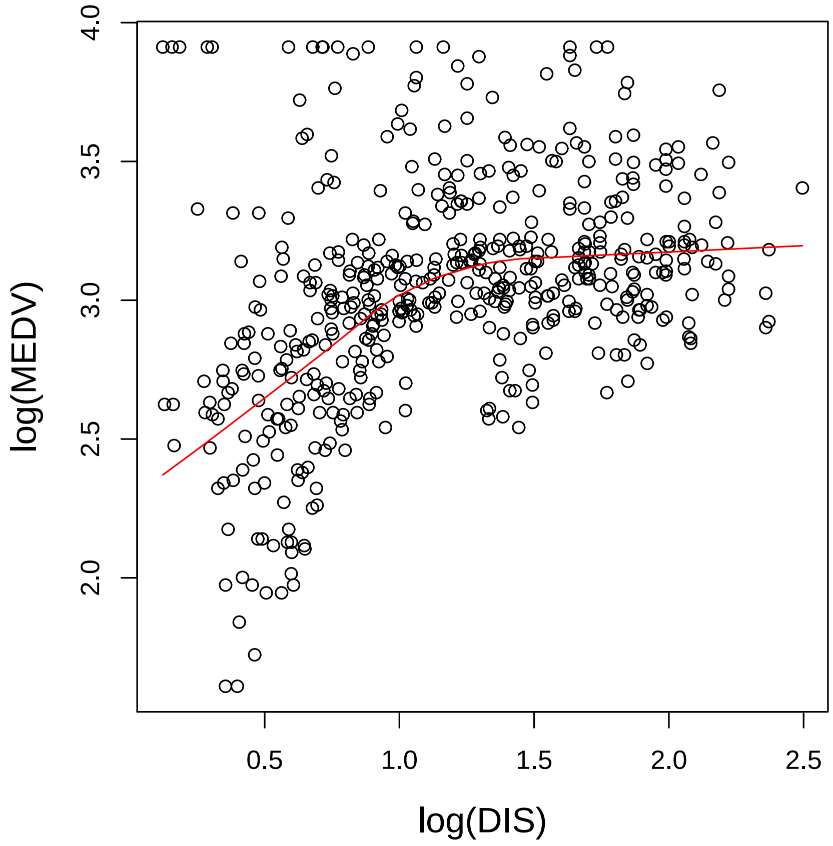
Harrison & Rubinfeld model for log(MEDV)

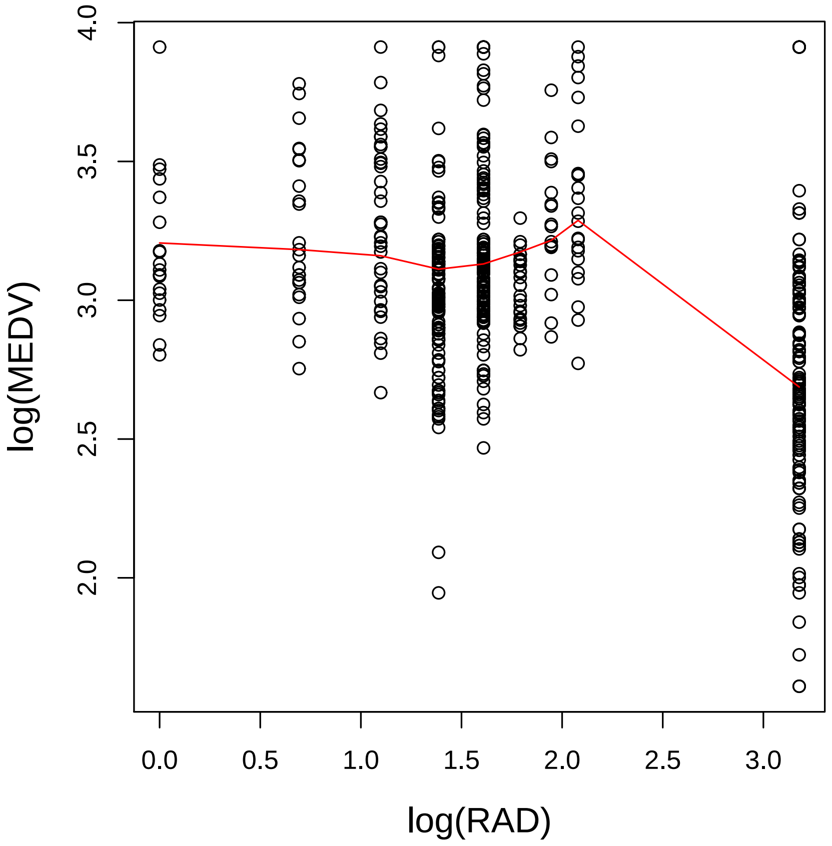
X	β	t	ρ	X	β	t	ρ
Constant	4.6	30.0		AGE	7.1E-5	0.1	-0.5
CRIM	-1.2E-2	-9.6	-0.5	log(DIS)	-2.0E-1	-6.0	0.4
ZN	9.2E-5	0.2	0.4	log(RAD)	9.0E-2	4.7	-0.4
INDUS	1.8E-4	0.1	-0.5	TAX	-4.2E-4	-3.5	-0.6
CHAS	9.2E-2	2.8	0.2	PT	-3.0E-2	-6.0	-0.5
NOX^2	-6.4E-1	-5.7	-0.5	B	3.6E-4	3.6	0.4
RM^2	6.3E-3	4.8	0.6	log(LSTAT)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

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

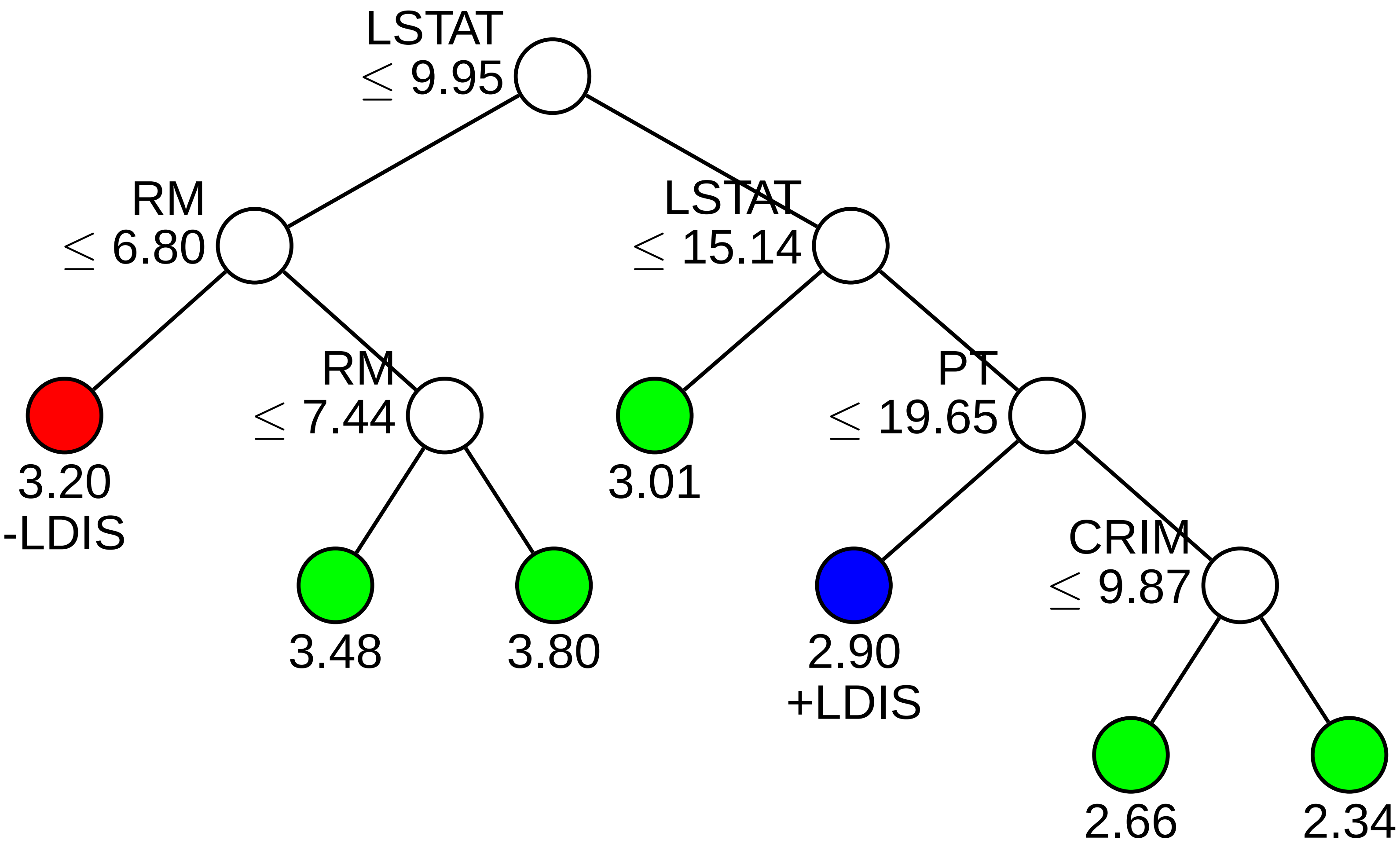
What can we conclude from this model?

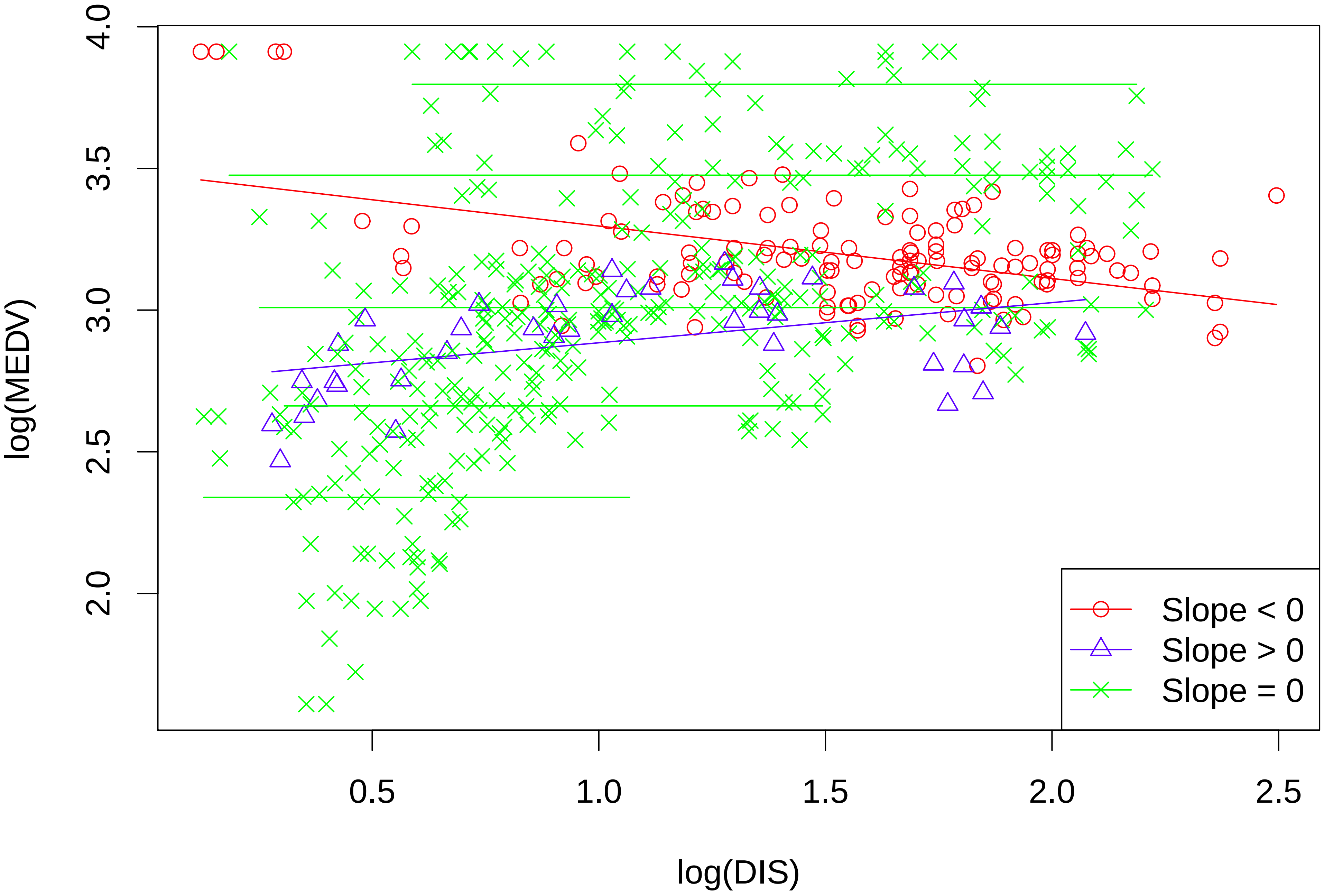
log(MEDV) vs. log(DIS) and log(RAD)



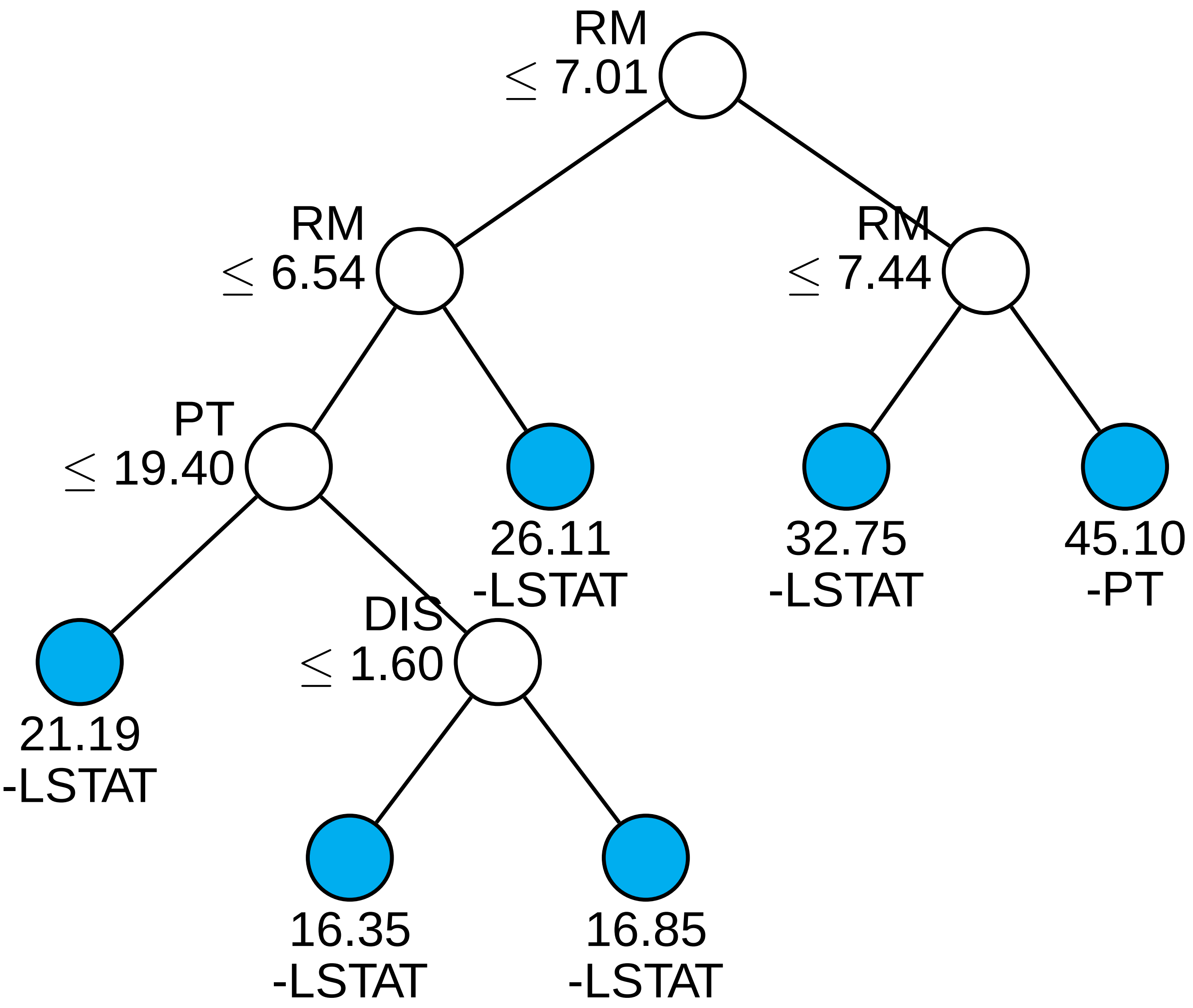


Model for log(MEDV) with log(DIS) as linear predictor

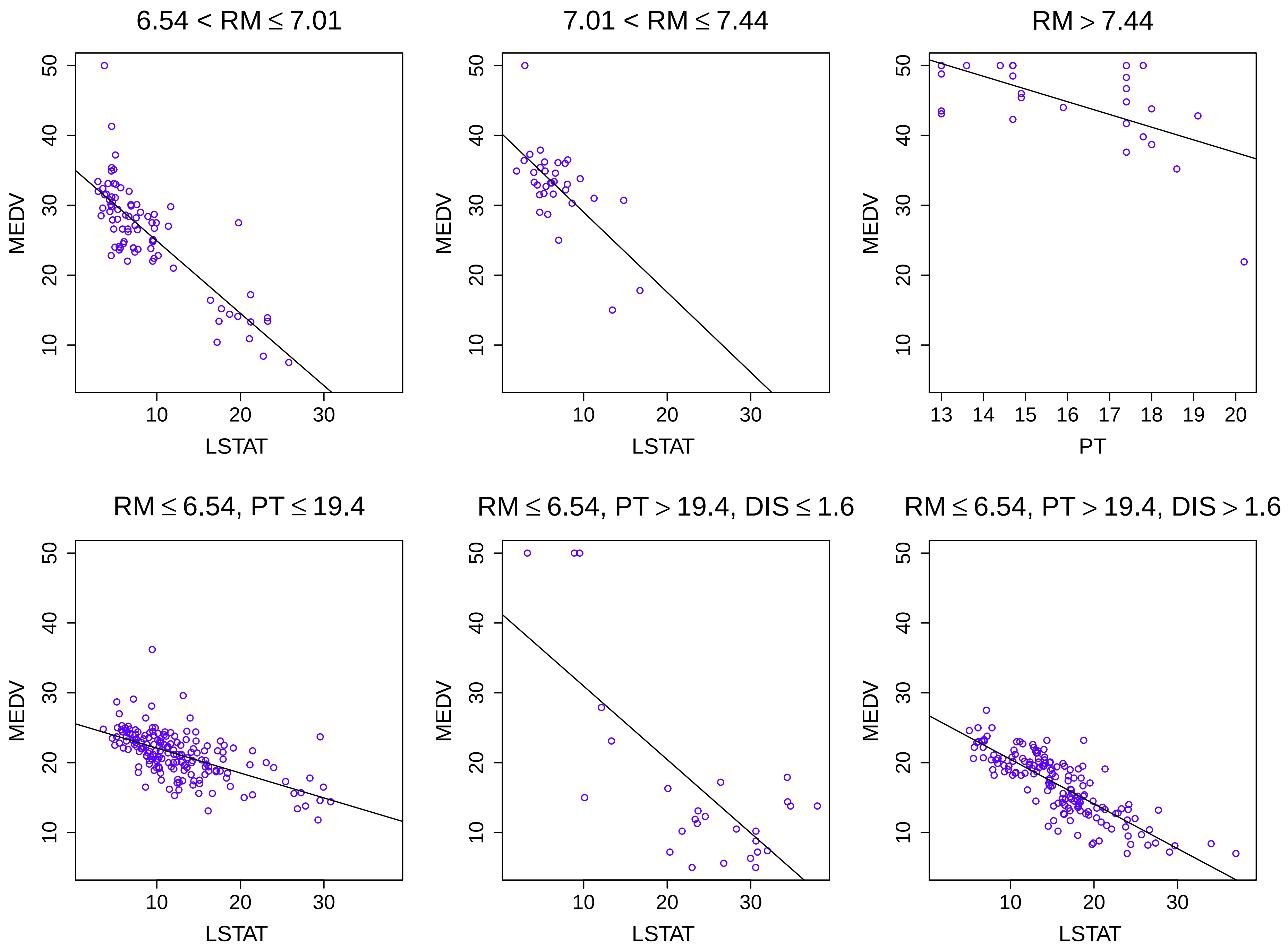



log(MEDV)
4.0
3.5
3.0
2.5
2.0
log(DIS)
0.5
1.0
1.5
2.0
2.5
Slope < 0
Slope > 0
Slope = 0

GUIDE piecewise simple linear model for MEDV



Mean MEDV and signed linear predictor beneath each node



6.54 < RM ≤ 7.01

7.01 < RM ≤ 7.44

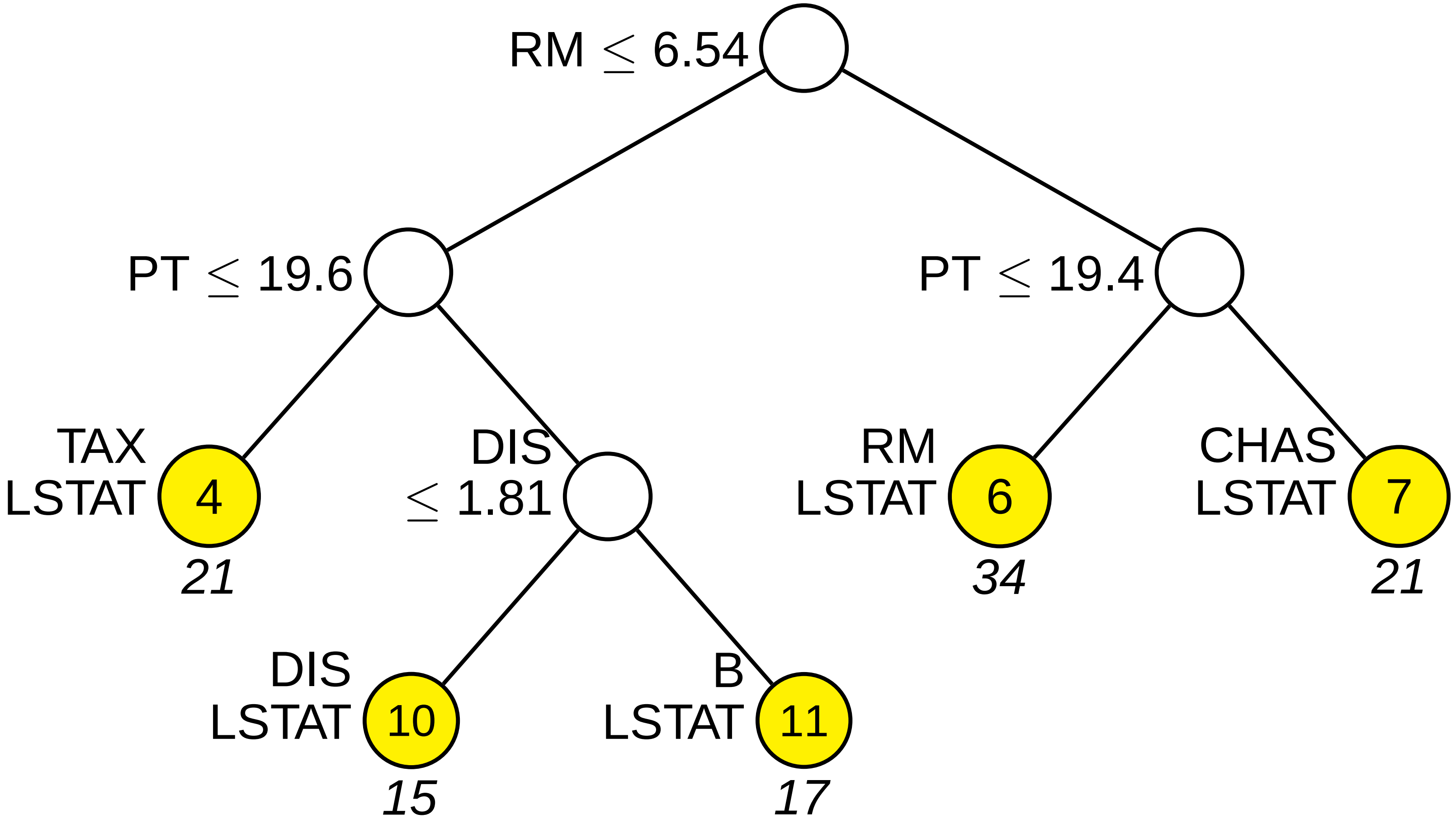
RM > 7.44

RM ≤ 6.54, PT ≤ 19.4

RM ≤ 6.54, PT > 19.4, DIS ≤ 1.6

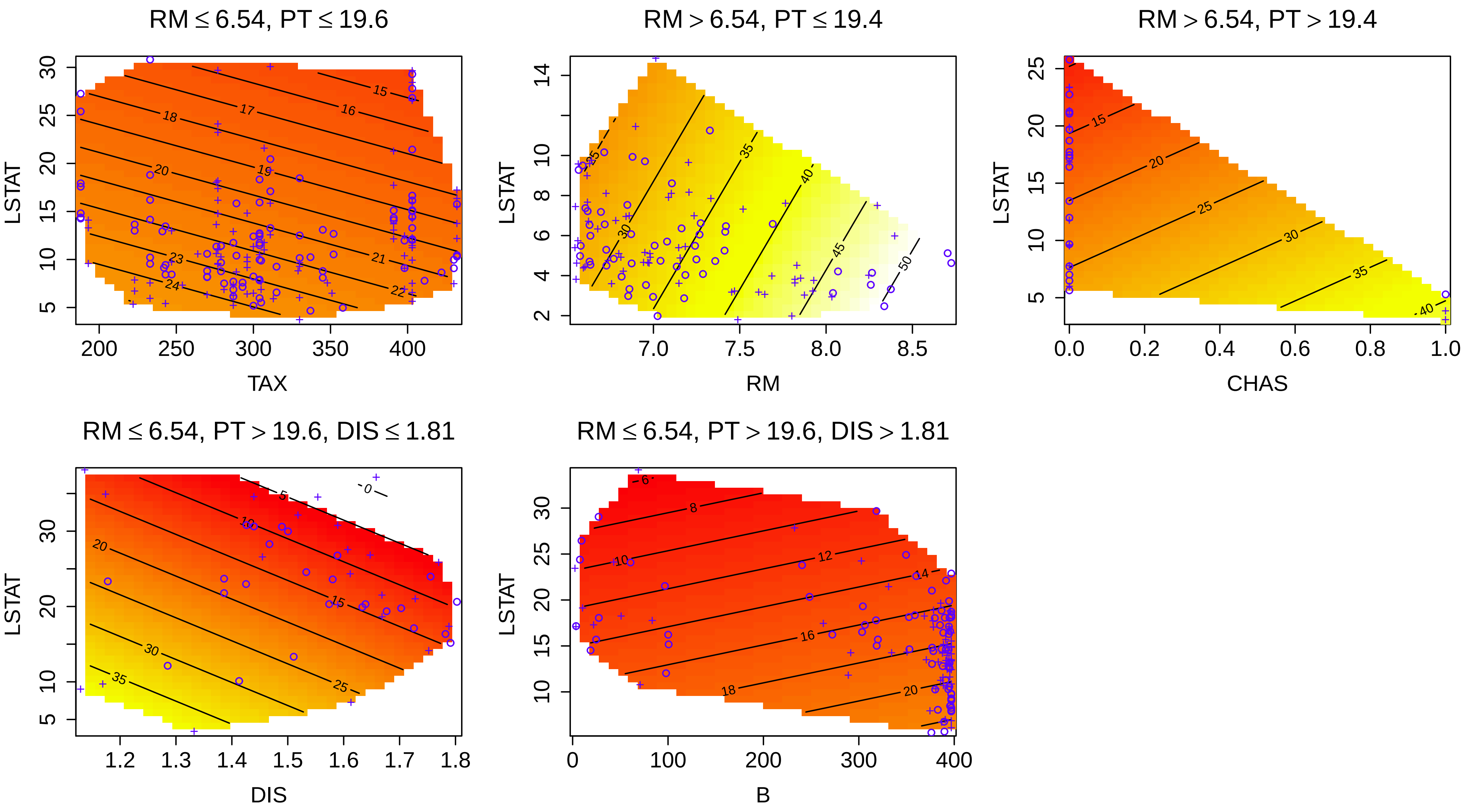
RM ≤ 6.54, PT > 19.4, DIS > 1.6

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

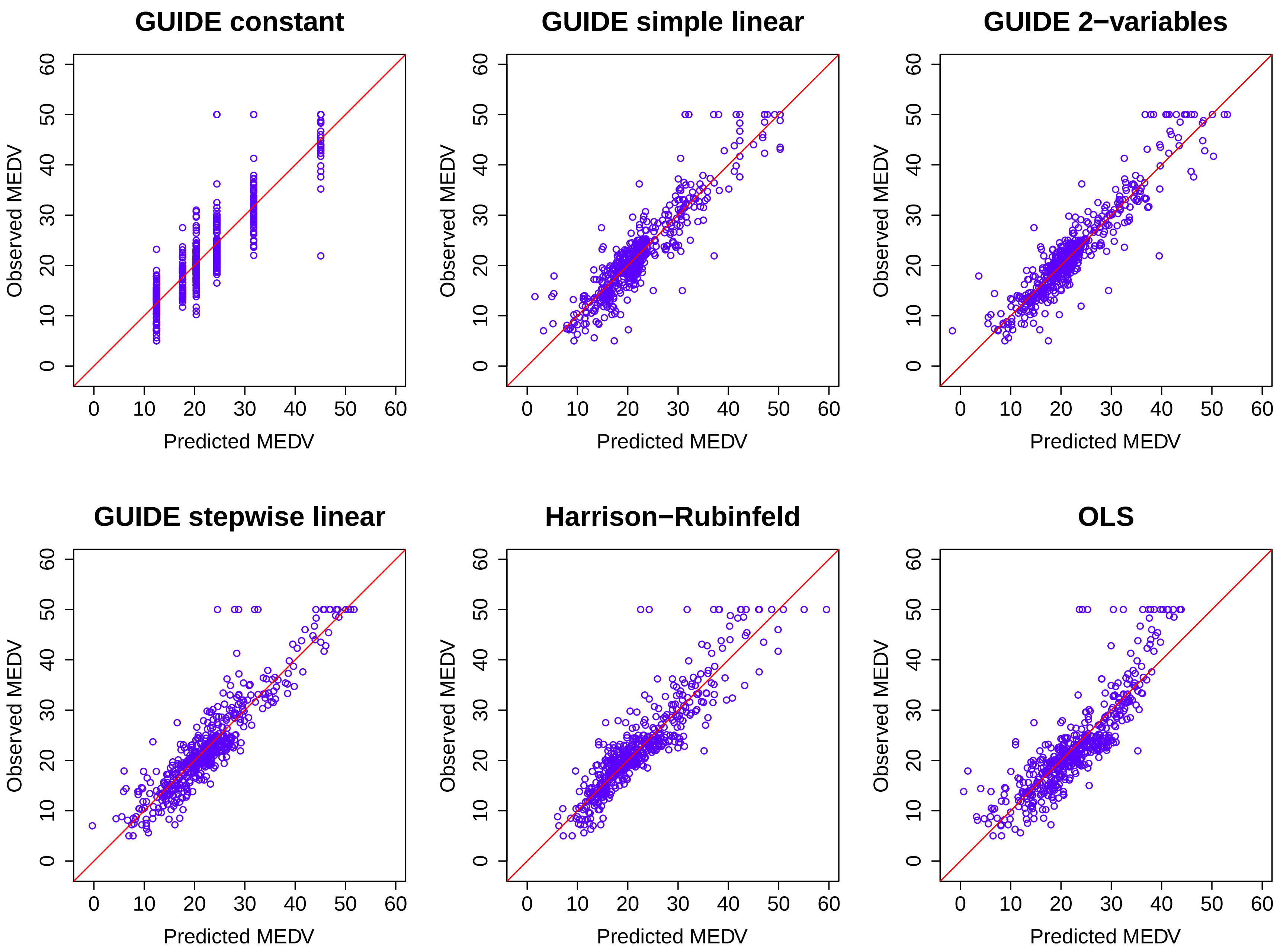
RM > 6.54, PT ≤ 19.4

RM > 6.54, PT > 19.4

RM ≤ 6.54, PT > 19.6, DIS ≤ 1.81

RM ≤ 6.54, PT > 19.6, DIS > 1.81

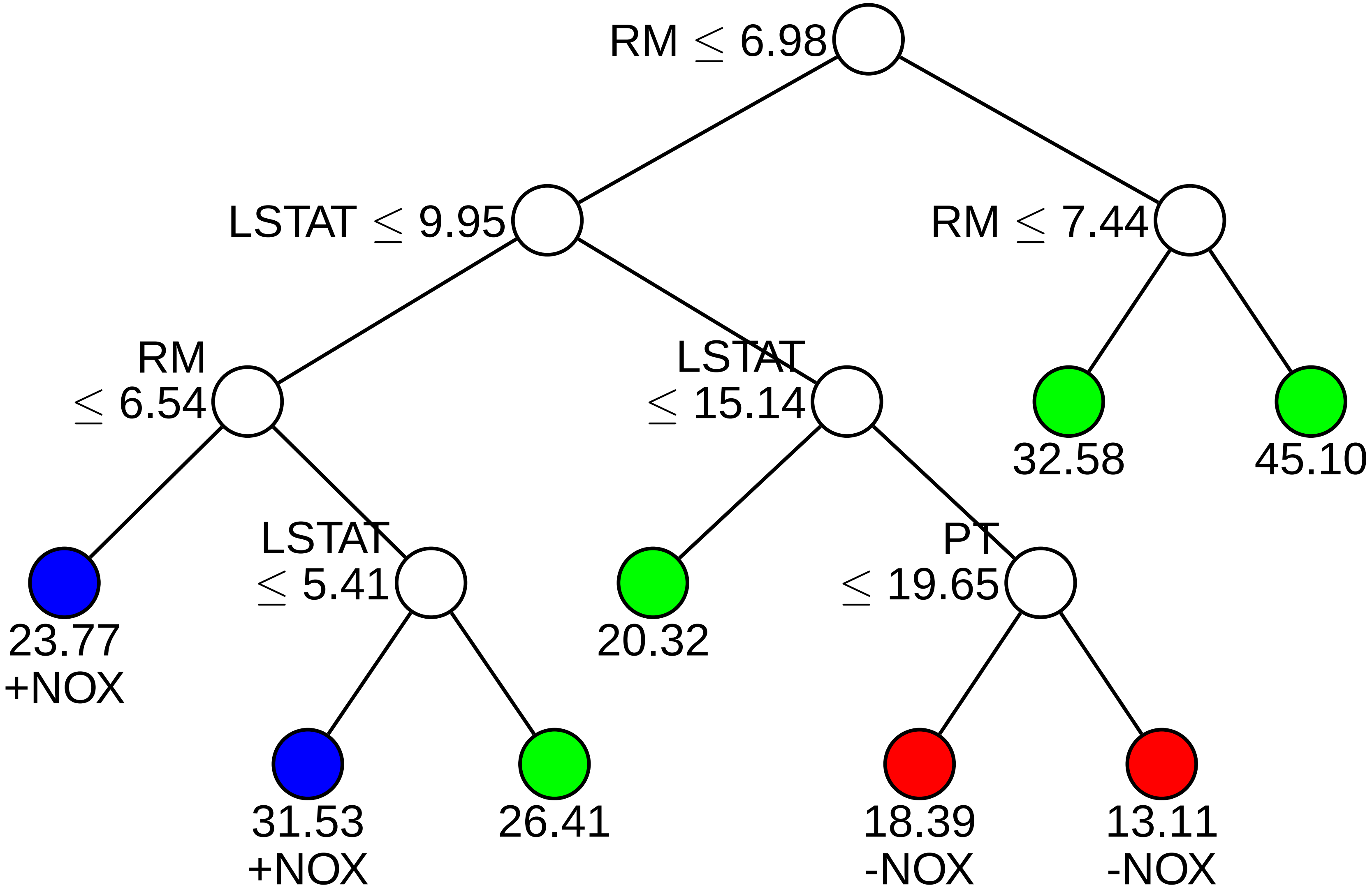
Comparison of models



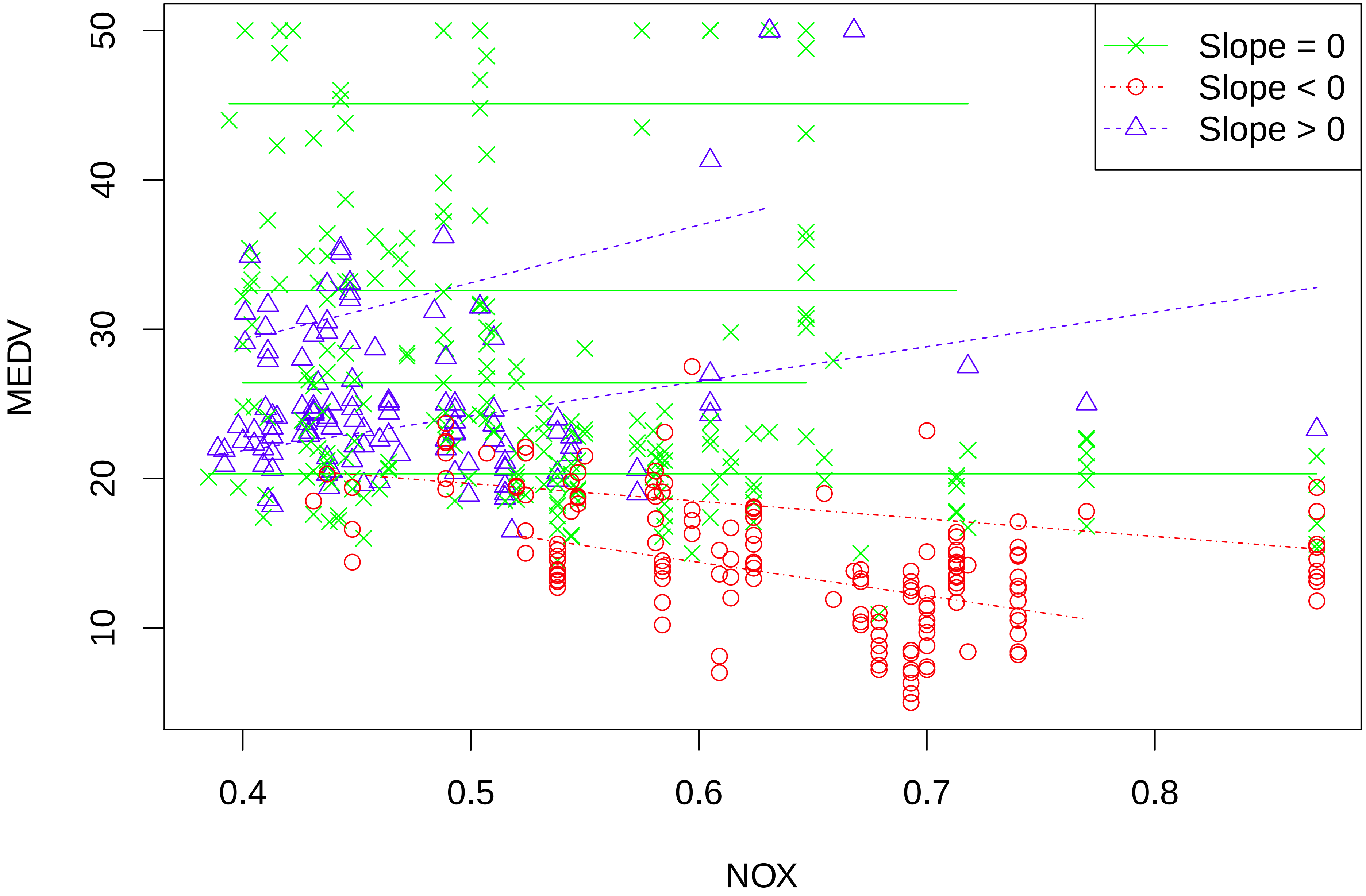
Why did Harrison & Rubinfeld use NOX^2 but not NOX?

	Estimate	P-value		Estimate	P-value
Constant	4.5e+00	<2e-16	AGE	4.0e-05	0.941
CRIM	-1.2e-02	<2e-16	log(DIS)	-2.0e-01	6e-08
ZN	1.2e-04	0.815	log(RAD)	8.9e-02	4e-06
INDUS	1.5e-04	0.947	TAX	-4.2e-04	6e-04
CHAS	9.3e-02	0.005	PT	-3.0e-02	6e-09
NOX	1.9e-01	0.855	B	3.6e-04	4e-04
NOX^2	-7.8e-01	0.311	log(LSTAT)	-3.7e-01	<2e-16
RM^2	6.2e-03	2e-06			

Model for MEDV with NOX as only linear predictor



MEDV vs NOX



Poisson regression

- y_i has a Poisson distribution with mean μ_i
- $\log \mu_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$
- $\mu_i = \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})$
- $\beta_0, \beta_1, \ldots, \beta_k$ estimated by maximizing the Poisson likelihood

$$\prod_{i=1}^{n} \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!}$$

Unreplicated 3 x 2 x 4 x 10 x 3 soldering experiment (Comizzoli et al., 1990; Chambers and Hastie, 1992)

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)

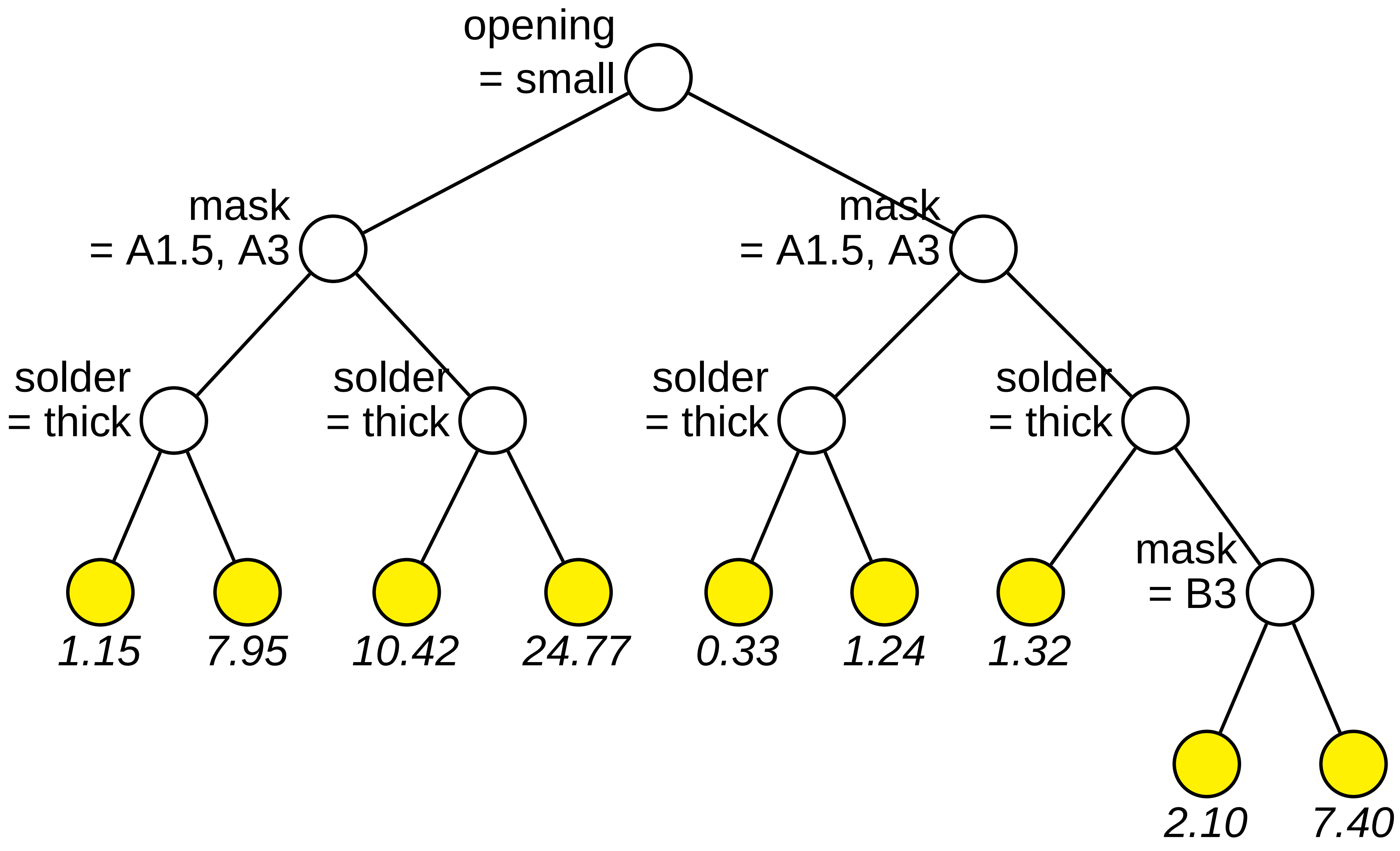
Full 2nd-degree Poisson loglinear model

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

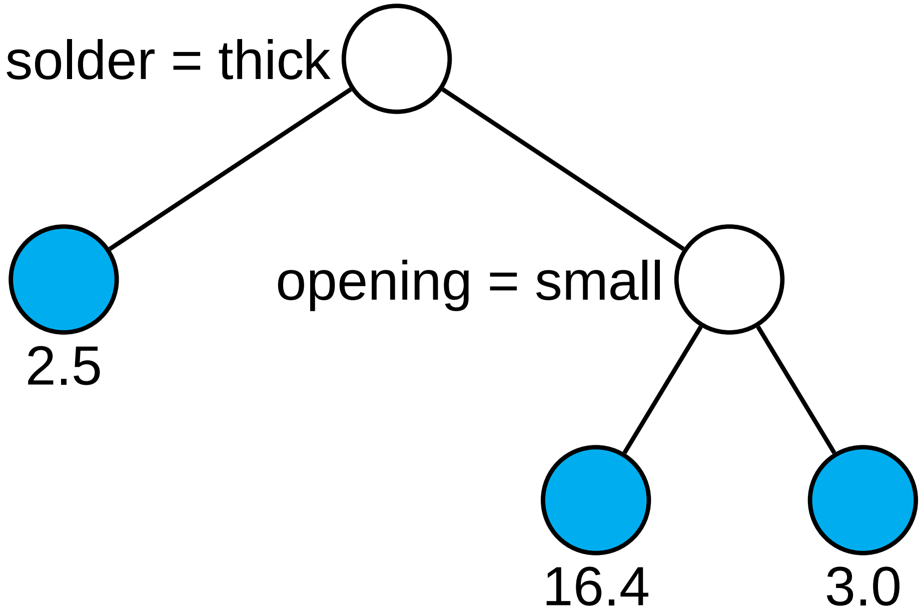
Regressor	Coef	t-stat	Regressor	Coef	t-stat
Constant	-2.668	-9.25			
maskA3	0.396	1.21	openmedium	0.921	2.95
maskB3	2.101	7.54	opensmall	2.919	11.63
maskB6	3.010	11.36	soldthin	2.495	11.44
padD6	-0.369	-5.17	maskA3:openmedium	0.816	2.44
padD7	-0.098	-1.49	maskB3:openmedium	-0.447	-1.44
padL4	0.262	4.32	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



Estimated mean number of solder skips given under each leaf node

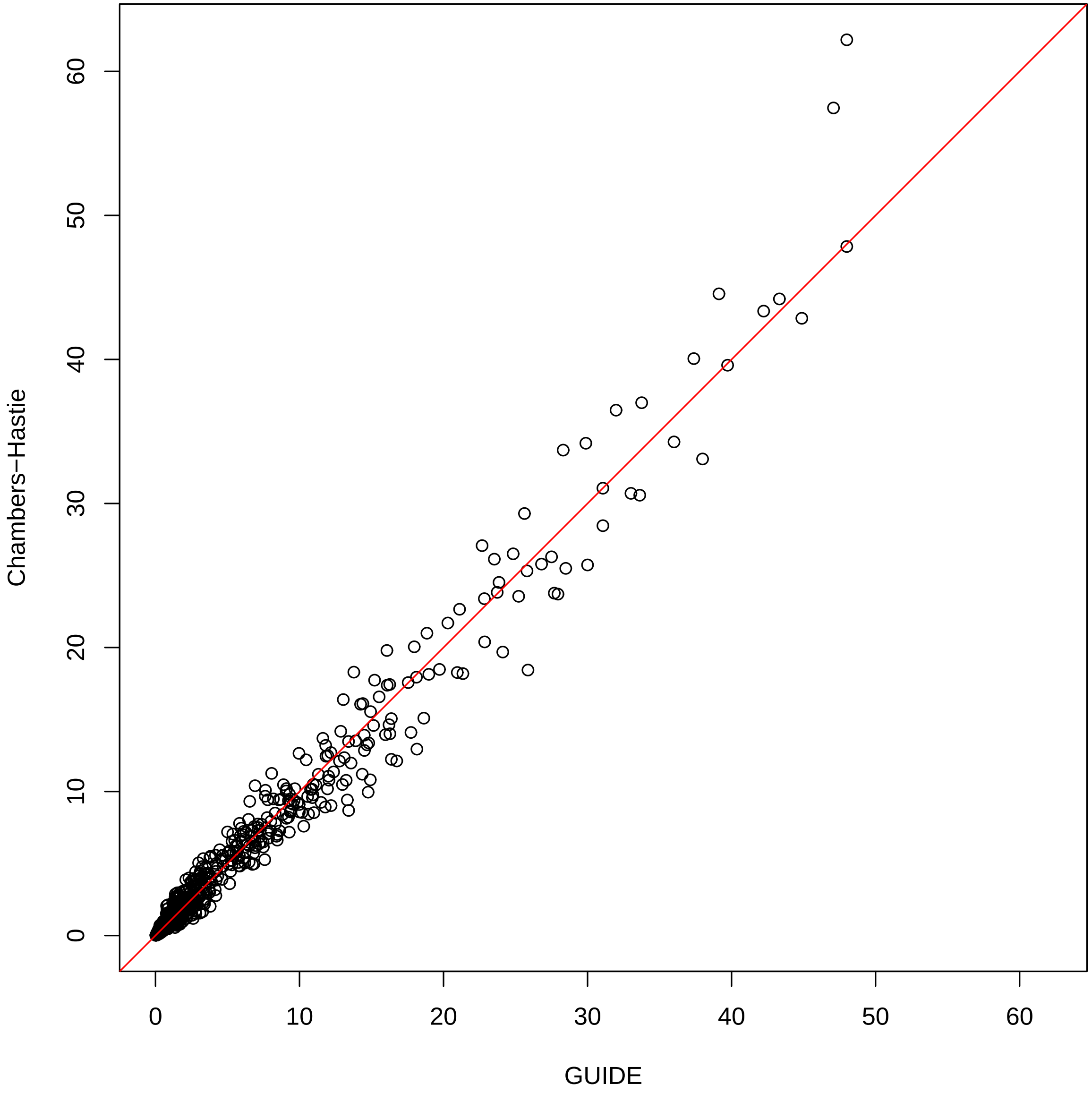
Regression coefficients

	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

Regression coefficients (cont'd.)

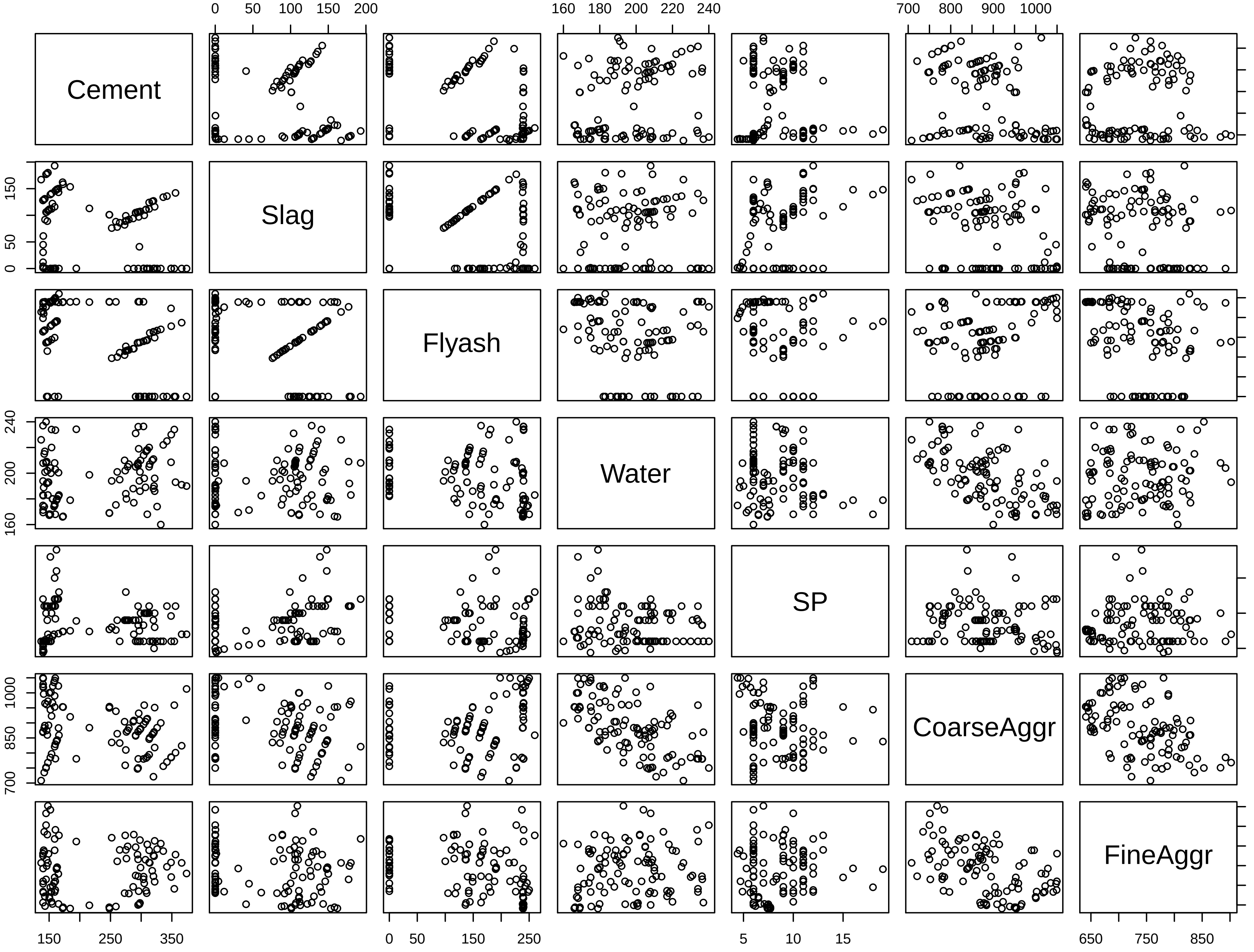
	solder = thick		solder = thin			
			opening = small		medium or large	
Regressor	Coef	t-stat	Coef	t-stat	Coef	t-stat
padD6	-0.32	-2.03	-0.25	-2.8	-0.80	-4.6
padD7	0.12	0.85	-0.15	-1.7	-0.19	-1.3
padL4	0.70	5.53	0.08	1.0	0.21	1.6
padL6	-0.40	-2.46	-0.72	-6.8	-0.82	-4.7
padL7	0.04	0.29	-0.65	-6.3	-0.76	-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

Chambers-Hastie vs. linear GUIDE fits



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
150
250
5
10
15
650
750
850
160
200
240
0
50
150
0
100
200
150
250
350
5
10
15
700
850
1000
650
750
850

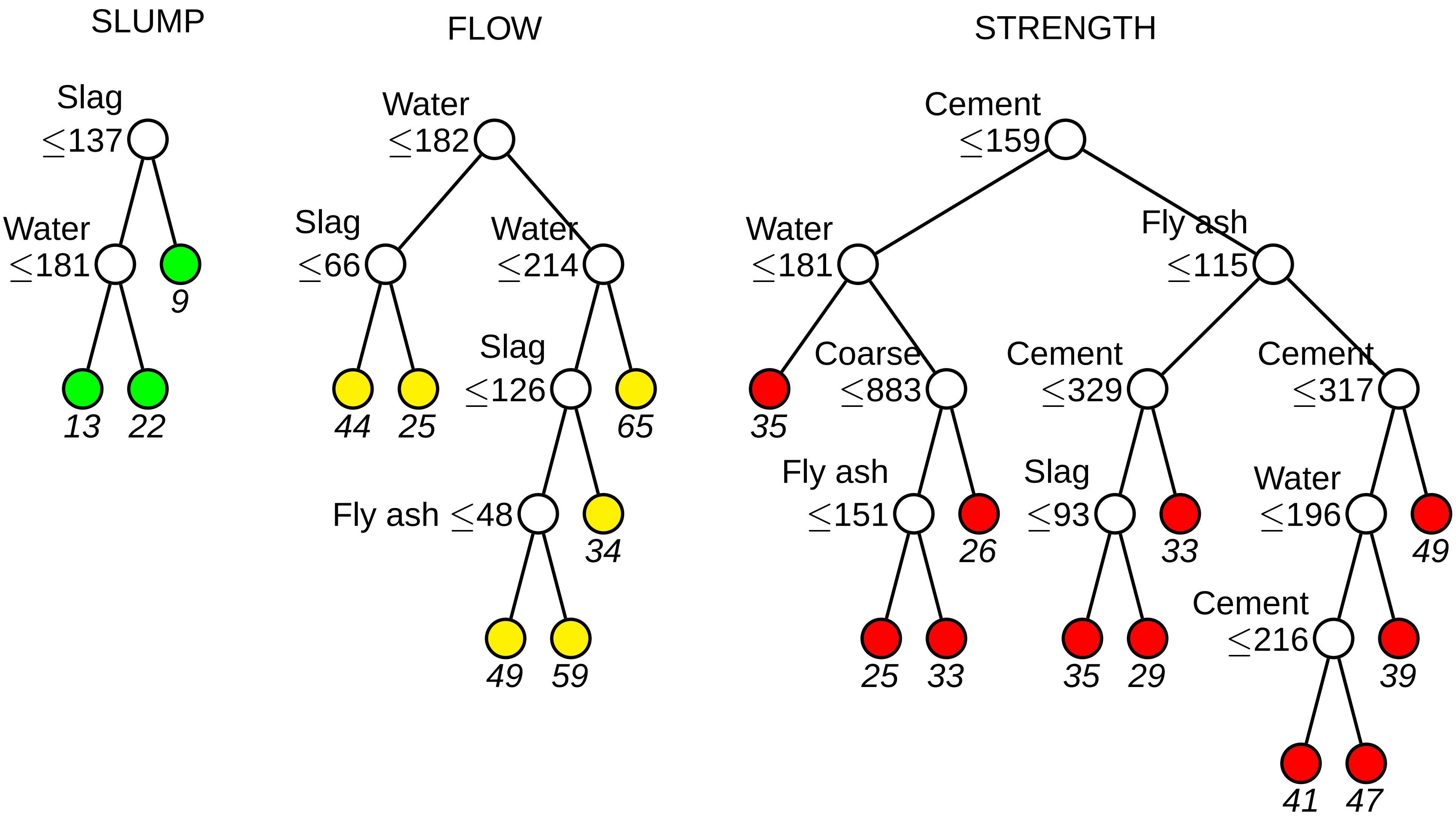
Separate linear models

	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

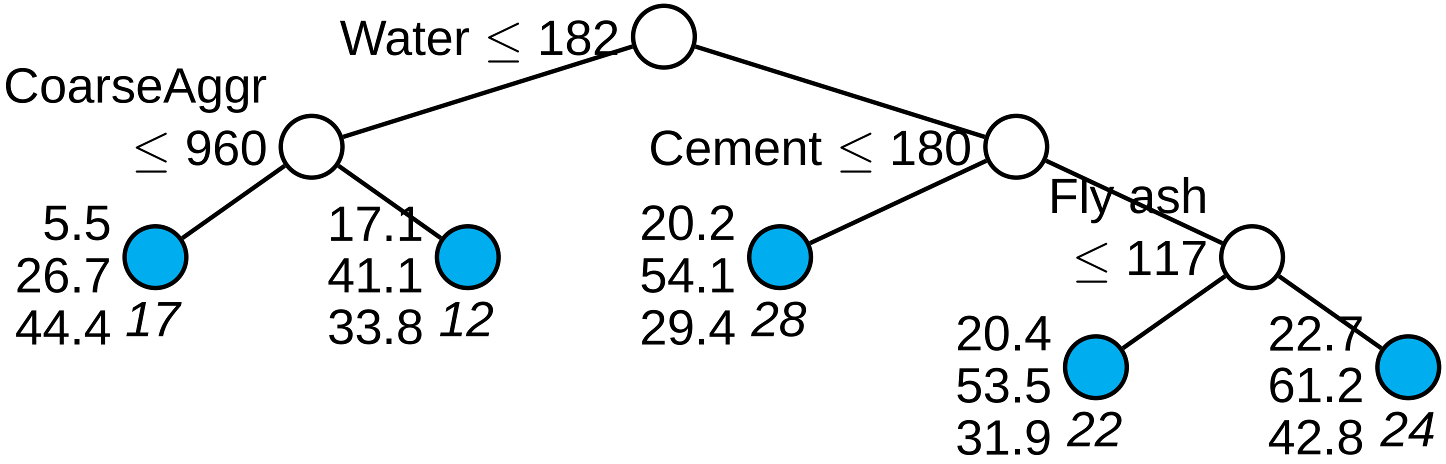
Regression models for Slump

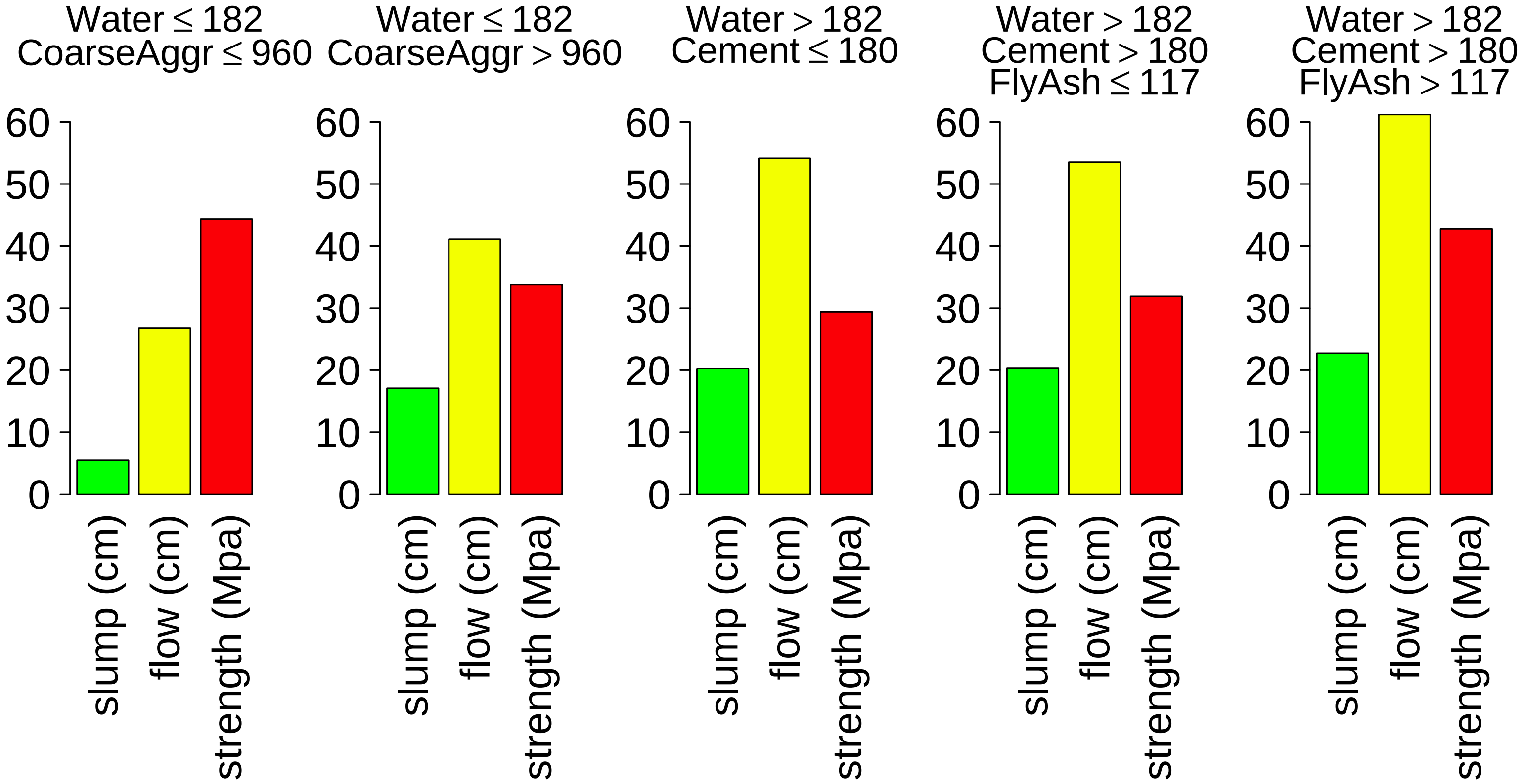
	Estimate	Std. Error	t value	Pr($>|t|$)
(Intercept)	-18.099	7.314	-2.475	0.01502 *
Water	0.199	0.036	5.455	3.56e-07 ***
Slag	-0.039	0.012	-3.227	0.00169 **
(Intercept)	11.370	9.683	1.174	0.243
Water	0.050	0.0486	1.025	0.308
Slag	-0.479	0.104	-4.604	1.23e-05 ***
Water:Slag	0.002	0.001	4.251	4.83e-05 ***

Separate regression trees



GUIDE tree


Water $\leq$ 182
CoarseAggr $\leq$ 960
Cement $\leq$ 180
Fly ash $\leq$ 117
5.5
26.7
44.4 *17*
17.1
41.1
33.8 *12*
20.2
54.1
29.4 *28*
20.4
53.5
31.9 *22*
22.7
61.2
42.8 *24*


Water ≤ 182
CoarseAggr ≤ 960
Water ≤ 182
CoarseAggr > 960
Water > 182
Cement ≤ 180
Water > 182
Cement > 180
FlyAsh ≤ 117
Water > 182
Cement > 180
FlyAsh > 117
0
10
20
30
40
50
60
slump (cm)
flow (cm)
strength (Mpa)

College tuition

- 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 in-state and out-of-state tuition
- 515 complete cases

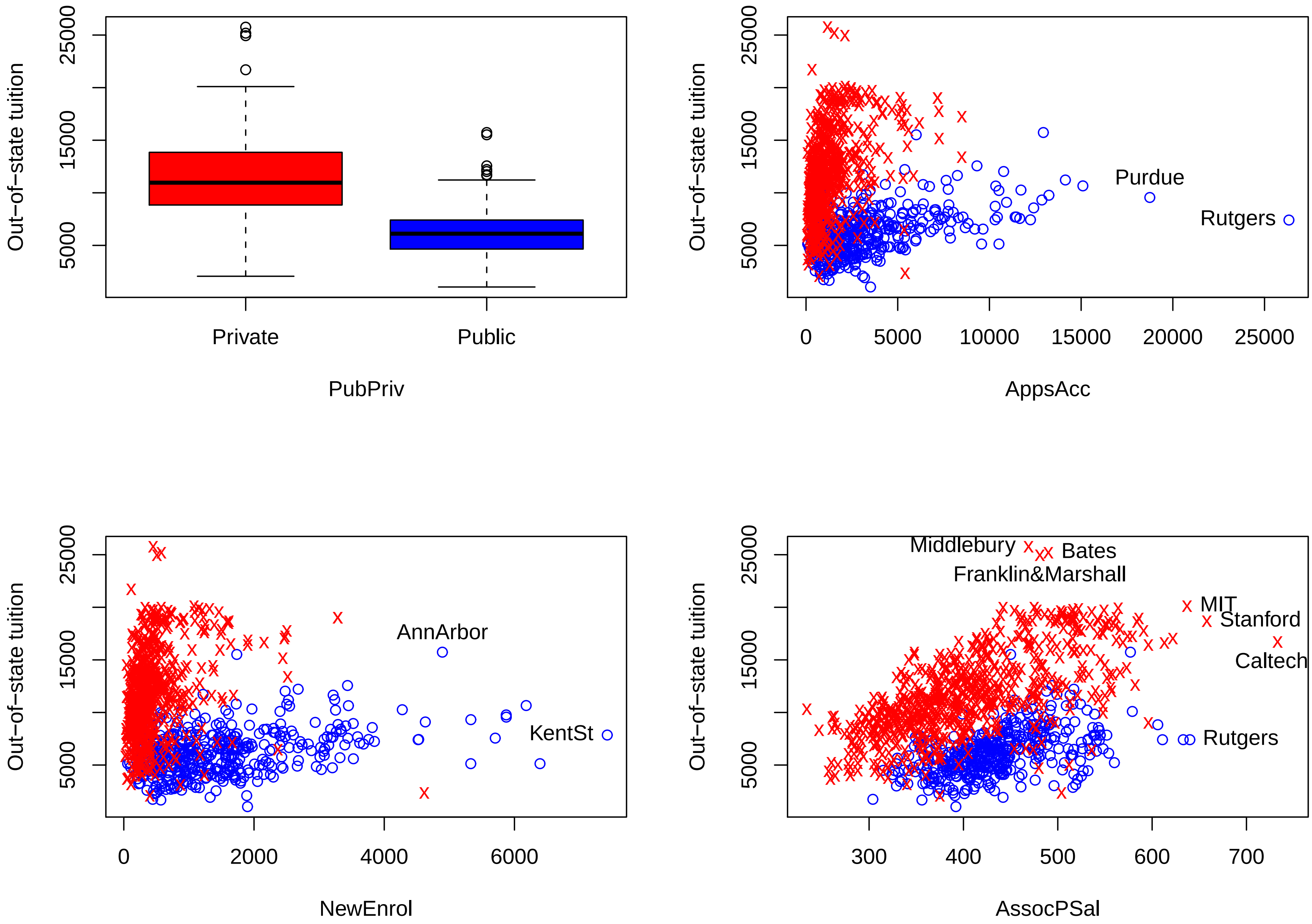
Explanatory variables for college data

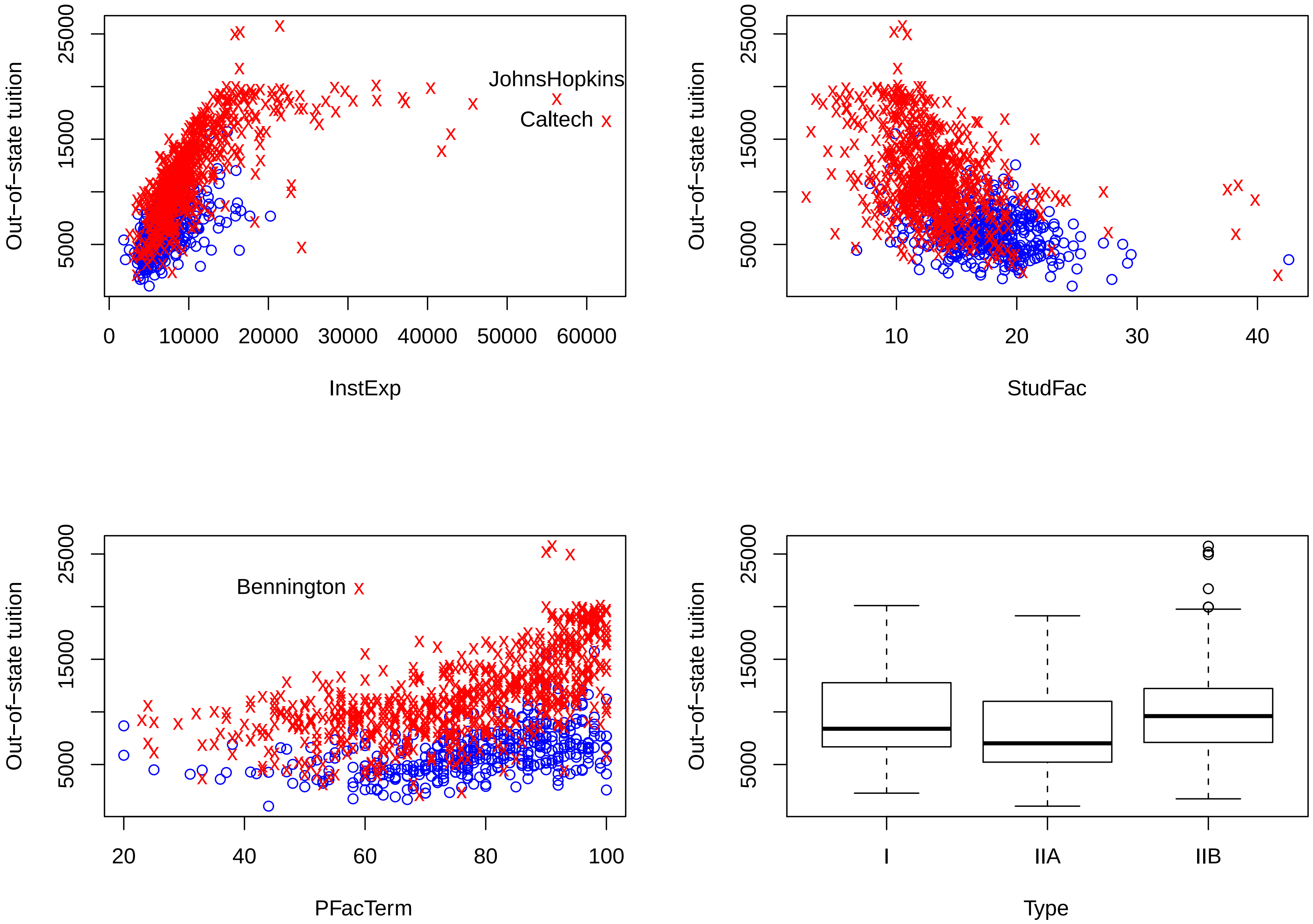
Name	Description	#Missing
PubPriv	Public or private college (binary)	0
CombSAT	Average Combined SAT score	471
AppsRec	Number of applications received	9
AppsAcc	Number of applicants accepted	9
NewEnrol	Number of new students enrolled	5
Top10	Percent new students from top 10% of H.S. class	183
Top25	Percent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

Explanatory variables for college data (cont'd)

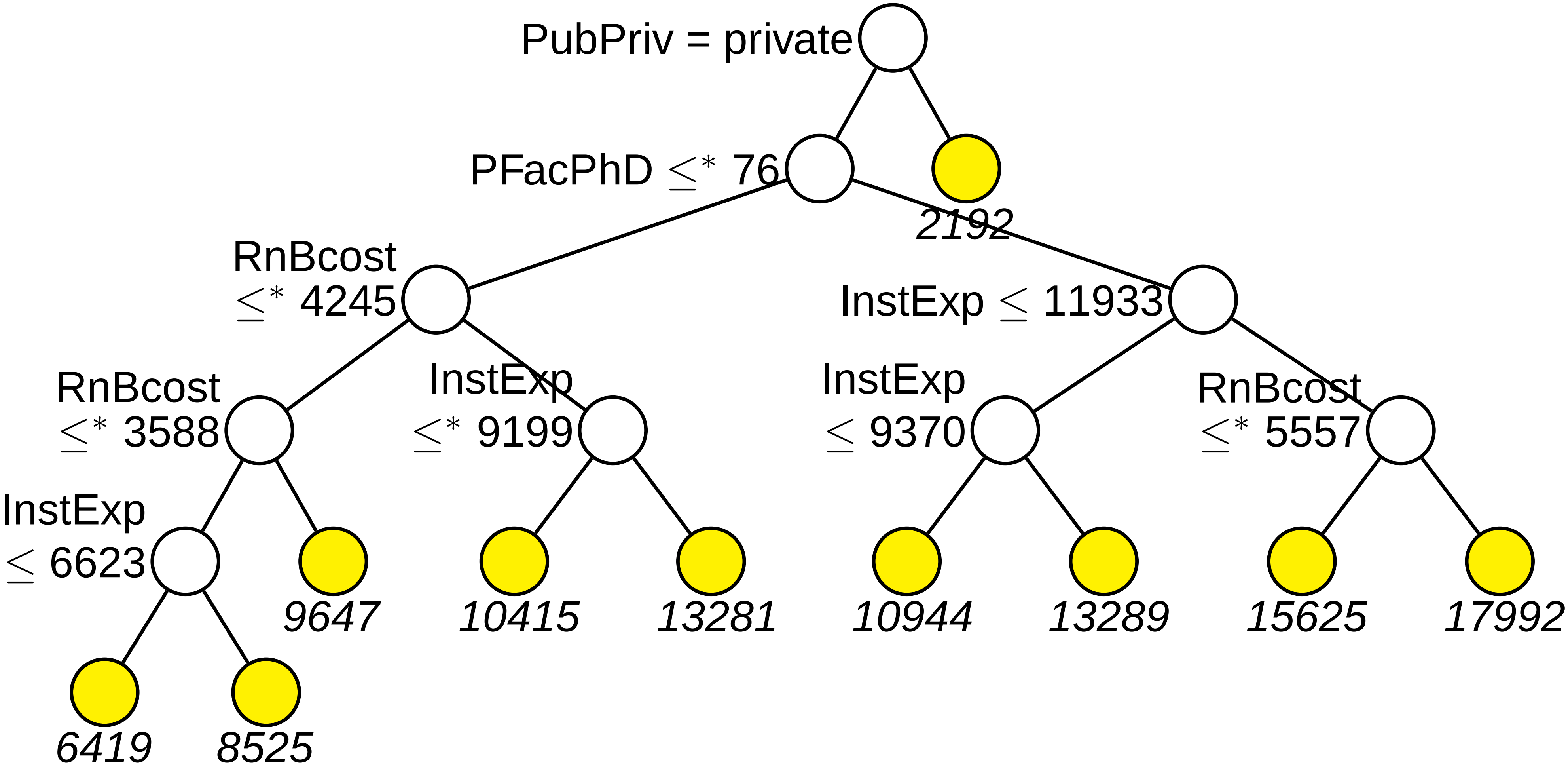
Name	Description	#Missing
RnBcost	Room and board costs	57
PFacPhD	Percent of faculty with Ph.D.'s	29
StudFac	Student/faculty ratio	2
InstExp	Instructional expenditure per student	24
GradRate	Graduation rate	69
Type	Type of college (I: PhD, IIA: master, or IIB: bachelor)	0
FullPSal	Average salary—full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

513 cases with complete observations

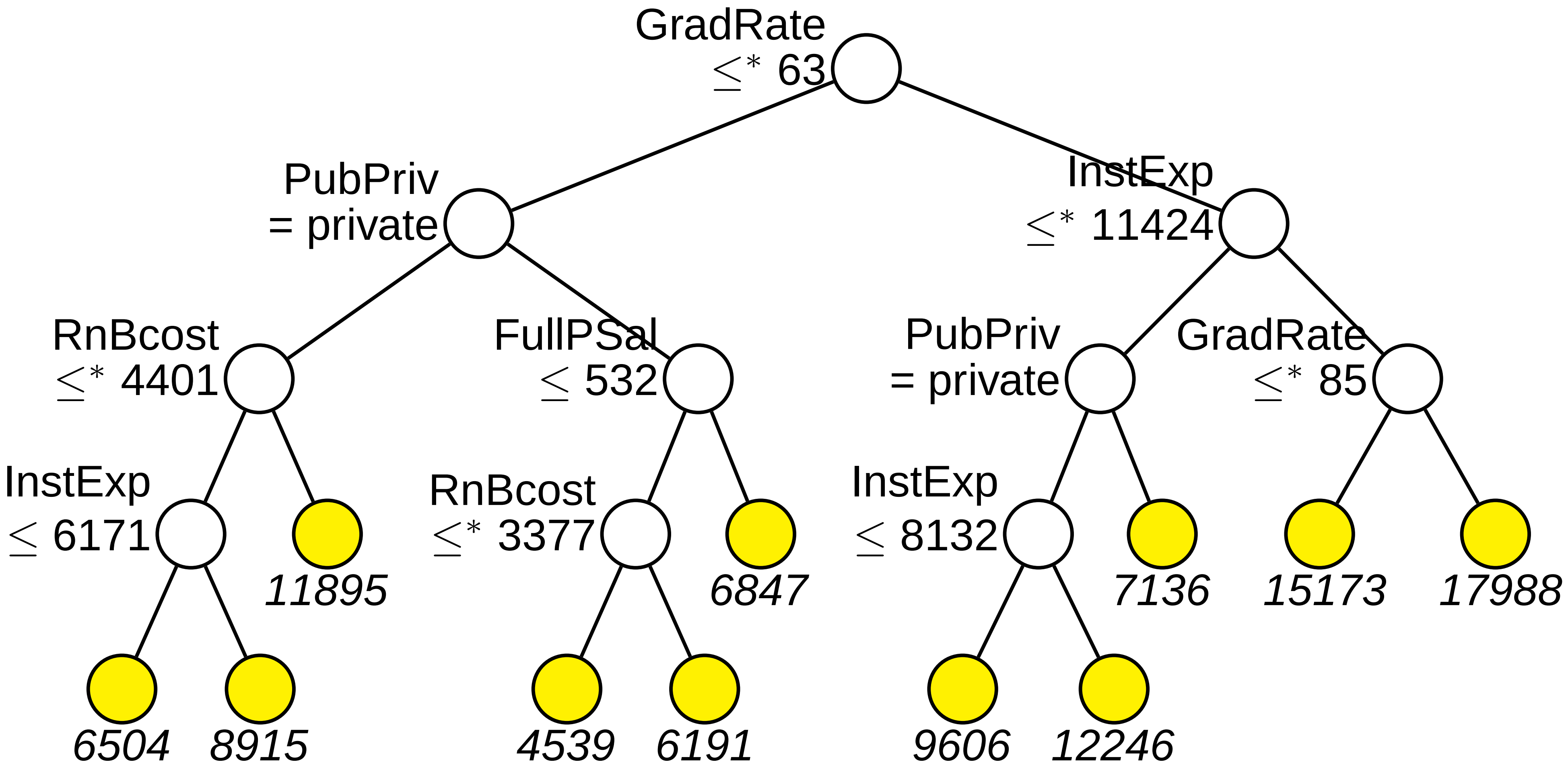

Out−of−state tuition
5000
15000
25000
Private
Public
PubPriv
Out−of−state tuition
5000
15000
25000
0
5000
10000
15000
20000
25000
AppsAcc
Purdue
Rutgers
Out−of−state tuition
5000
15000
25000
0
2000
4000
6000
NewEnrol
AnnArbor
KentSt
Out−of−state tuition
5000
15000
25000
300
400
500
600
700
AssocPSal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


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

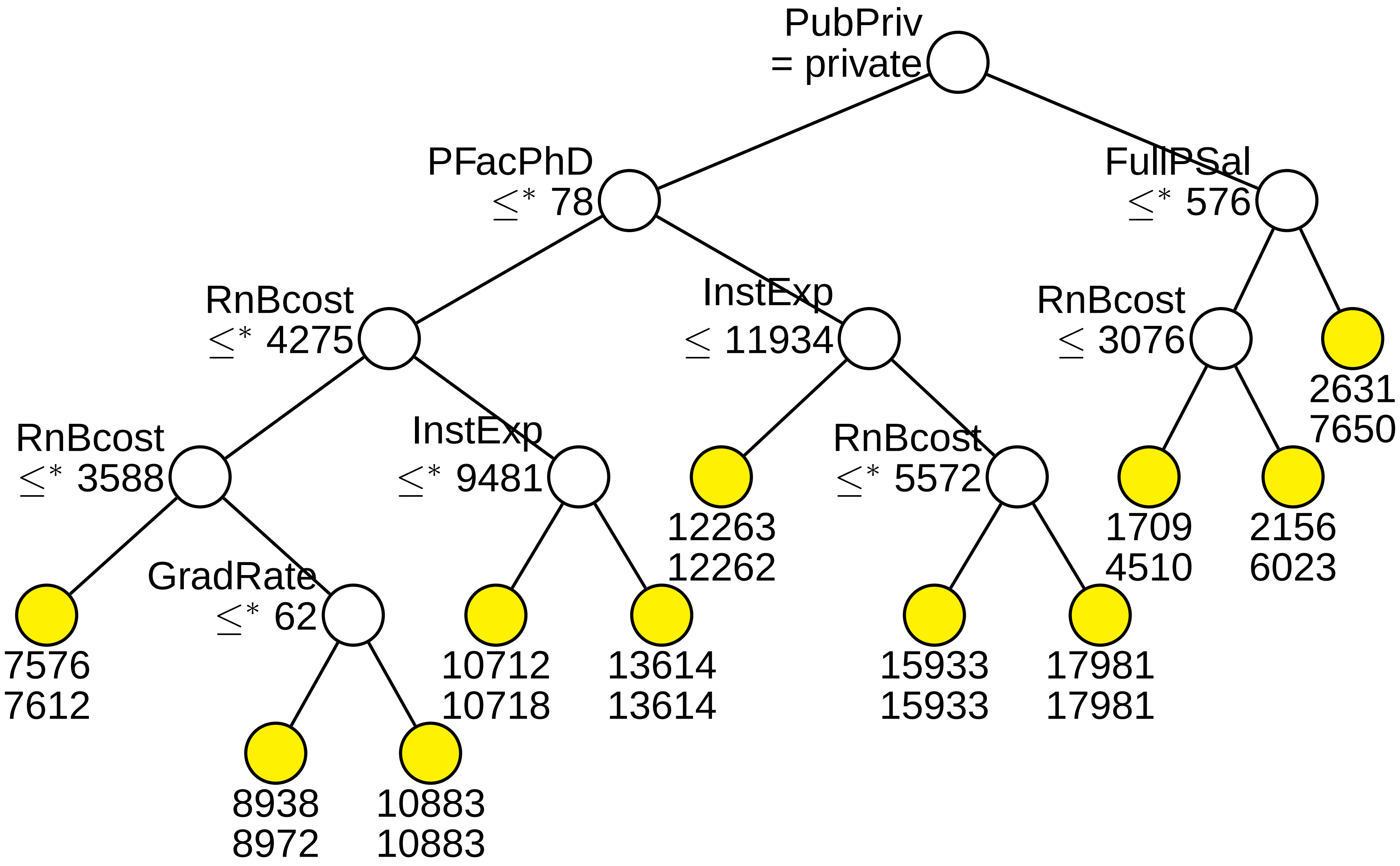
Model for in-state tuition



Model for out-of-state tuition



In-state (upper) and out-of-state (lower) tuition



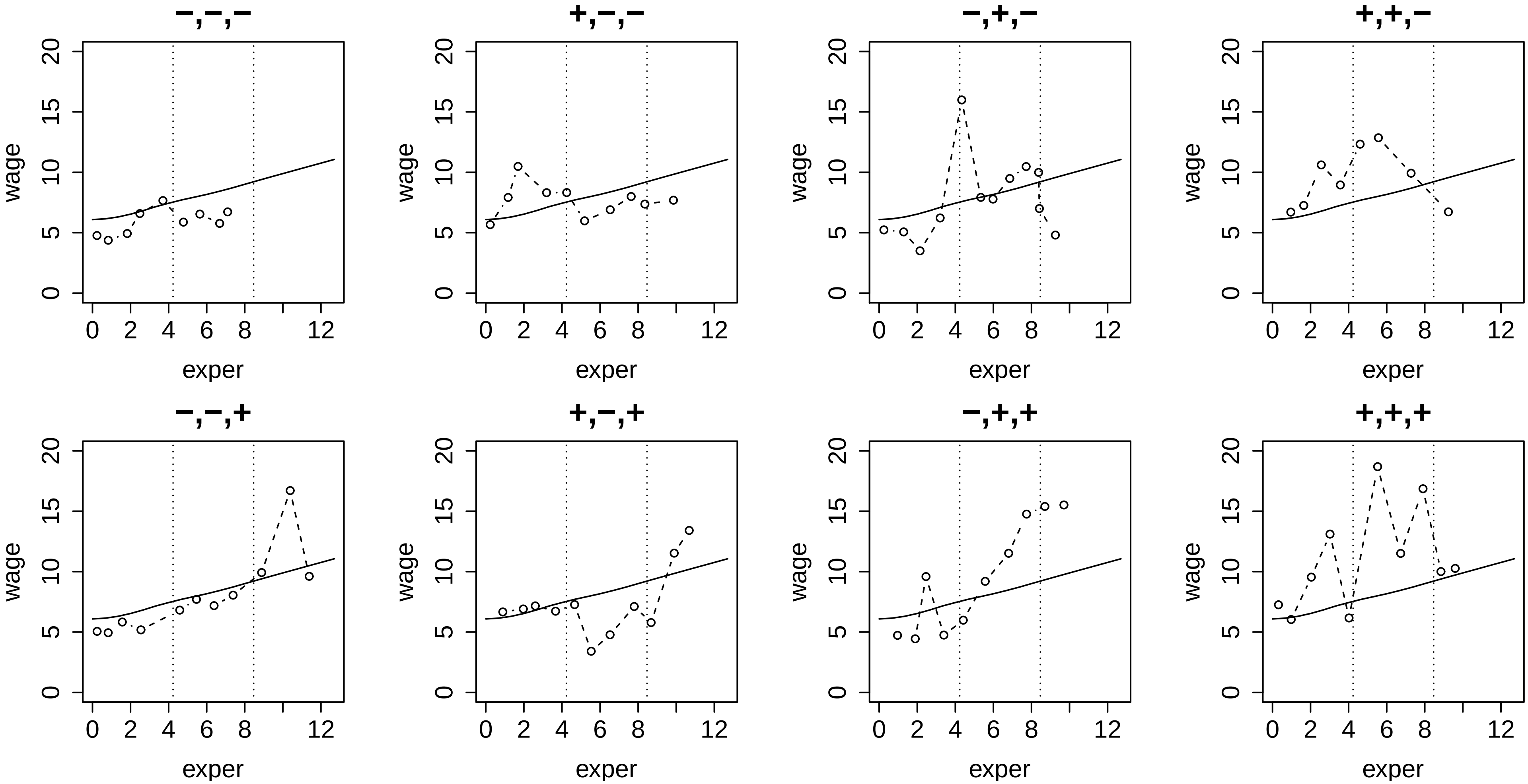
Hourly wage of high-school dropouts

- 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 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:

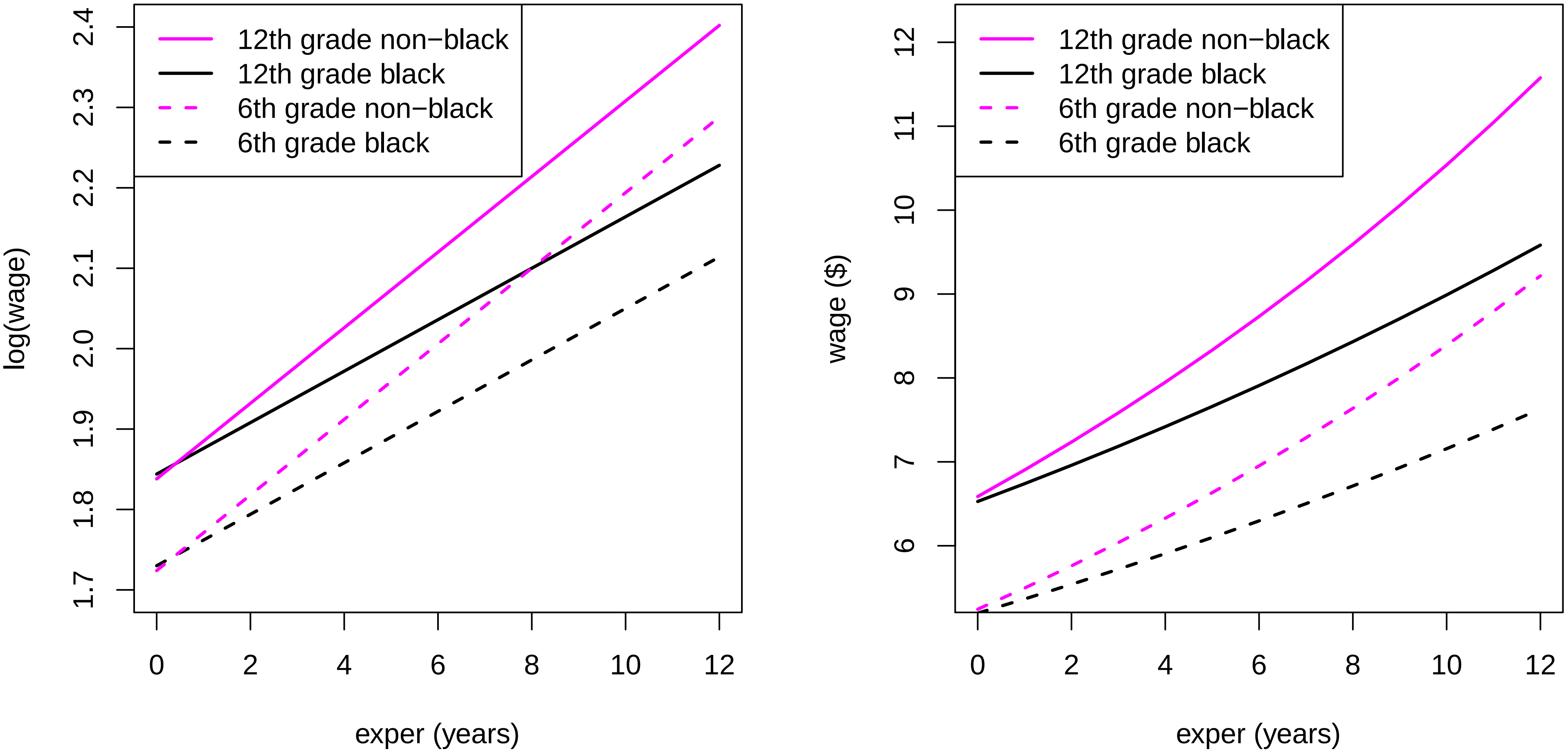
1. Random (subject) intercepts and slopes $b_0 \sim N(0, \sigma_0^2)$ and $b_1 \sim N(0, \sigma_1^2)$; $\epsilon \sim N(0, \sigma^2)$; all independent
2. Log transformation of `wage` to address skewness, linearize individual wage trajectories, and overcome range restriction
3. Predictions of `wage` requires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

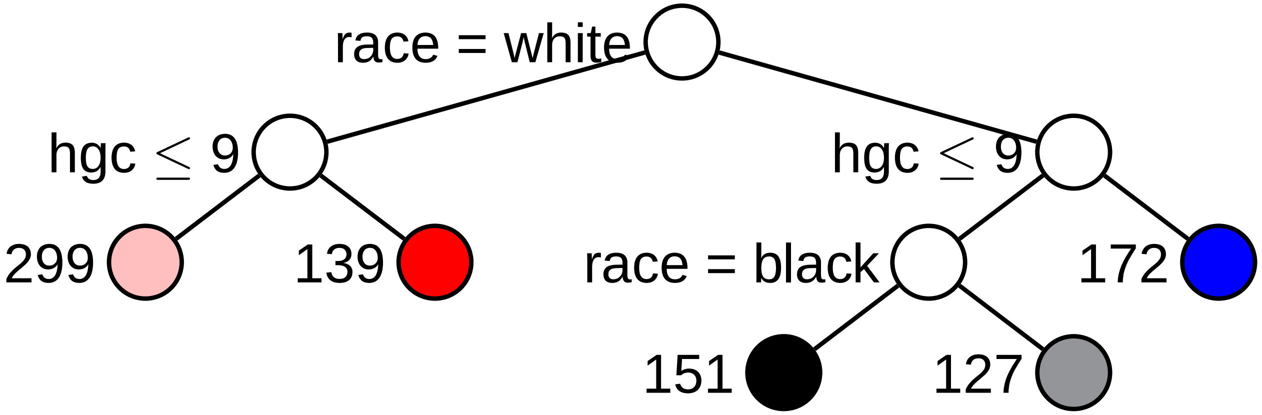
Coefficients of fixed effect terms

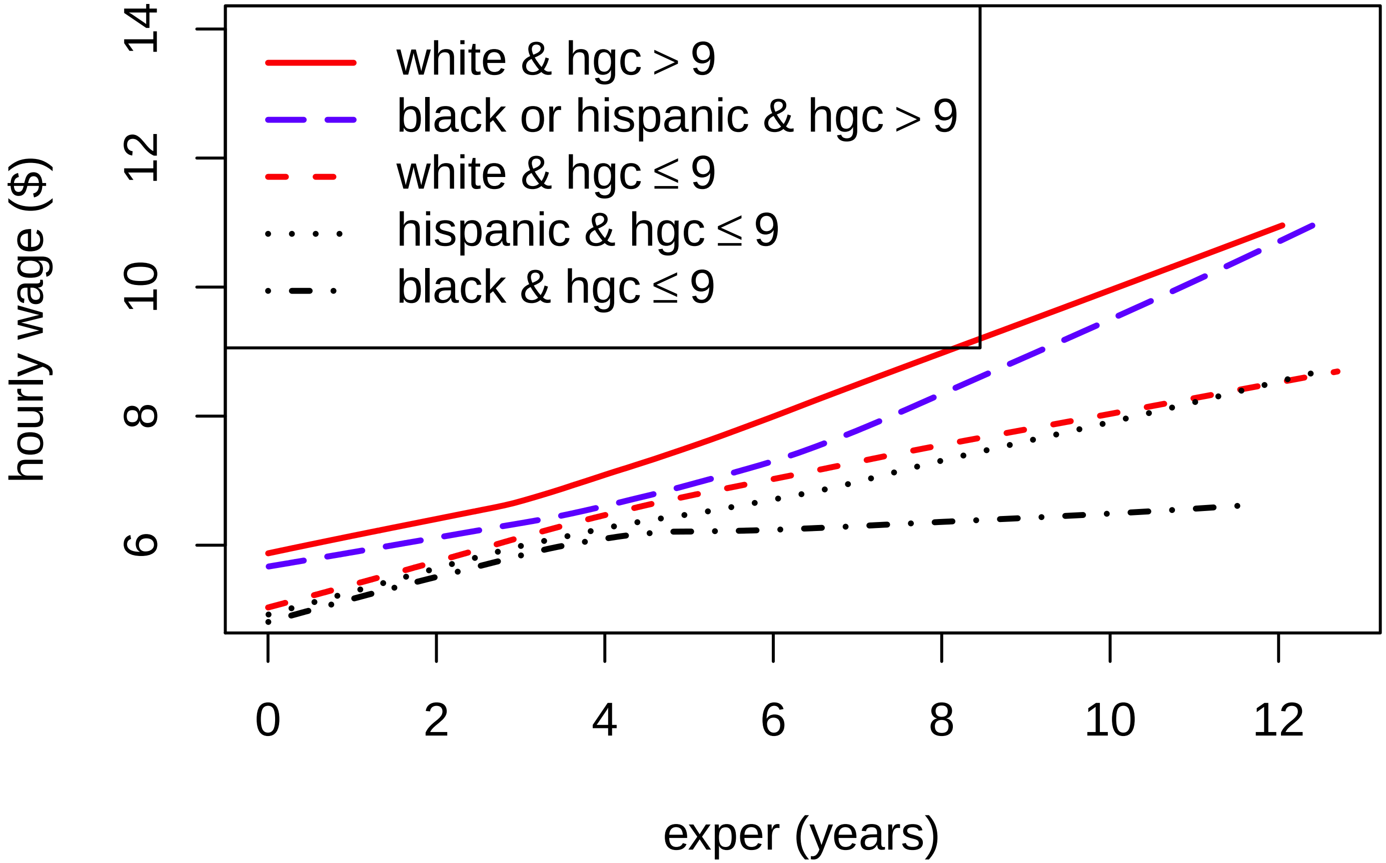
	Value	Std.Error	DF	t-value	p-value
(Intercept)	1.382	0.059	5511	23.43	0.000
hgc	0.038	0.006	884	5.94	0.000
exper	0.047	0.003	5511	14.57	0.000
black	0.006	0.025	884	0.25	0.804
hisp	-0.028	0.027	884	-1.03	0.302
exper$\times$black	-0.015	0.006	5511	-2.65	0.008
exper$\times$hisp	0.009	0.006	5511	1.51	0.131

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

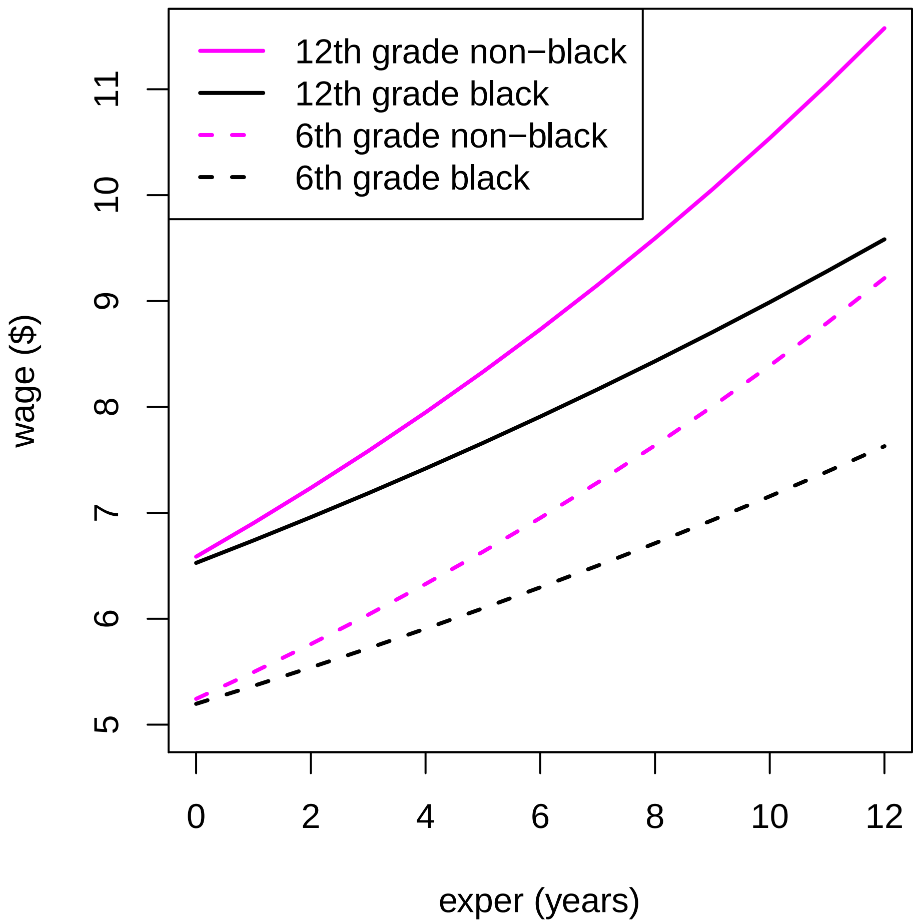
Sample LME fits



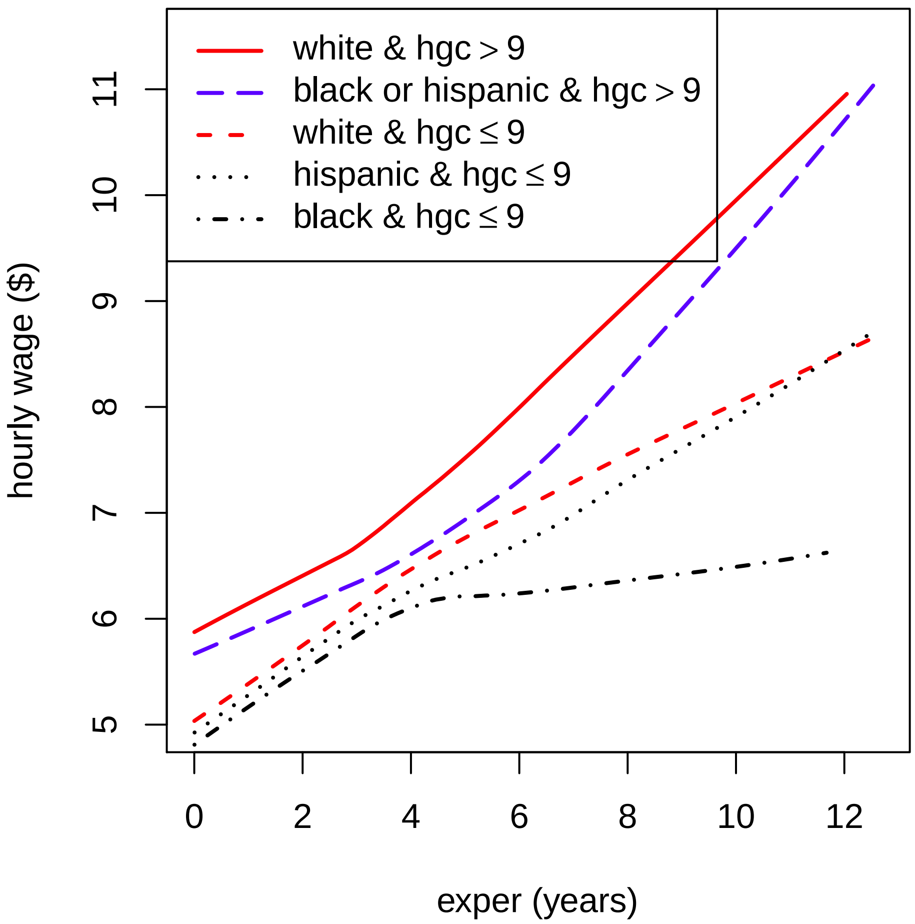

race = white
hgc ≤ 9
hgc ≤ 9
299
139
race = black
172
151
127


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



Linear mixed model



GUIDE model

Conclusions

1. Parametric models are often constrained by range restrictions, missing value patterns, distributional assumptions, and number and variety of predictor variables
2. Regression tree models are not so constrained
3. Regression tree models can be used to check parametric model assumptions as well as suggest the functional forms to use
4. Regression tree models can be visualized by looking at tree structures and lowess plots
5. Regression tree models cannot be used for “statistical inference” in the traditional sense, because no model is assumed
6. Regression tree models are meant for data exploration and prediction

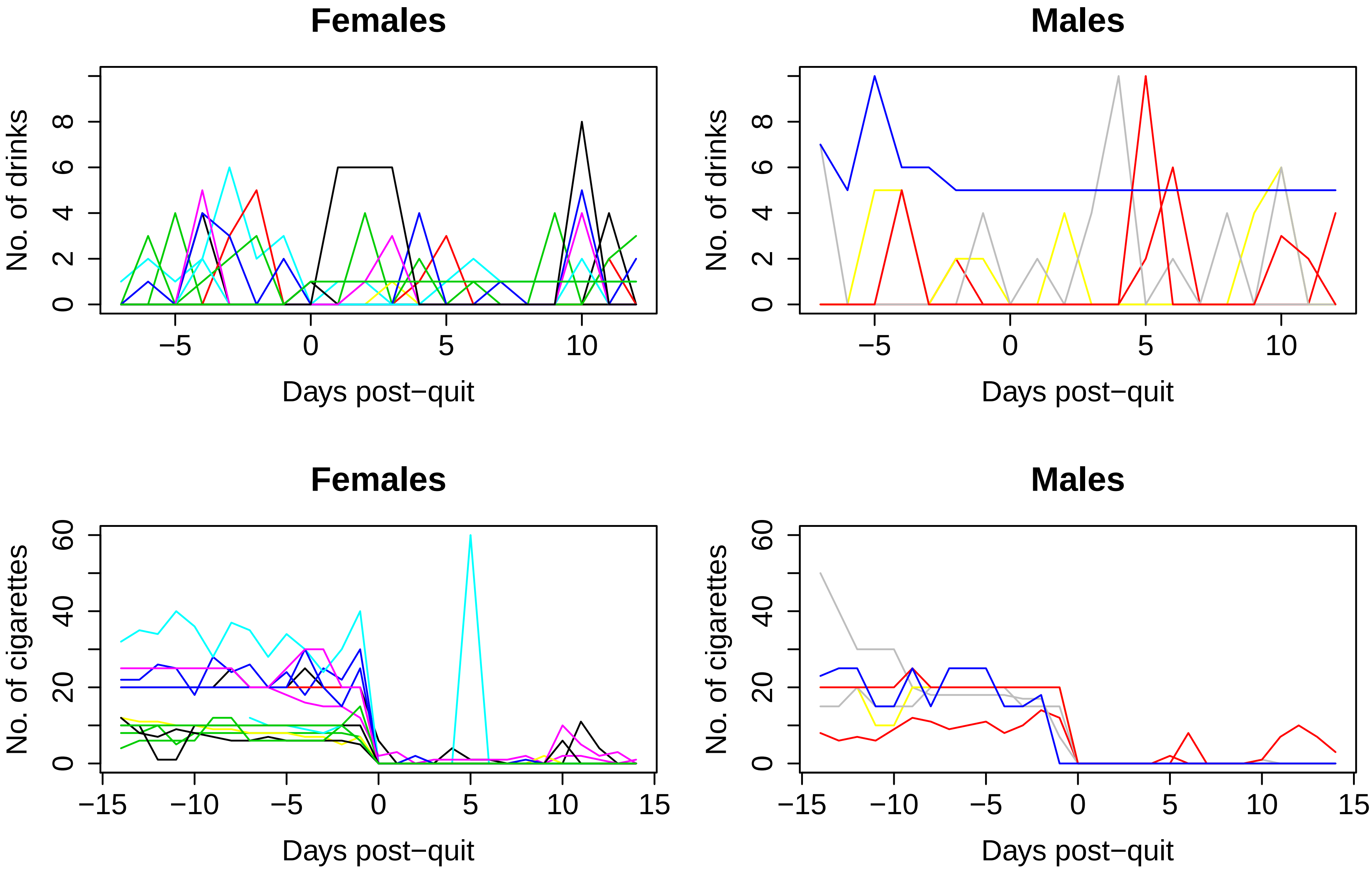
Ecological momentary assessment (EMA) in a smoking cessation trial

- Responses are number of drinks and cigarettes consumed daily for two weeks before and after quit day for 1470 persons
- 135 explanatory variables with 0–1308 missing values
- 32–63% persons missing drink responses in 8–14 days pre-quit and 13–14 days post-quit

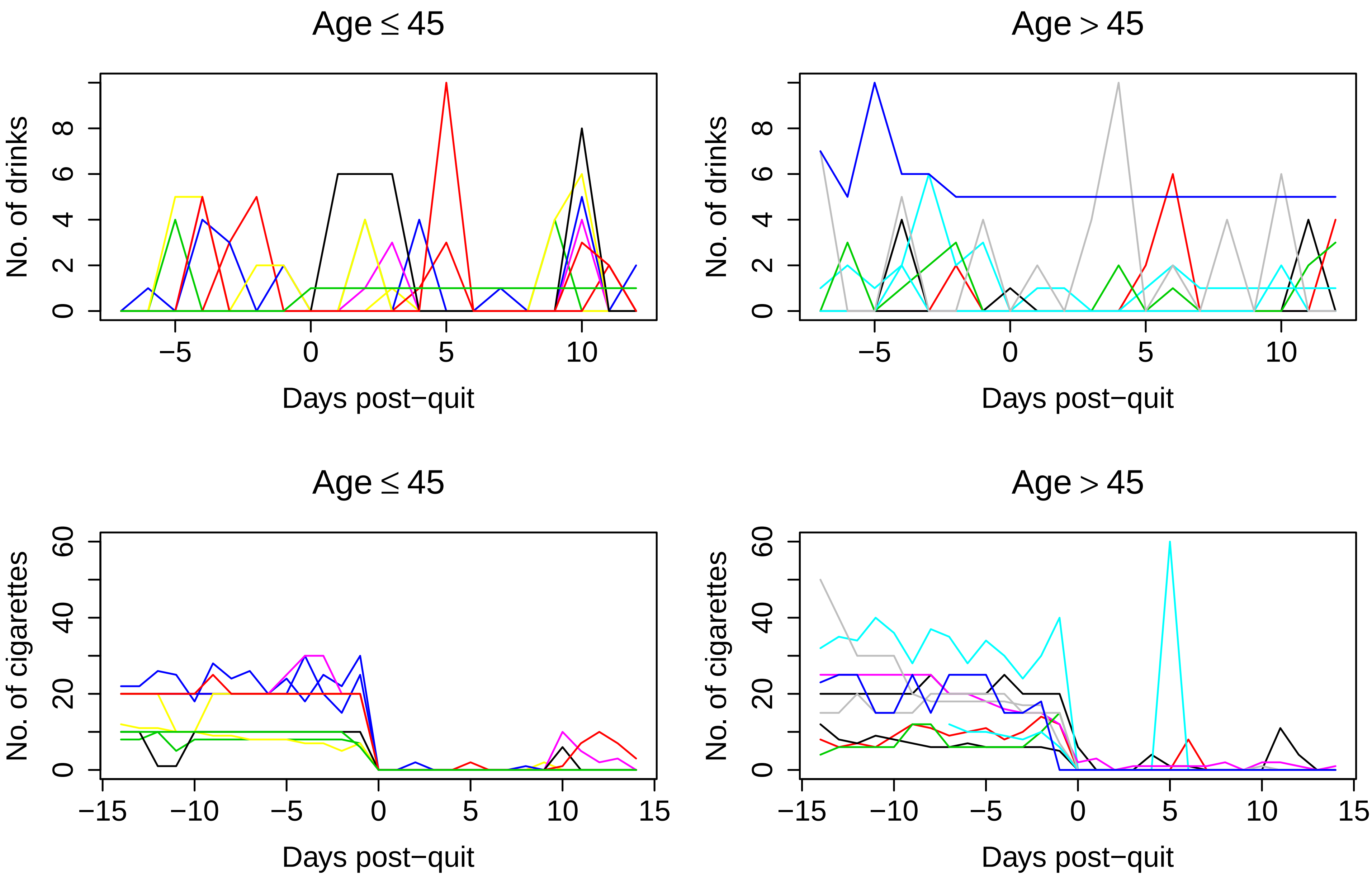
135 explanatory variables

Age, gender, marital status, education, income, race, age 1st cigarette, years smoked, cigarette type, various measures of emotional reaction to smoking, living and working environments w.r.t. smoking, number of times tried quitting, quitting methods, numerous FTND, PRISM, and WISDM scores, baseline health and physical measurements (e.g., blood pressure, BMI), treatment, past drinking frequency, etc.

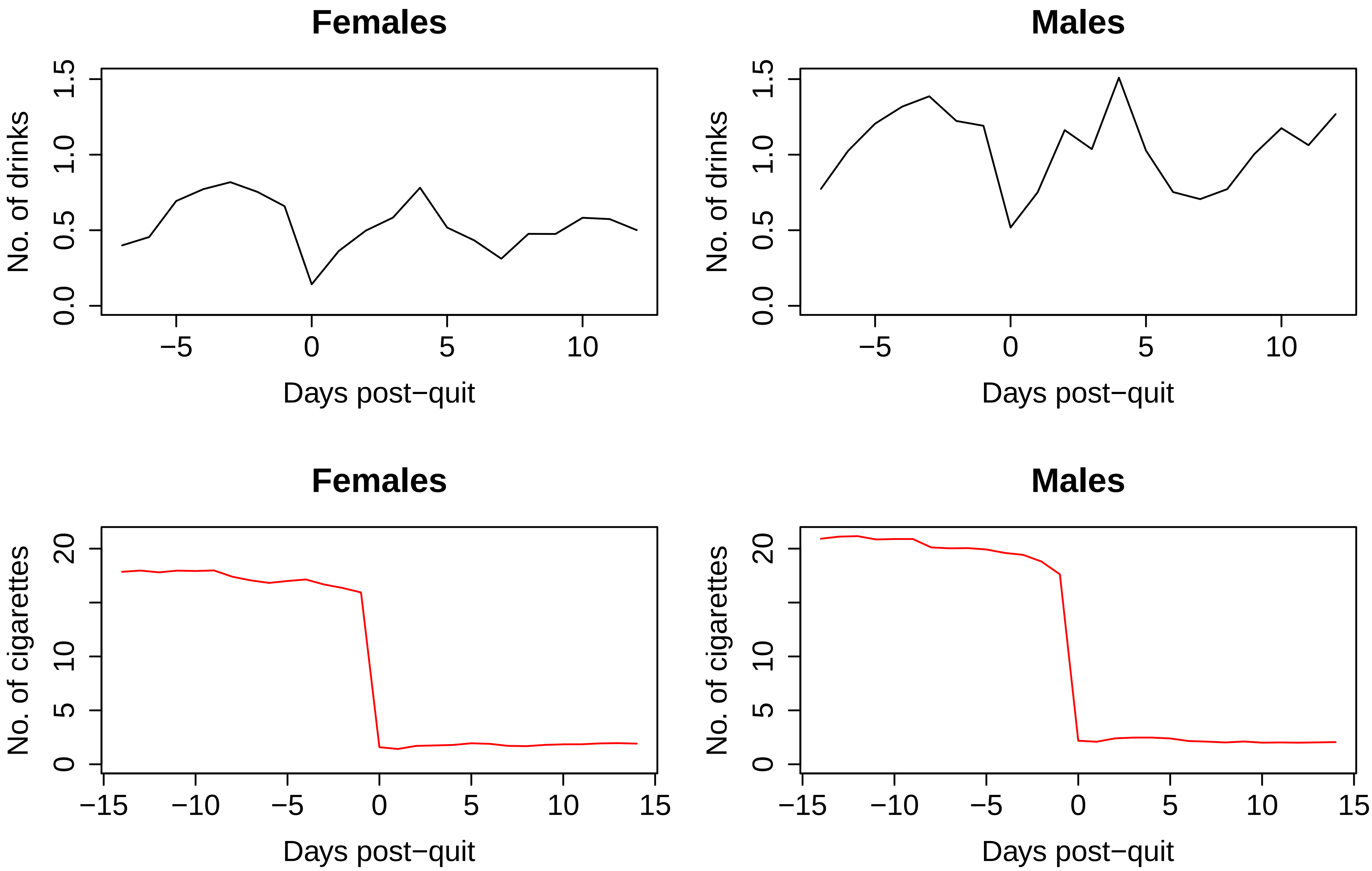
20 drinking and smoking profiles by gender



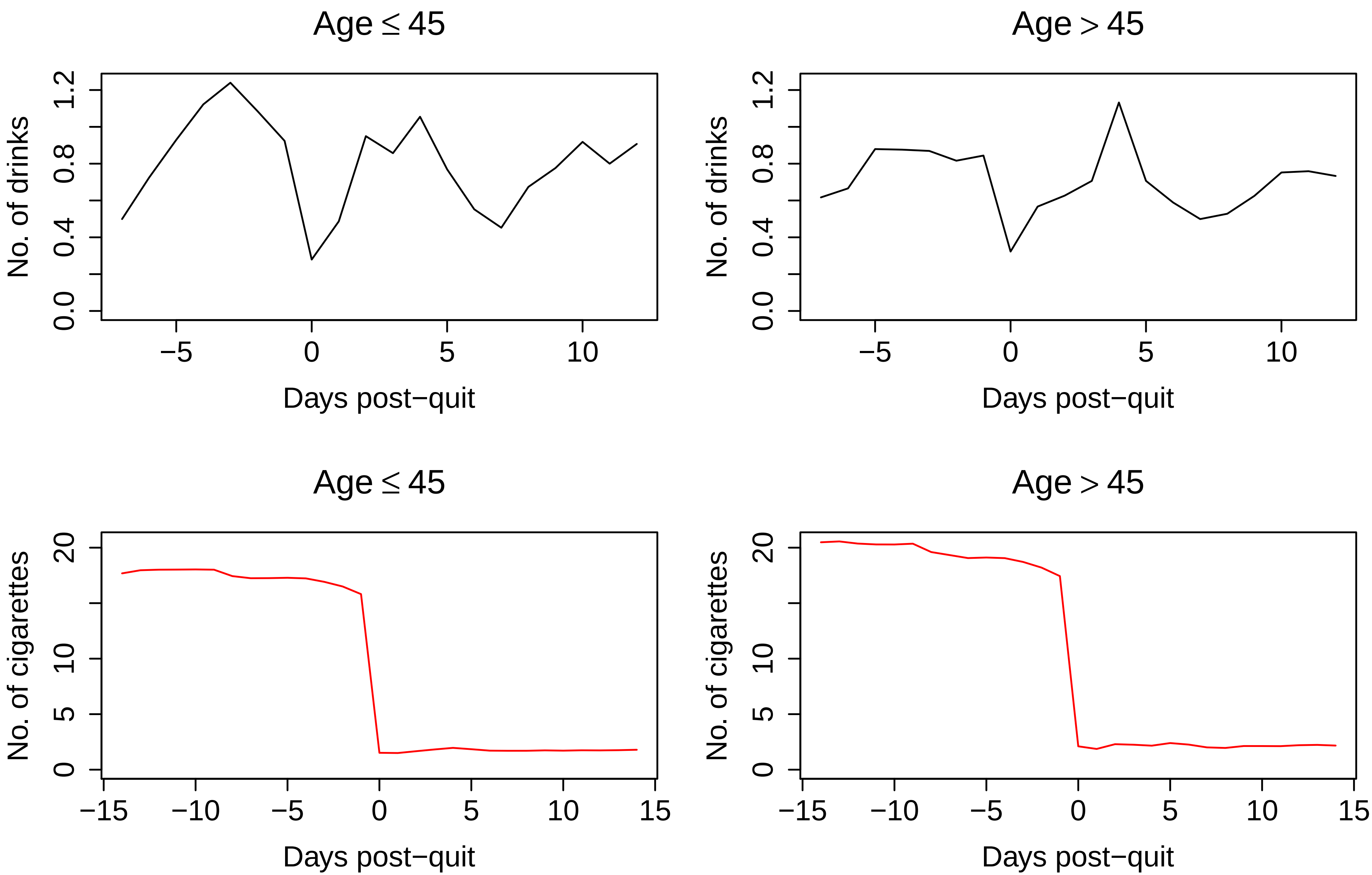
20 drinking and smoking profiles by age group



Mean drinking and smoking profiles by gender



Mean drinking and smoking profiles by age gp



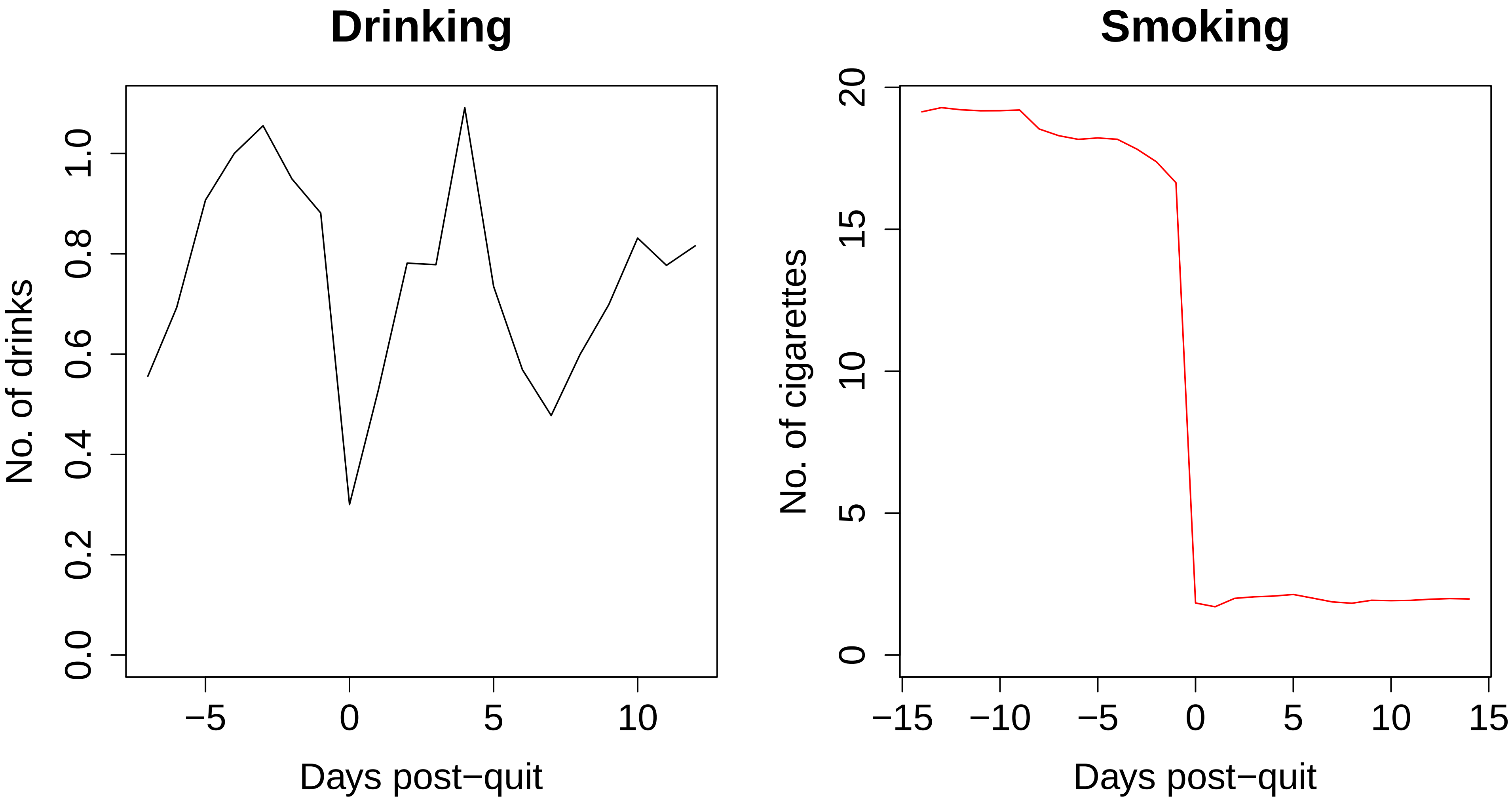
Age ≤ 45

Age > 45

Age ≤ 45

Age > 45

Mean profiles over all subjects



General linear mixed-effects model

Let $\mathbf{y}_i$ be a t-vector for subject i ($i = 1, \ldots, n$). Assume that

$$\begin{aligned} \mathbf{y}_i &= X_i\boldsymbol{\beta} + Z_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i \\ \mathbf{b}_i &\sim N(\mathbf{0}, D) \\ \boldsymbol{\varepsilon}_i &\sim N(\mathbf{0}, \Sigma) \end{aligned}$$

where

1. X_i and Z_i are $(t \times p)$ and $(t \times q)$ matrices of known covariates,
2. $\boldsymbol{\beta}$ is an unknown p-vector containing the fixed effects,
3. $\mathbf{b}_i$ is an unknown q-vector containing the random effects,
4. $\boldsymbol{\varepsilon}_i$ is an unknown t-vector of error components,
5. D and Σ are $(q \times q)$ and $(t \times t)$ unknown covariance matrices,
6. $\mathbf{b}_1, \ldots, \mathbf{b}_n, \boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n$ are mutually independent.

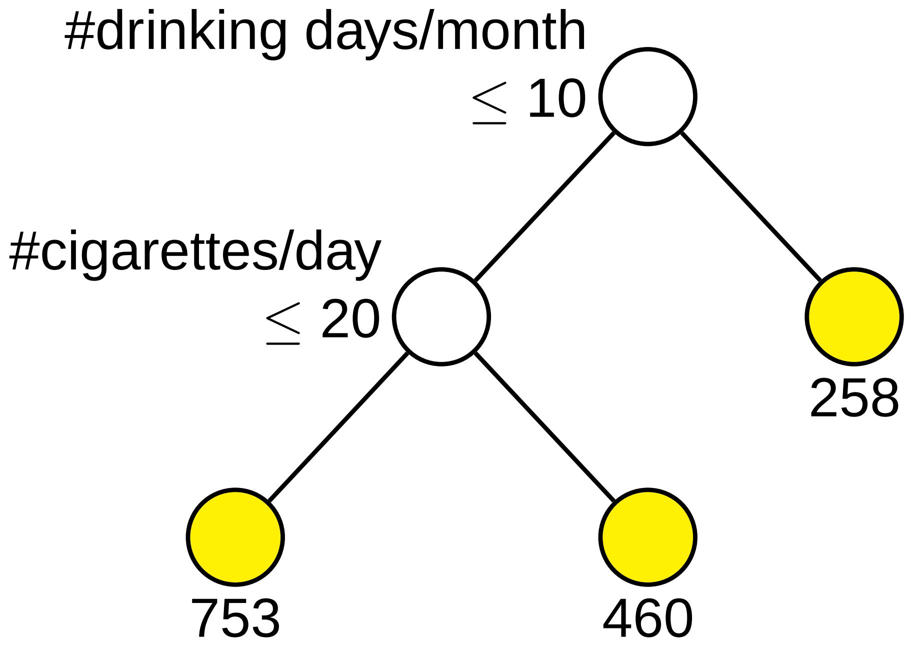
Difficulties with longitudinal data models

1. How to parametrize the drinking and smoking profile shapes?
2. How to justify normal errors when the responses are nonnegative integers? Mixing Poisson distributions with normal random effects is very complicated.
   But even Poisson may not be justifiable.
3. How to deal with so many explanatory variables?
4. How to select variables?
5. How to estimate covariance matrices with so many missing values?
6. How to deal with drinking and smoking profiles jointly?
7. What can be gained from fitting a standard linear mixed-effects model? Tests and confidence intervals? For a model that is almost surely wrong?
8. What do we want anyway?

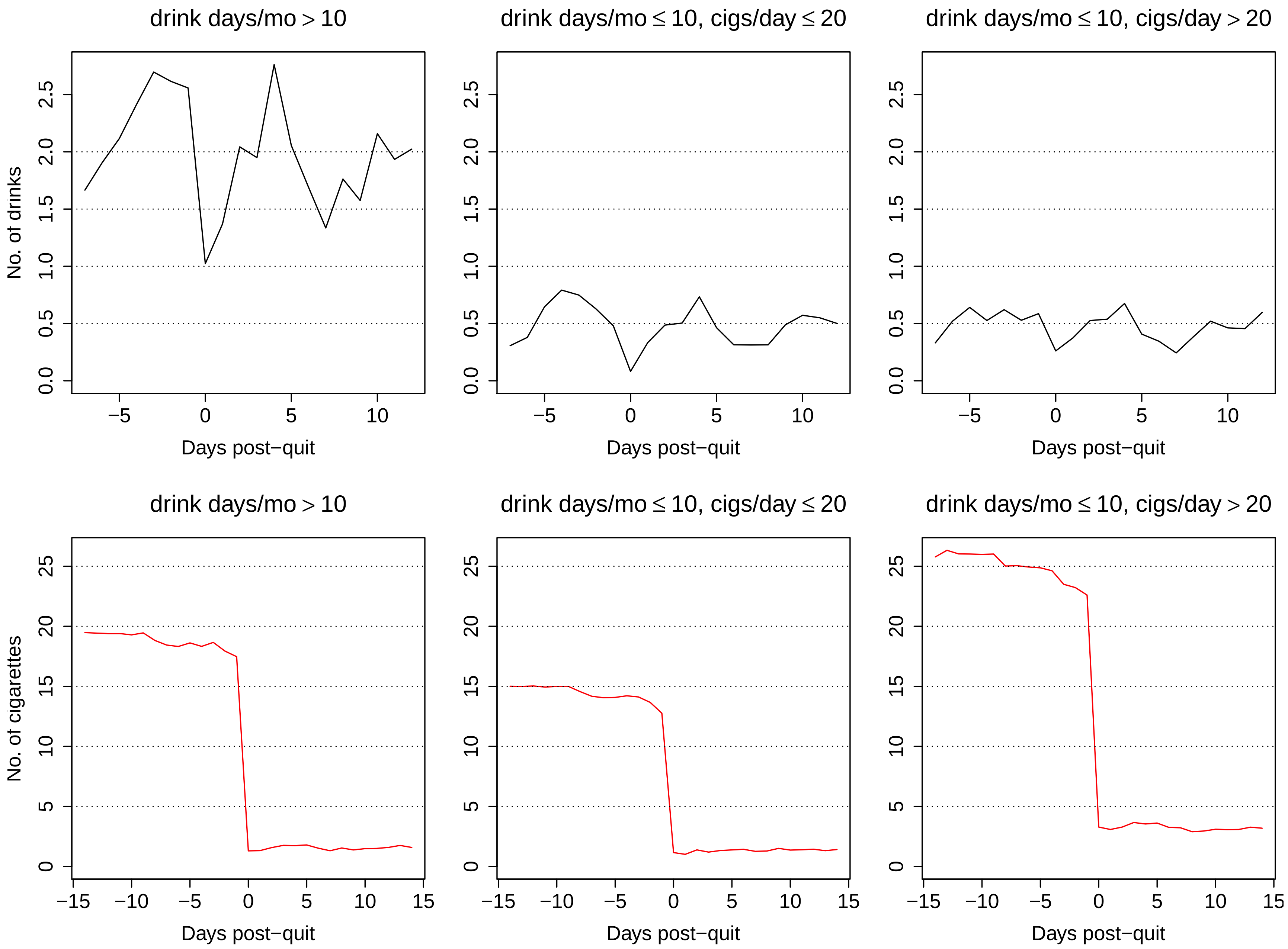
Simpler approaches do not work too

1. One approach is to test for differences in average slopes between subgroups (e.g., by gender, race, marital status, treatment).
2. How to form subgroups for continuous variables (e.g., age, income, years smoked, FTND scores)? How many subgroups?
3. Does a "slope" make sense when profiles are not straight lines?
4. Does a test of significance of a parameter (e.g., slope) make sense if it does not pertain to a statistical model? The "average slope" is then determined by the sample design.
5. What about the multiple testing problem with 100s or 1000s of tests?
6. This approach only tests for main effects. What about interactions?
7. How does this approach lead to a predictive model for the data?
8. What to do if we wish to study drinking and smoking profiles jointly?

GUIDE longitudinal regression tree



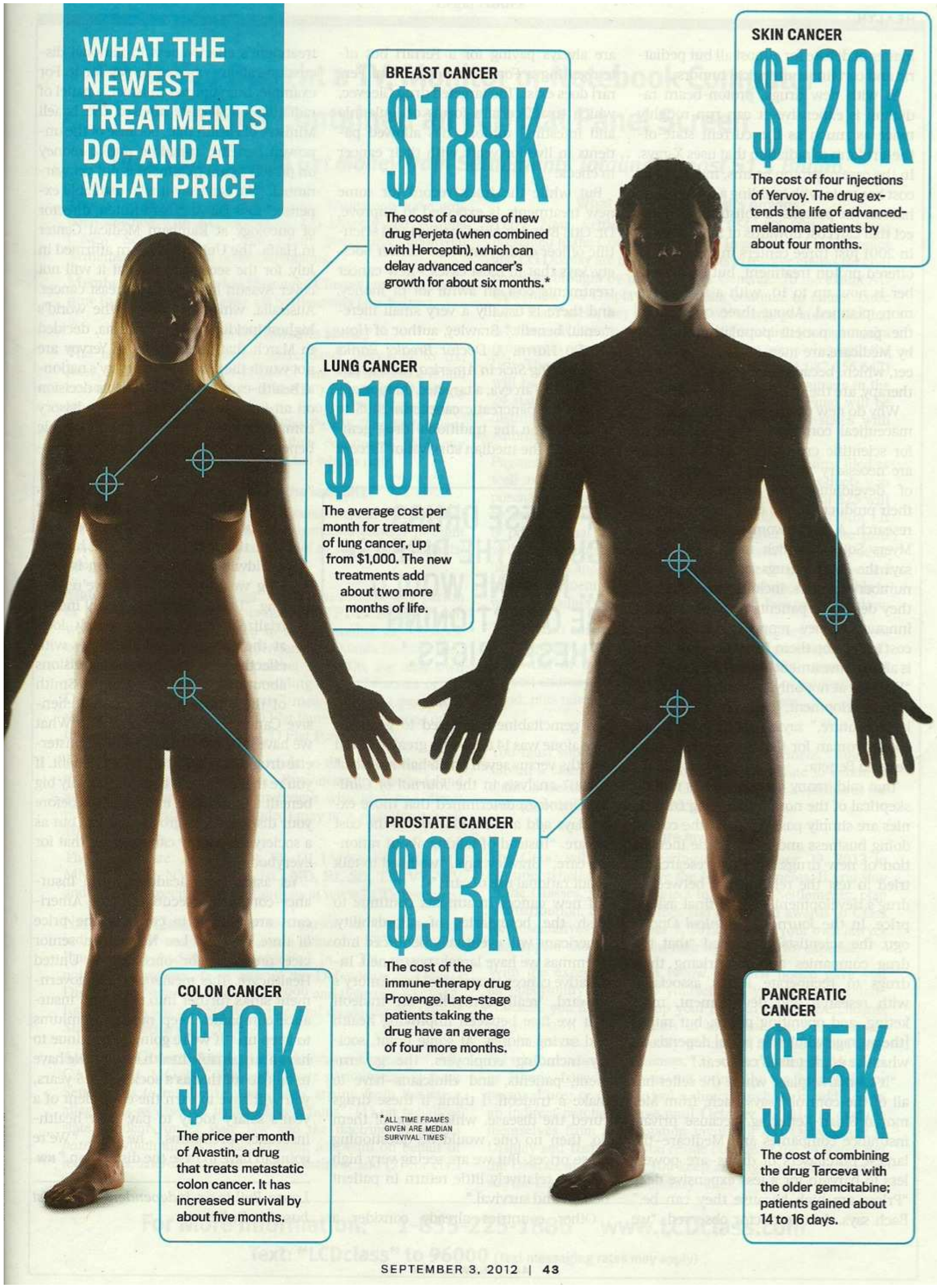
Sample size beneath node


drink days/mo > 10
drink days/mo ≤ 10, cigs/day ≤ 20
drink days/mo ≤ 10, cigs/day > 20
No. of drinks
0.0
0.5
1.0
1.5
2.0
2.5
−5
0
5
10
Days post−quit
drink days/mo > 10
drink days/mo ≤ 10, cigs/day ≤ 20
drink days/mo ≤ 10, cigs/day > 20
No. of cigarettes
0
5
10
15
20
25
−15
−10
−5
0
5
10
15
Days post−quit

Why are drugs so costly?

- It used to take $200 million and 7 years to bring a new drug to market
- Now the cost is $1.5 billion to $2 billion and the timeline can be as long as 15 years

— Wisconsin State Journal, Sep. 13, 2012



WHAT THE NEWEST TREATMENTS DO—AND AT WHAT PRICE

BREAST CANCER

$188K

The cost of a course of new drug Perjeta (when combined with Herceptin), which can delay advanced cancer's growth for about six months.*

SKIN CANCER

$120K

The cost of four injections of Yervoy. The drug extends the life of advanced-melanoma patients by about four months.

LUNG CANCER

$10K

The average cost per month for treatment of lung cancer, up from $1,000. The new treatments add about two more months of life.

PROSTATE CANCER

$93K

The cost of the immune-therapy drug Provenge. Late-stage patients taking the drug have an average of four more months.

COLON CANCER

$10K

The price per month of Avastin, a drug that treats metastatic colon cancer. It has increased survival by about five months.

*ALL TIME FRAMES GIVEN ARE MEDIAN SURVIVAL TIMES

PANCREATIC CANCER

$15K

The cost of combining the drug Tarceva with the older gemcitabine; patients gained about 14 to 16 days.

SEPTEMBER 3, 2012 | 43

Three frontiers in cancer treatment (Newsweek, Sep 3, 2012)

1. Targeted therapies. Traditional chemotherapy is notorious for side effects because it wields destruction indiscriminately throughout the body. Targeted theapies are designed to hit cancer cells only.

- *Perjeta* targets a protein produced in excess amounts in some breast cancers
- *Avastin* hinders the ability of a tumor to form new blood vessels to feed itself

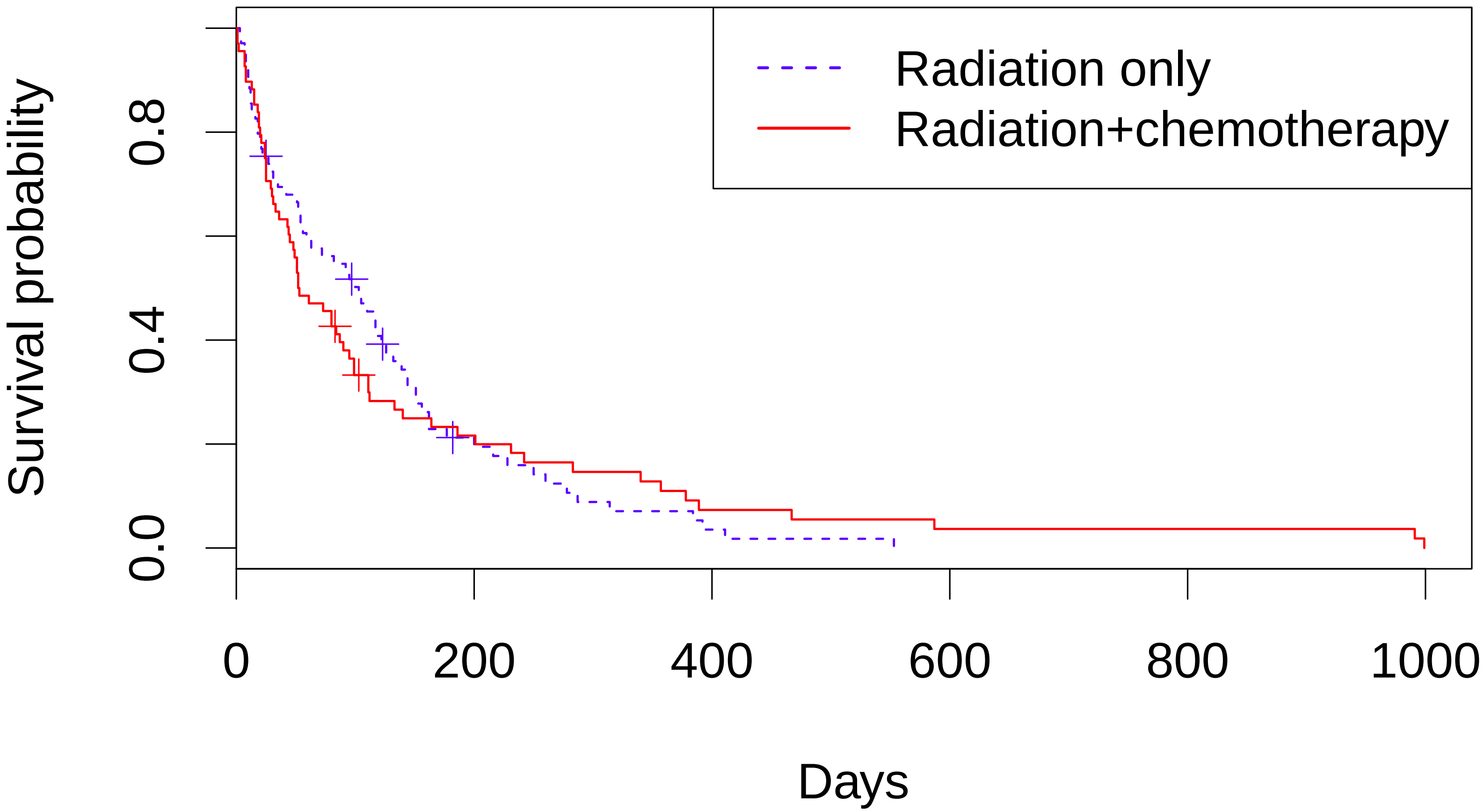
Cancer is more genetically crafty than researchers once imagined

- Scientists may build a drug to hit one target, but a tumor may employ lots of yet-undiscovered genetic tricks to stay alive
- Instead of a magic bullet, any particular tumor may need lots of magic bullets

2. **Immune therapy.** Drugs have appeared that can spur the body’s own immune cells to attack the tumor
   - *Provenge* works this way for prostate cancer
   - *Yervoy* does this for advanced melanoma
3. **Radiation.** Protons instead of X-rays to kill cancer cells
   - Many doctors believe that protons offer better precision and is able to get rid of tumors without collateral damage to nearby tissues
   - But whether it has fewer side effects than traditional radiation is still unclear
   - Proton-beam radiation is expensive; a cyclotron that harvests the protons can cost more than $150 million to build

Lung cancer data

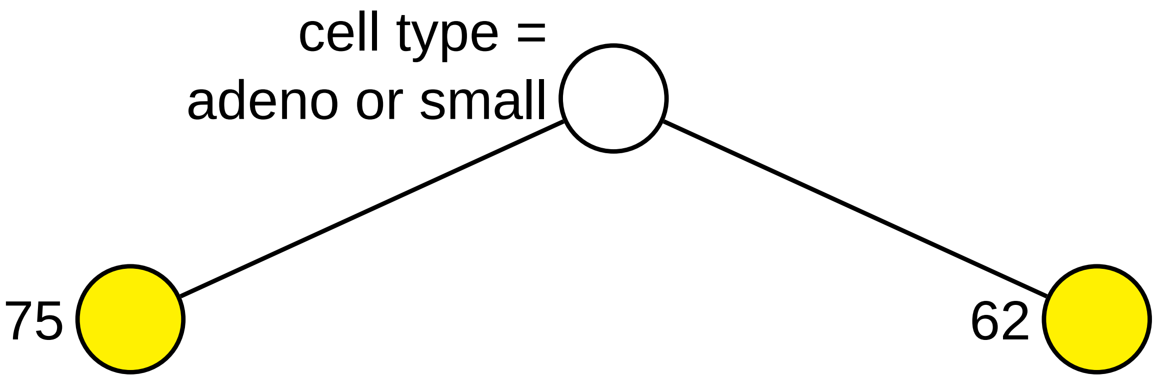
- Randomized clinical trial of 137 male Veterans Administration patients with advanced lung cancer
- Response is days of survival (1–999, 128 uncensored, 9 censored)
- Six predictor variables with no missing values:
  1. Cell type (squamous cell, large cell, small cell, and adenocarcinoma)
  2. Karnofsky performance status (10–99)
  3. Time in months from diagnosis to start of therapy (1–87)
  4. Age in years (34–81)
  5. Prior therapy (0=no, 10=yes)
  6. Treatment (1=radiation, 2=radiation+chemotherapy)
- Kalbfleisch and Prentice (1980, 223–224)

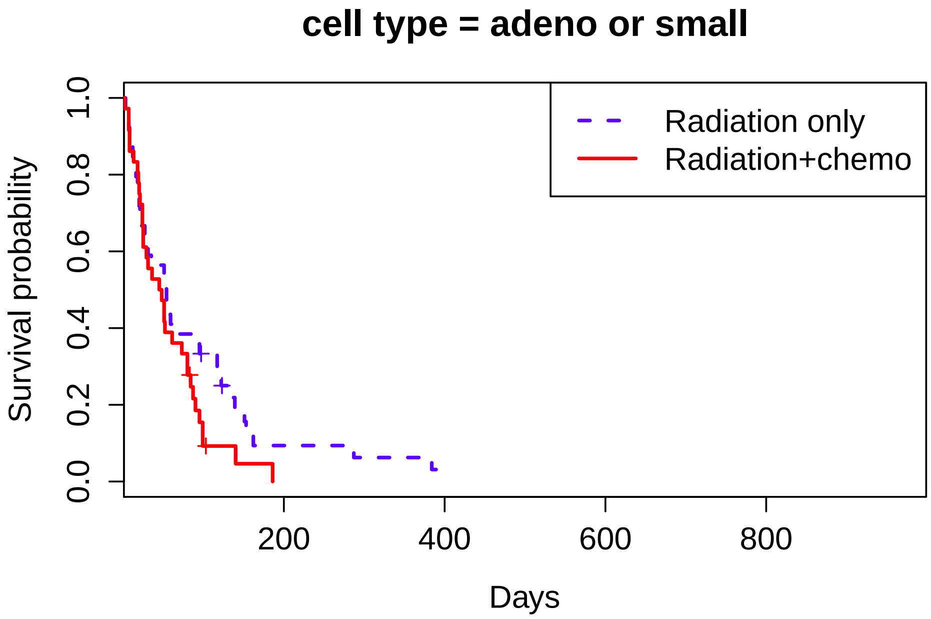


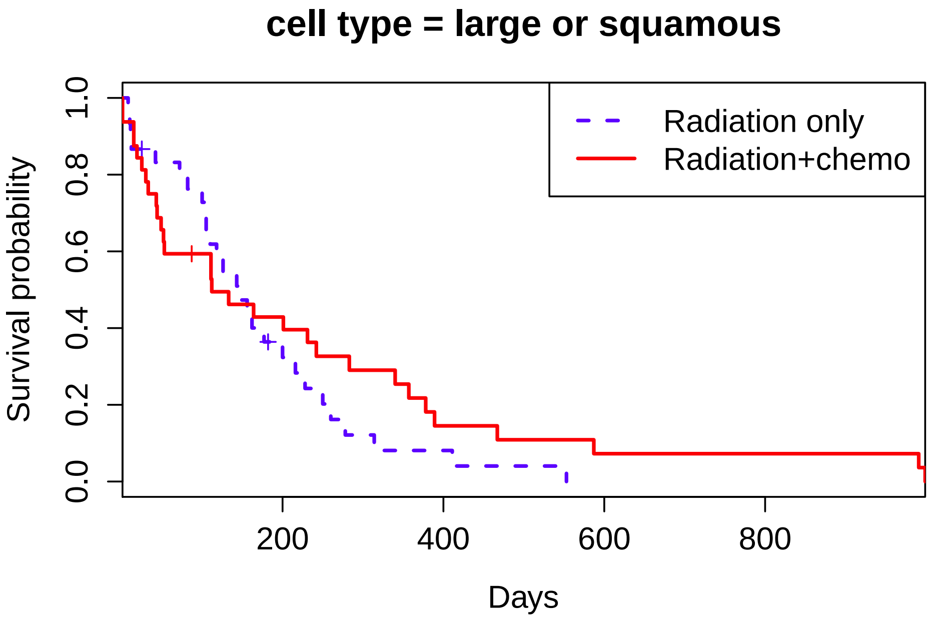
Predictor	Coef	p-val.	Predictor	Coef	p-val.
rx	0.2950	0.16	karno	-0.0328	0.0000
cell=large	-0.7950	0.009	months	0.0001	0.99
cell=small	-0.3350	0.23	age	-0.0087	0.35
cell=squam.	-1.2000	0.0001	priorrx	0.0072	0.76

Subgroup identification

- Is there a subgroup where radiation+chemotherapy is better than radiation?
- Key to personalized medicine or targeted therapy

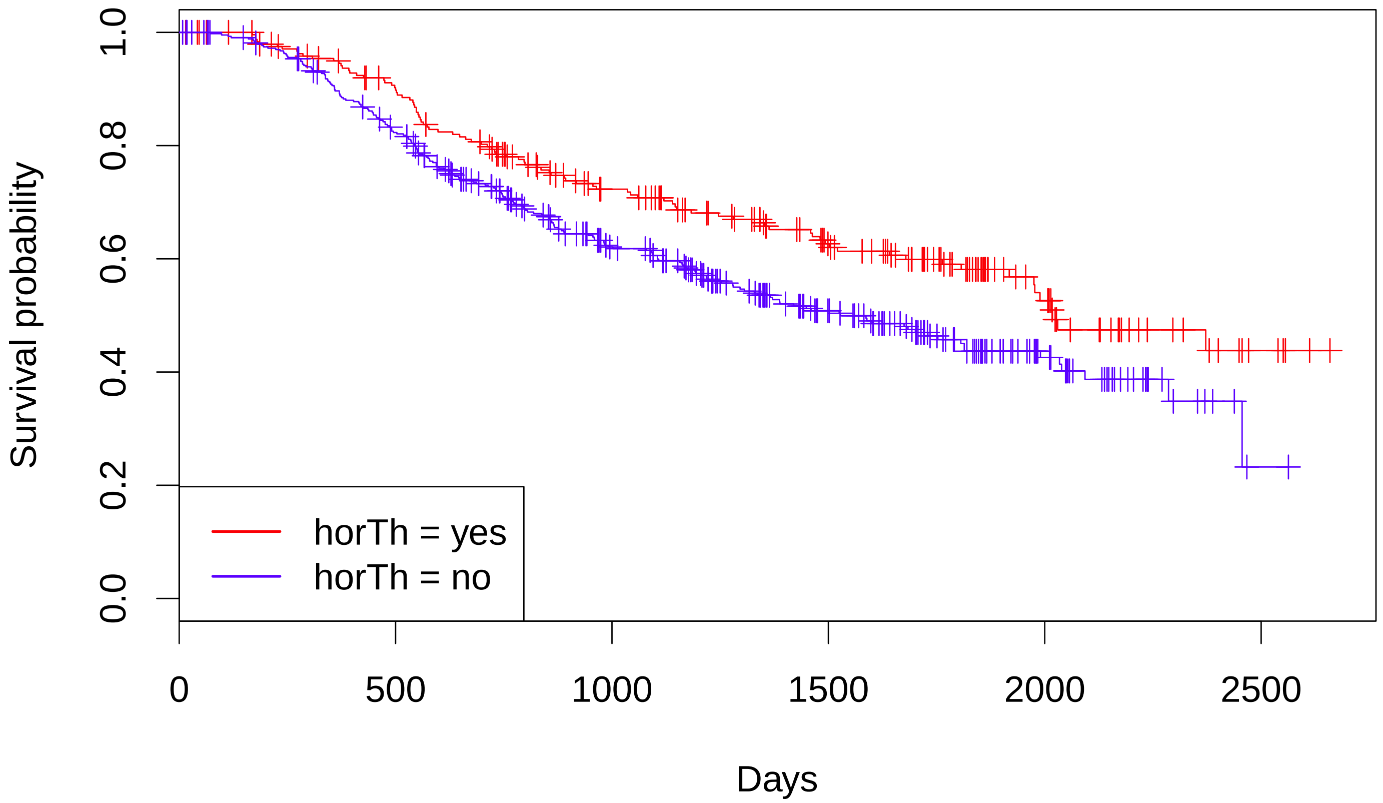



cell type = adeno or small


cell type = large or squamous

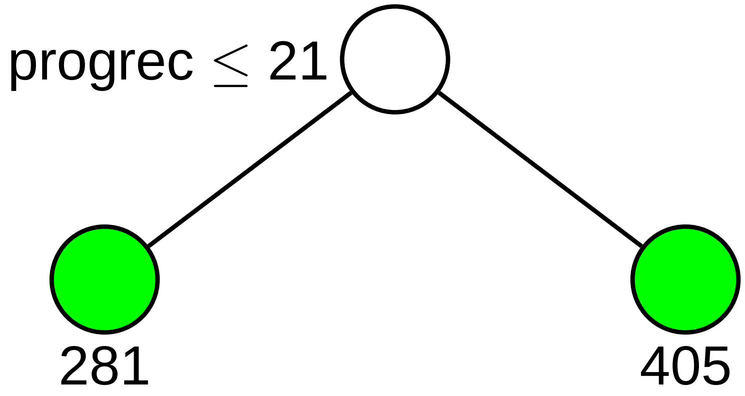
Breast cancer data

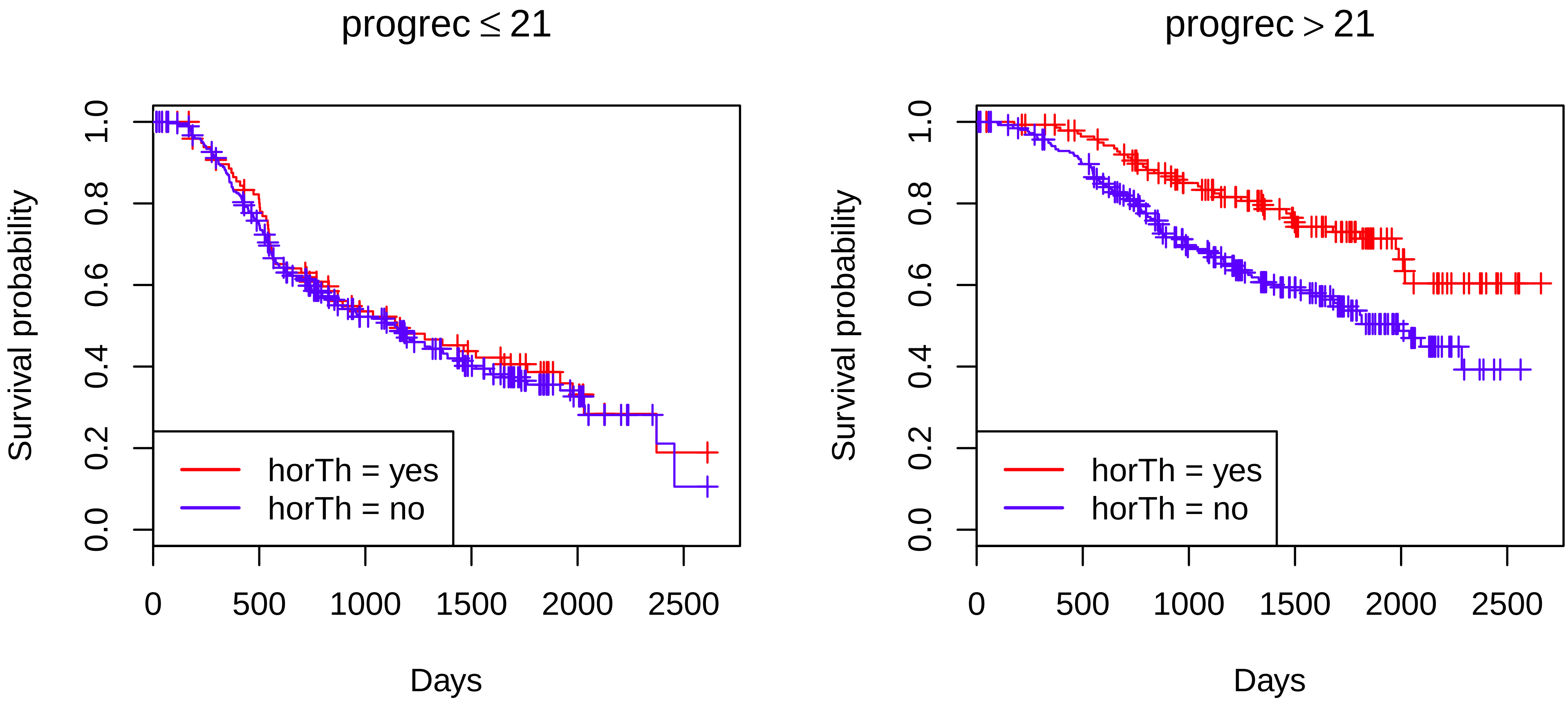
- Randomized clinical trial of 686 subjects with primary node positive breast cancer (Jul 1984—Dec 1989)
- Response is recurrence-free survival time (8–2659 days, 299 uncensored, 387 censored)
- Eight predictor variables with no missing values:
  1. **horTh** (hormone therapy, yes/no)
  2. **age** (21–80 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)
- Schumacher et al. (1994); data from **ipred** R package



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?


progrec $\leq$ 21
281
405

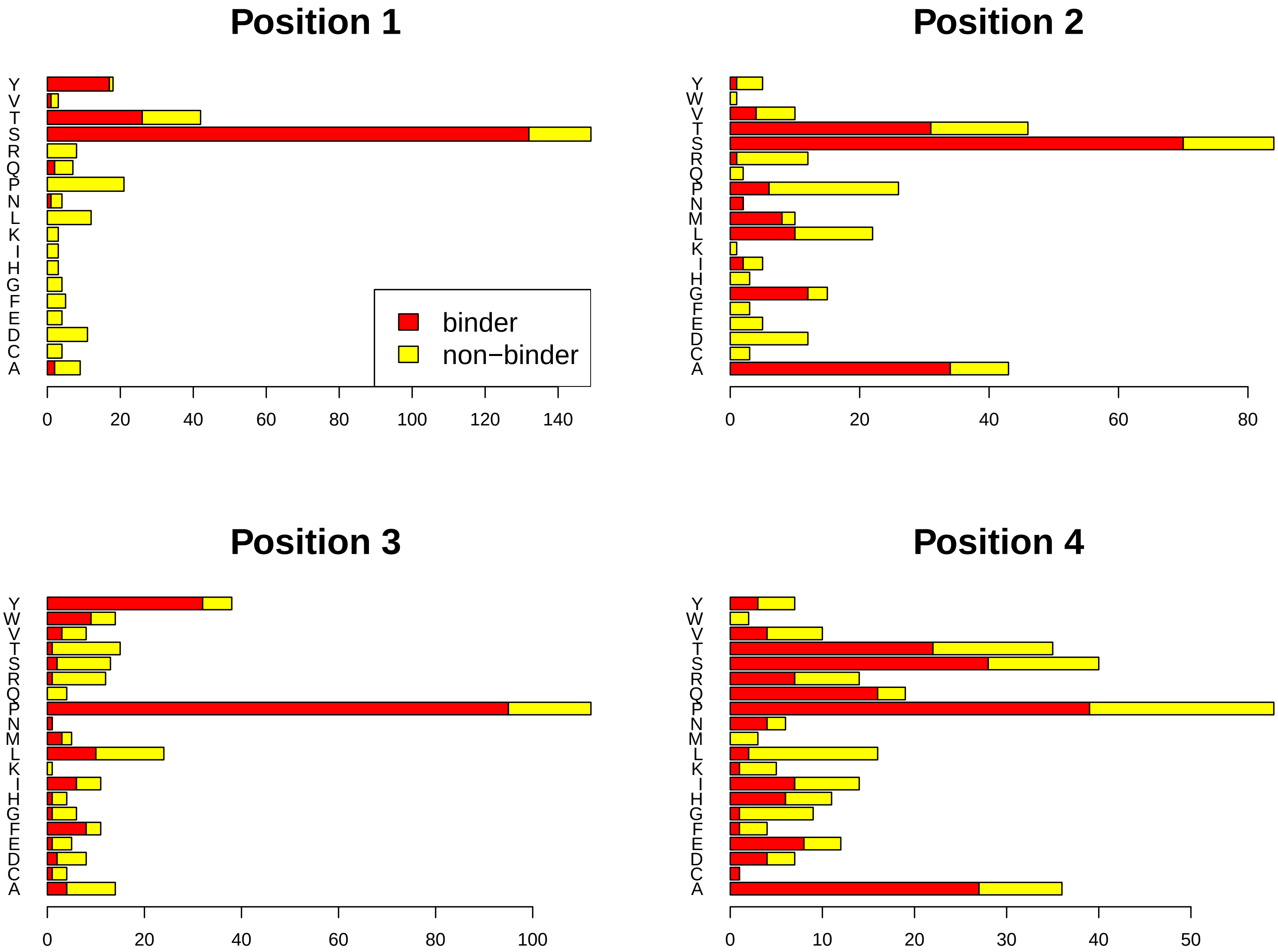

progrec ≤ 21
progrec > 21
Survival probability
0.0
0.2
0.4
0.6
0.8
1.0
0
500
1000
1500
2000
2500
Days
horTh = yes
horTh = no

Classification with 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



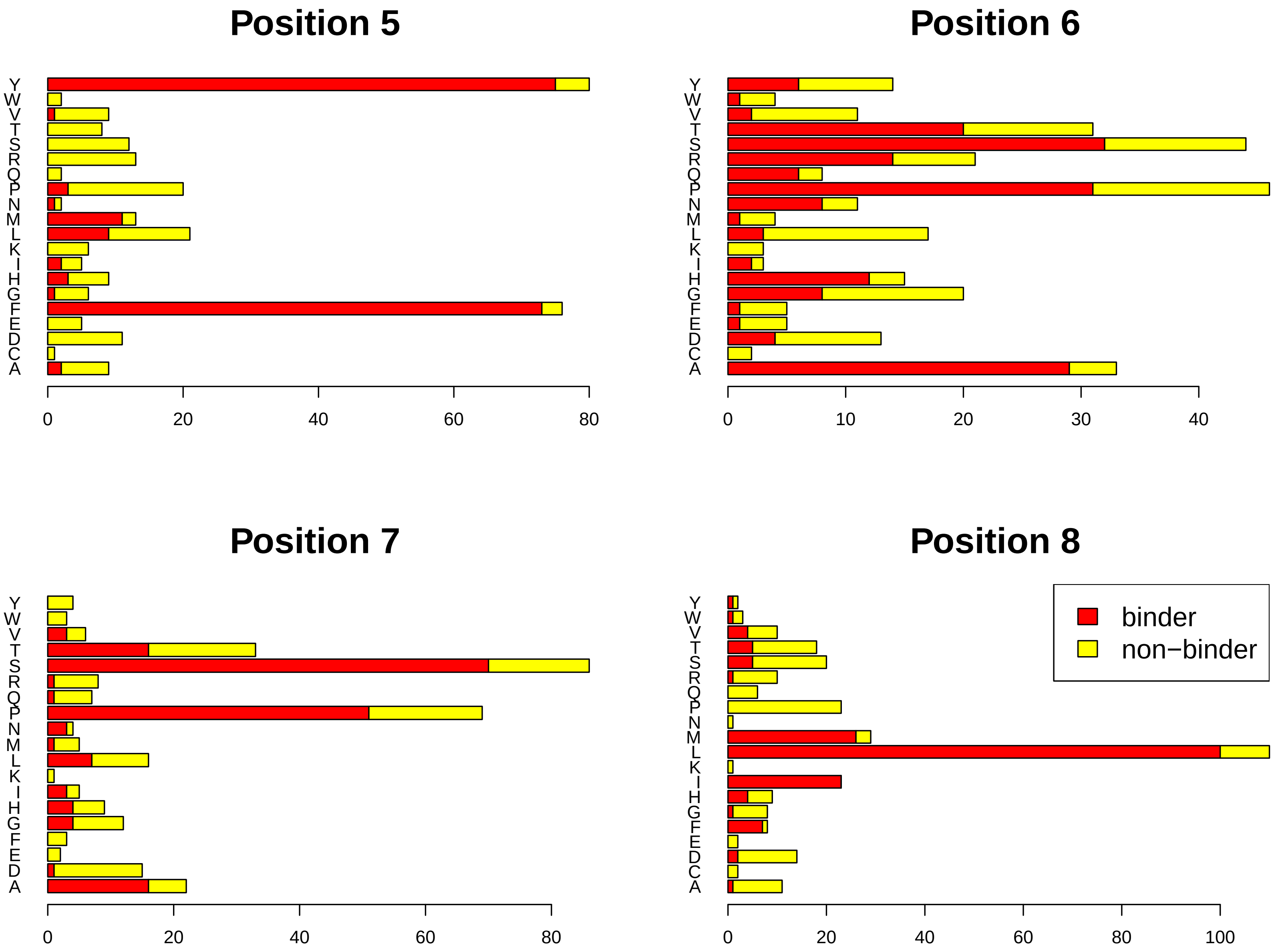
Position 1

Position 2

Position 3

Position 4

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



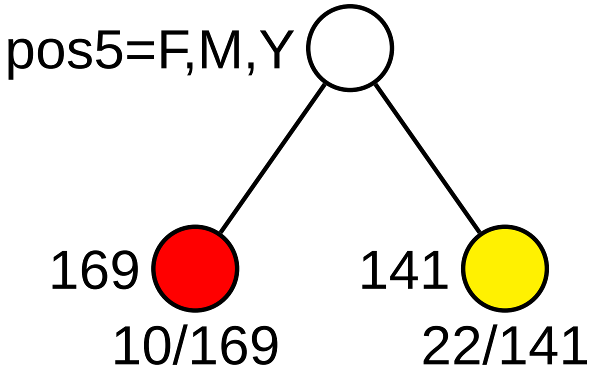
Position 5

Position 6

Position 7

Position 8

GUIDE classification tree for peptide data



Red denotes binder, yellow denotes non-binder

Numbers beneath nodes are misclassified/sample size

Importance scores of position variables

Score	Variable	Rank
46.1	pos1	1
45.9	pos8	2
42.7	pos5	3
18.5	pos2	4
18.2	pos3	5
11.0	pos7	6
6.8	pos6	7
5.7	pos4	8

Variables with scores above 1.0 are considered important

Classification with unequal costs: Credit card data

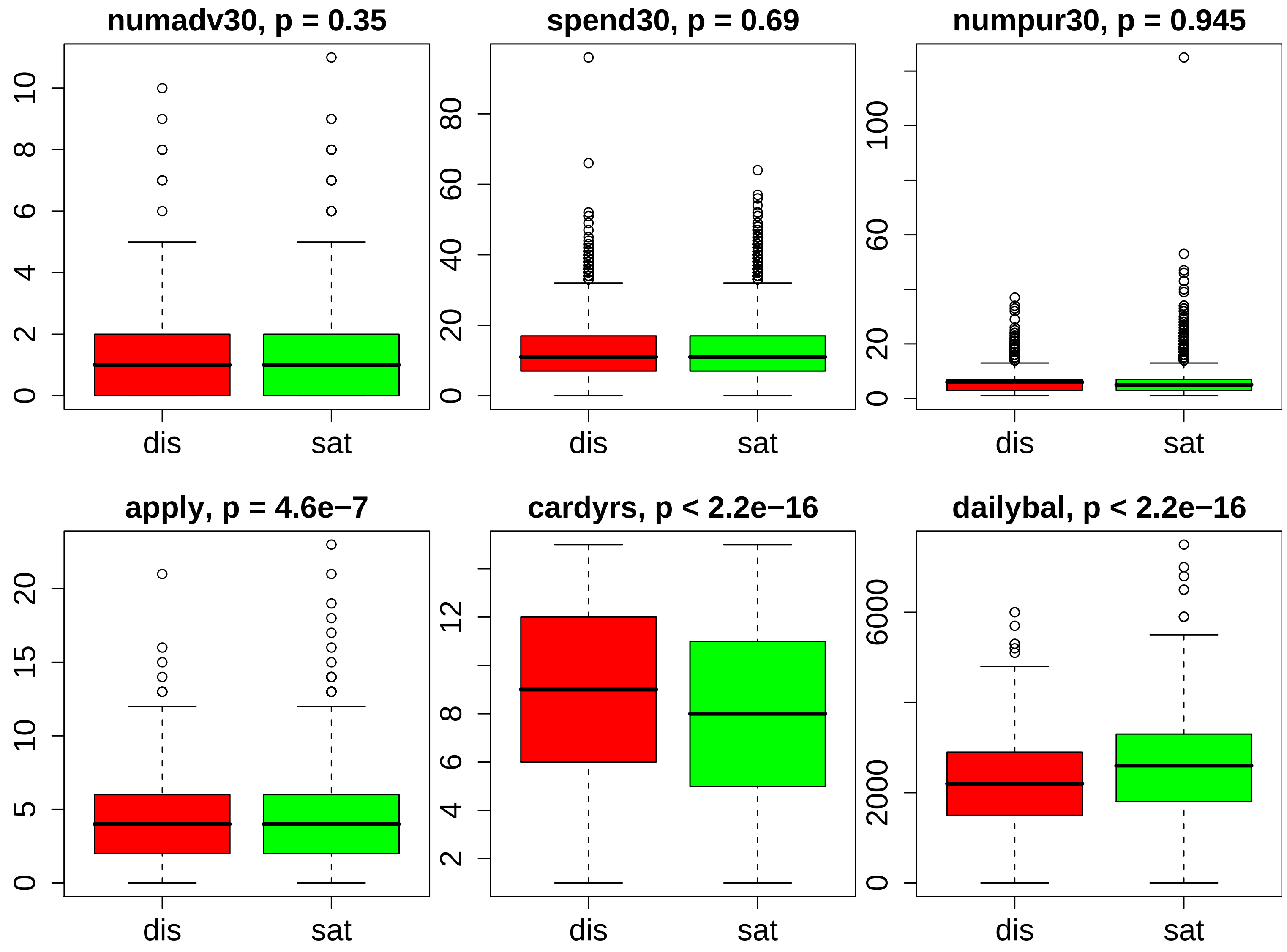
- Goal: A major credit card company wants to find out why 14.8% of its card holders are dissatisfied
- Data: 22,242 card holder records with information on 24 predictor variables
- Missing values: 1,752 records contain one or more missing values; 0.34% missing values overall
- Response variable: whether 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 × that of satisfied (more emphasis on identifying dissatisfied card holders)

Predictor variables for credit card data

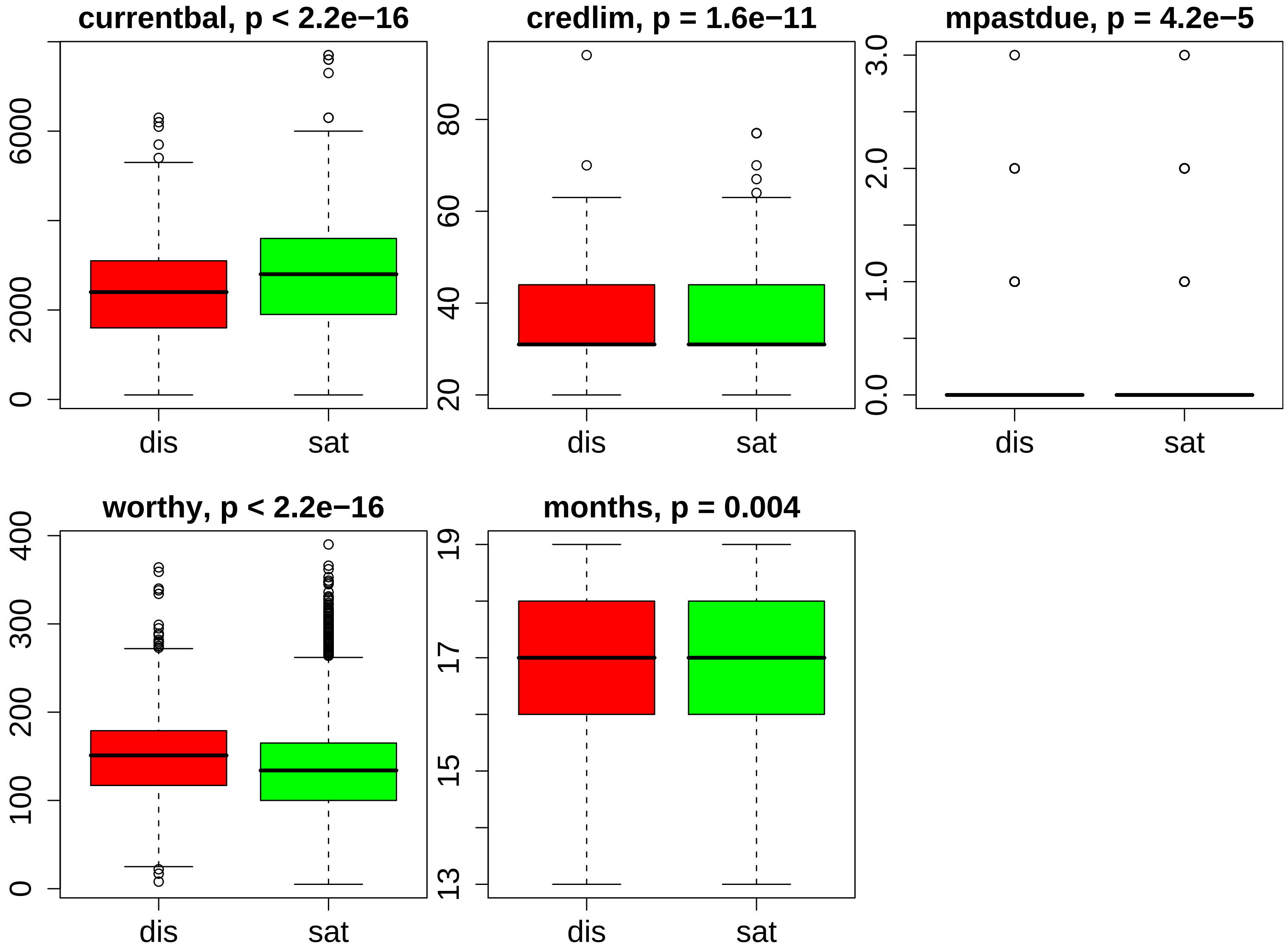
numadv30	How many times did you get cash advances in last 30 days?
spend30	How much money did you spend on purchases in last 30 days? ($)
numpur30	How many times did you make purchases in last 30 days?
over30	Have you gone over limit in last 30 days? (1=yes 0 = no)
otherbal	How much balance do you carry 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



t-tests of ordered predictors (cont'd.)



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 ($p = 1.5 \times 10^{-13}$)								
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.)

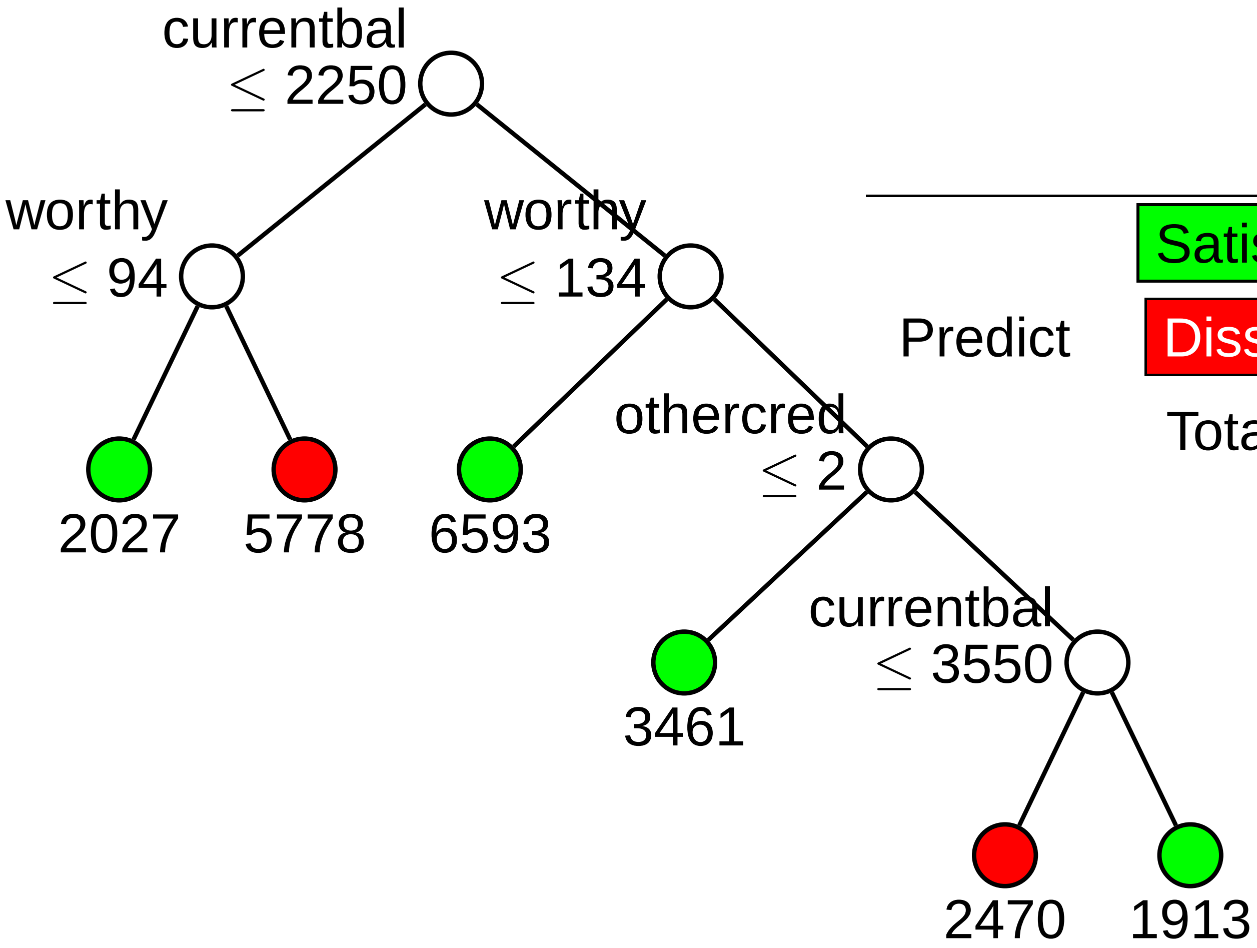
	apr (p = 0.002431)								
Satisfied	4	5	6	7	8	9	10	11	15
Yes	164	5	273	36	459	4	145	17386	482
No	24	6	42	11	59	1	27	3044	74

	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 P(Dissatisfied)

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

Equal priors or unequal cost (5.5:1) GUIDE tree



		True	
		Satis.	Diss.
	Satis.	12545	1449
Predict	Diss.	6409	1839
	Total	18954	3288

Properties of an ideal classifier

High predictive accuracy: able to classify unseen cases with low error

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

Correct, unbiased inference: inferences about predictor variables should be correct and unbiased

Fast training time: classification rule should be reasonably quick to construct

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

Definitions

Definition 1 *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* $1, 2, \ldots, J$

Let

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

Definition 2 *A classifier is a partition of* $\mathcal{X}$ *into* J *disjoint subsets* $A_1, \ldots, A_J$*,* $\mathcal{X} = \cup_j A_j$ *such that for every* $\mathbf{x} \in A_j$*, the predicted class is* j

Definition 3 *A* ***learning*** *or* ***training sample*** $\mathcal{L}$ *consists of data* $(\mathbf{x}_1, j_1), \ldots, (\mathbf{x}_N, j_N)$ *on* N *cases where* $\mathbf{x}_n \in \mathcal{X}$ *and* $j_n \in \mathcal{C}$*,* $n = 1, \ldots, N$*, i.e.,*

$$\mathcal{L} = \{(\mathbf{x}_1, j_1), \ldots, (\mathbf{x}_N, j_N)\}$$

Types of predictor variables

- A variable is called **categorical** or **nominal** if it takes values in a finite set not having any natural ordering (e.g., hair color, occupation, marital status)
- A variable is called **non-categorical** if its values are ordered and they can be represented as numbers (e.g., age, income, severity of pain)

Standard classification methods

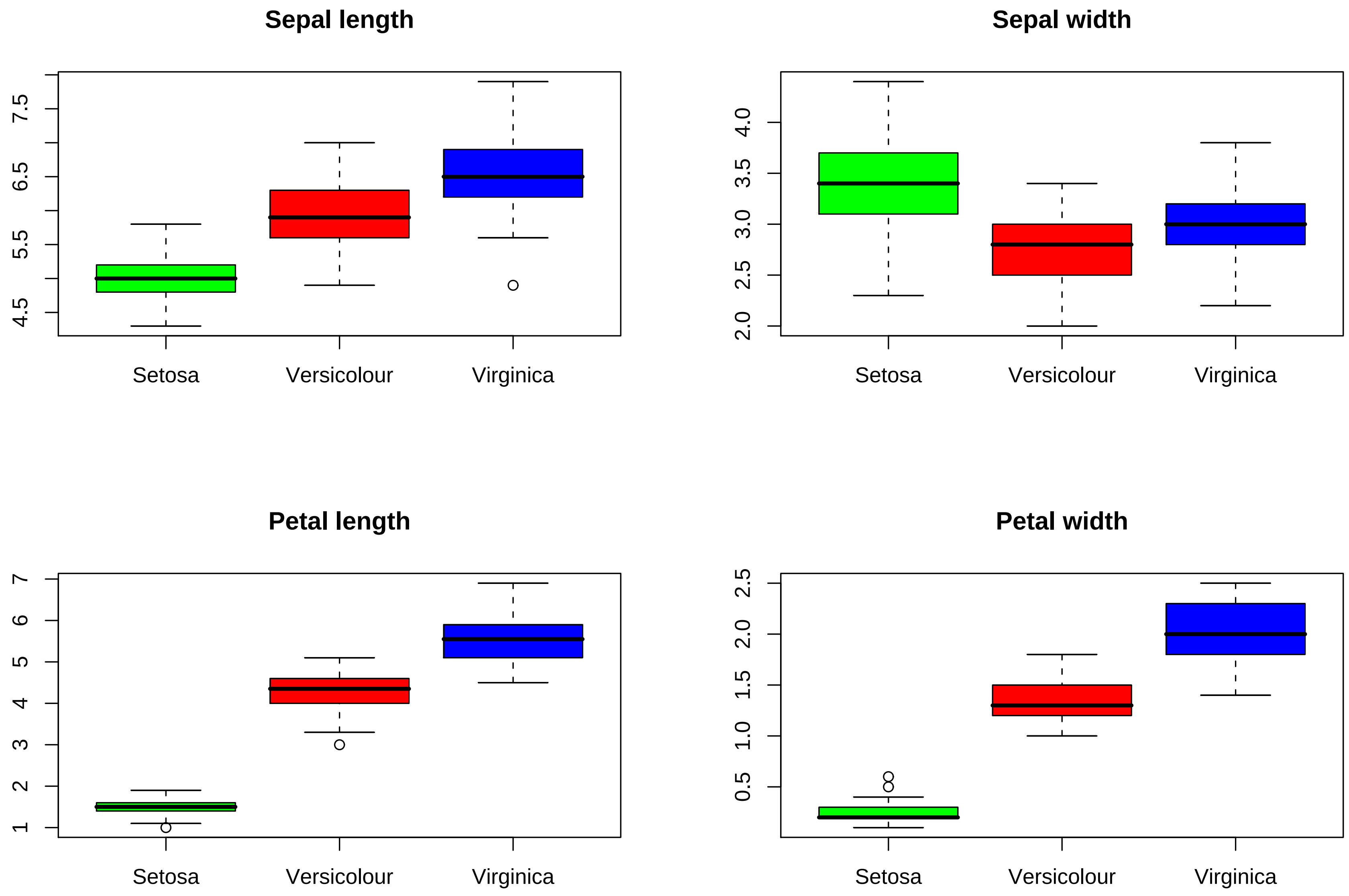
- Linear discriminant analysis (LDA)
- Quadratic discriminant analysis (QDA)
- Density estimation methods
- Nearest-neighbor methods
- Logistic regression
- Neural networks
- Support vector machines
- Fuzzy set theory

These may produce accurate classifiers, but they do not provide much insight into the roles of the variables

Fisher’s iris data

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

Boxplots of iris data



LDA for two groups

- Let $\bar{\mathbf{x}}_i$, $\mathbf{S}_i$ be the sample mean and covariance matrix of group $i = 1, 2$
- Let $\mathbf{S} = (n_1 + n_2 - 2)^{-1}[(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2]$
- Let $z = a_1 x_1 + \ldots + a_p x_p = \mathbf{a}'\mathbf{x}$
- The two-sample t-statistic based on z is

$$t_{\mathbf{a}} = \frac{\mathbf{a}'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)}{\sqrt{\mathbf{a}'\mathbf{S}\mathbf{a}(n_1^{-1} + n_2^{-1})}}$$

- Fisher's LDA chooses $\mathbf{a}$ to maximize

$$t_{\mathbf{a}}^2 = \left(\frac{n_1 n_2}{n_1 + n_2}\right)\left(\frac{\mathbf{a}'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{a}}{\mathbf{a}'\mathbf{S}\mathbf{a}}\right)$$

- Maximizing $\mathbf{a}$ is proportional to $\mathbf{S}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$

LDA for g groups

- Let $\bar{\mathbf{x}}_i$, $\mathbf{S}_i$ be the sample mean and covariance matrix of group i based on sample size n_i ($i = 1, \ldots, g$)

- Let $n = \sum_i n_i$, $\bar{\mathbf{x}} = n^{-1} \sum_i n_i \bar{\mathbf{x}}_i$ and

$$\mathbf{W} = \frac{\sum_i (n_i - 1)\mathbf{S}_i}{n - g}, \quad \mathbf{B} = \frac{\sum_i n_i (\bar{\mathbf{x}}_i - \bar{\mathbf{x}})(\bar{\mathbf{x}}_i - \bar{\mathbf{x}})'}{g - 1}$$

- One-way ANOVA F-statistic for $z = \mathbf{a}'\mathbf{x}$ is $F_{\mathbf{a}} = (\mathbf{a}'\mathbf{B}\mathbf{a})/(\mathbf{a}'\mathbf{W}\mathbf{a})$

- Maximizing $\mathbf{a}$ is eigenvector associated with largest eigenvalue of $\mathbf{W}^{-1}\mathbf{B}$

- Let r be the number of positive eigenvalues ($r \leq g - 1$)

- Discriminant coords are $z_i = \mathbf{a}_i'\mathbf{x}$, where $\mathbf{a}_1, \ldots, \mathbf{a}_r$ are eigenvectors with eigenvalues $c_1 \geq \ldots \geq c_r > 0$ satisfying

$$\mathbf{a}_j'\mathbf{W}\mathbf{a}_k = 0, \quad j \neq k$$

LDA for g groups (cont'd.)

- Let $\mathbf{T}$ be the sample covariance matrix for the $\mathbf{x}$ variables ignoring the class variable, i.e., $\mathbf{T} = \mathbf{B} + \mathbf{W}$
- If $\mathbf{T}^{-1}$ exists, the discriminant coord directions are also the solutions of the generalized eigenvalue problem

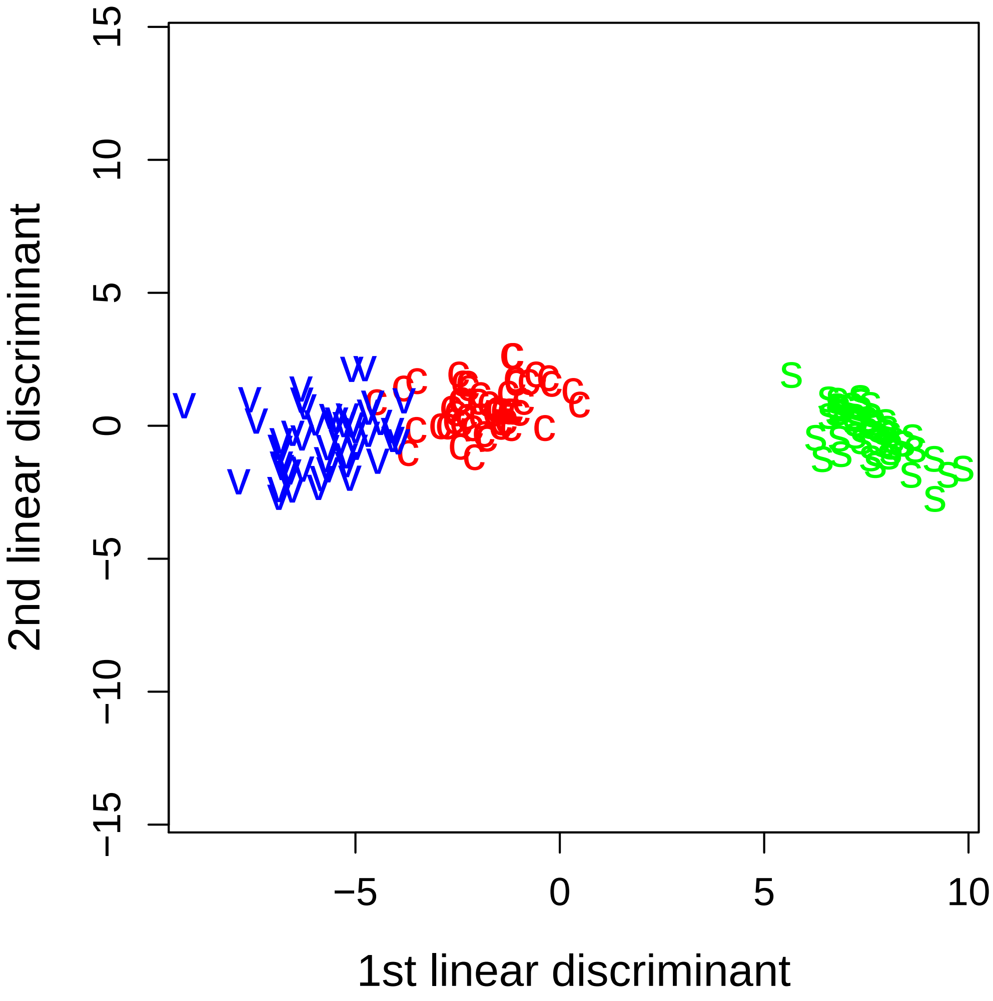
$$\mathbf{Ba} = \lambda \mathbf{Ta}$$

Linear discriminant functions

Variable	Setosa	Versicolour	Virginica
Sepal length	2.354	1.570	1.245
Sepal width	2.359	0.707	0.369
Petal length	-1.643	0.521	1.277
Petal width	-1.740	0.643	2.108
Constant	-86.308	-72.853	-104.368

Classify into the group with largest linear discriminant function

Plot of iris data in first 2 discriminant coords



s = Setosa, c = Versicolour, v = Virginica

Misconceptions and difficulties with LDA

1. LDA cannot be used if data are not multivariate normal with constant covariance matrix — FALSE
   - Normality ensures optimality but LDA often gives reasonable results even if data are non-normal or covariance matrix is not constant
2. LDA cannot be used for categorical variables — FALSE
3. Solution in terms of linear or quadratic discriminant functions is hard to interpret — TRUE

Difficulties with histogram density estimation

- Curse of dimensionality: Number of cells increases rapidly with number of dimensions. Thus a very large sample size is needed to prevent cells from having zero observations.
- Each cell boundary is a discontinuity. Beyond boundary cells, estimate falls abruptly to zero.

Difficulties with kernel density estimation

Density estimate with kernel φ and bandwidth h has the form

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

where $\int h^{-1} \varphi(|x|/h)\, dx = 1$ and $\varphi(0) = \max \varphi(|x|)$

1. How to choose bandwidth h for finite sample size?
   (a) Optimal h depends on criterion (e.g., mean squared error or integrated mean squared error)
   (b) Optimal h for one class may not be optimal for another
   (c) Optimal h depends on form of kernel
   (d) What is optimal in one part of $\mathcal{X}$ may not be optimal in another
2. How to choose form of kernel φ?

Difficulties with k-nearest-neighbor

1. Data must be stored and recalled each time a new case is classified—the classifier cannot be constructed beforehand

2. Computation of nearest-neighbor distances expensive in high dimensions

3. Choice of metric (distance) usually arbitrary

4. What metric for categorical variables?

5. Choice of k is unknown for finite n

Difficulties with logistic regression

- Hard to determine parametric form of model
- No really effective goodness of fit tests and few model selection techniques
- Usually assumes class probabilities are polynomial functions of predictor variables
- Categorical predictors are transformed to 0-1 dummy vectors, possibly causing a large number of degrees of freedom to be used up

Difficulties with neural networks, support vector machines and fuzzy set theory

- Classifiers are like black boxes — hard to interpret
- Difficult to choose the network topology and initial weights
- Categorical predictors are treated via dummy vectors as in logistic regression

Estimates of misclassification error

Resubstitution estimate:

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

This is usually overly optimistic

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

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

This is unbiased and computationally efficient

V-fold cross-validation estimate:

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

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

4. The V-fold cross-validation estimate is

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

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,

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

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

Nonnegative decrease in impurity

Theorem 1 *Let $\phi(p_1, \ldots, p_J)$ be a strictly concave function on $0 \le p_j \le 1$, $j = 1, \ldots, J$, and $\sum_j p_j = 1$. Let*

$$i(t) = \phi(p(1|t), \ldots, p(J|t)).$$

For any split s,

$$\Delta i(s, t) \ge 0$$

with equality if and only if

$$p(j|t_L) = p(j|t_R) = p(j|t), \quad j = 1, \ldots, J$$

Proof: Breiman et al. (1984, pp. 126–127)

The Gini index and entropy possess this property

Shortcut for categorical splits with 2 classes

Theorem 2 *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*

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

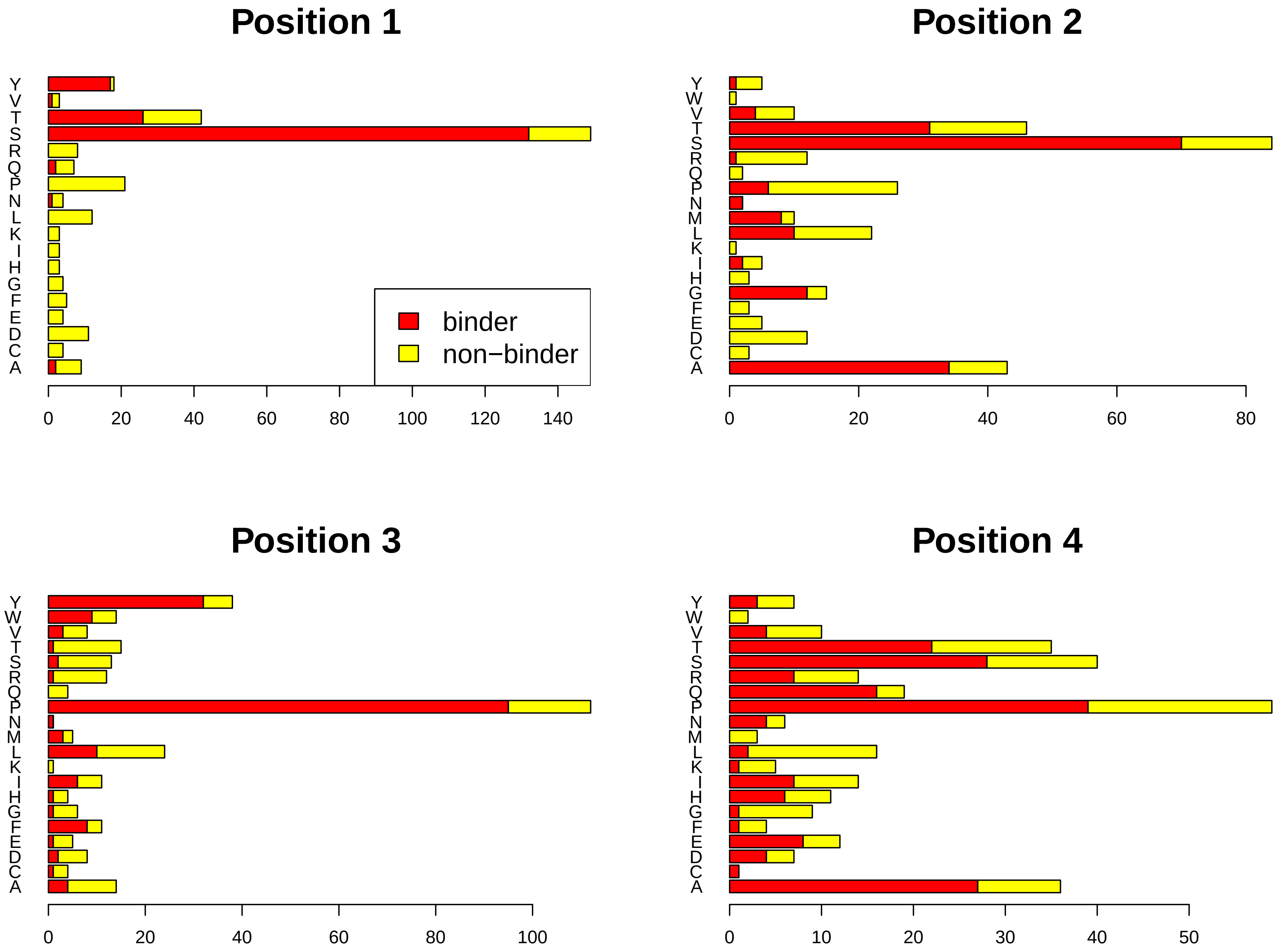
Then the split on X that maximizes the decrease in impurity is one of the splits:

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

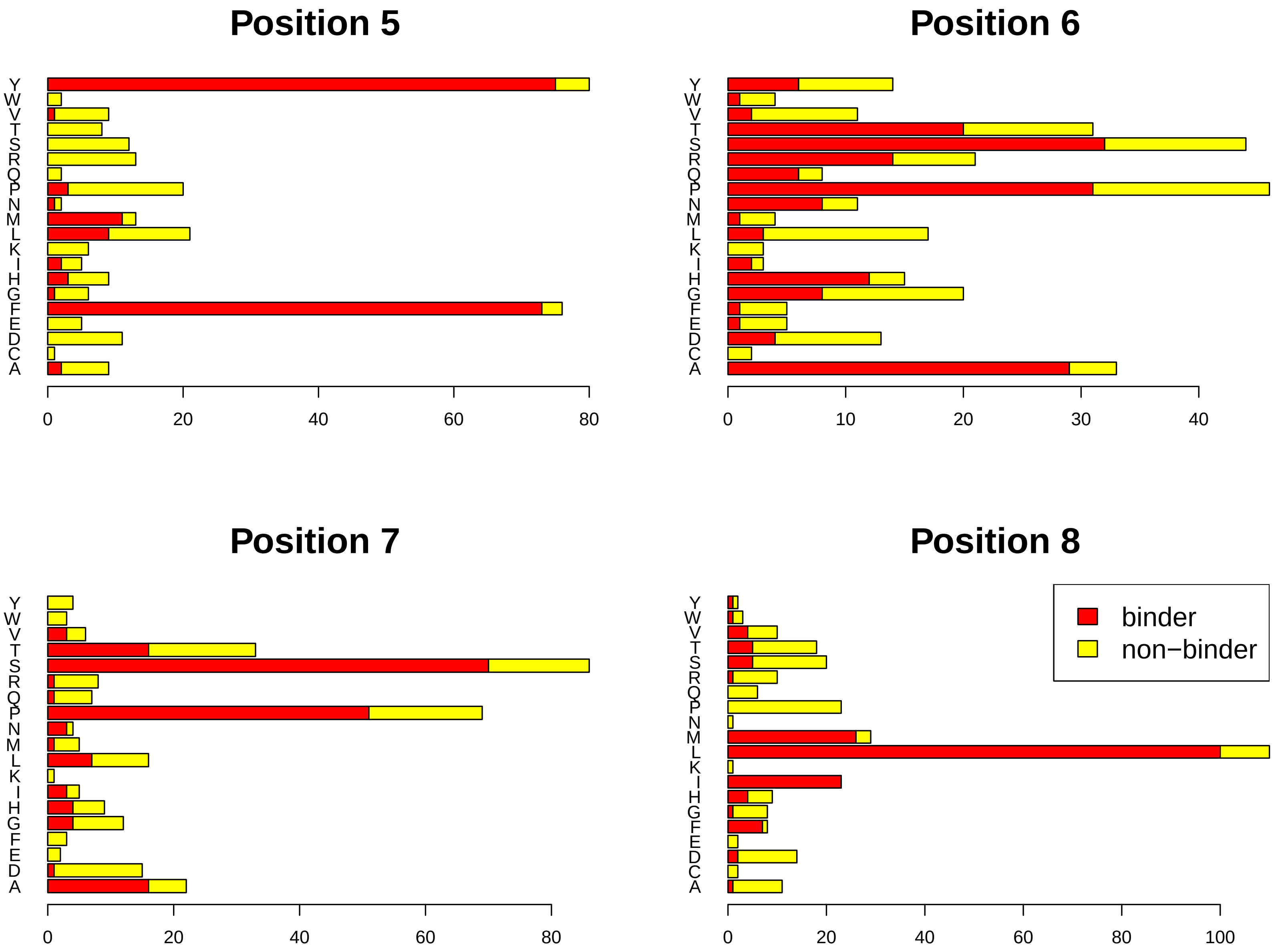
Proof: Breiman et al. (1984, Section 9.4)

Note: This result reduces the search from $2^{L-1} - 1$ subsets to $L - 1$ subsets

distributions of peptide-binding data



distributions of peptide-binding data (cont’d.)


Position 5

Position 6

Position 7

Position 8

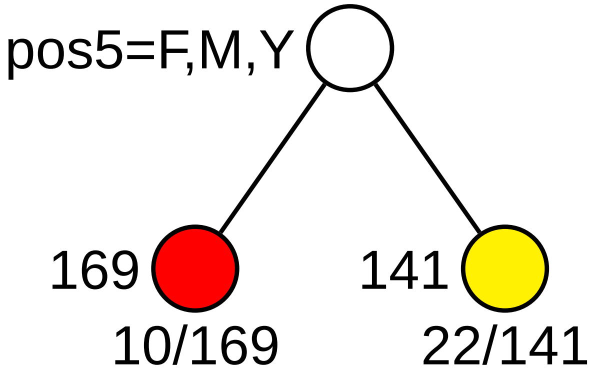
Ordered levels of Pos1 by P(Y = 0)

	Class					Class			
Level	0	1	Total	Prop.	Level	0	1	Total	Prop.
Y	1	17	18	0.056	E	4	0	4	1
S	17	132	149	0.114	F	5	0	5	1
T	16	26	42	0.381	G	4	0	4	1
V	2	1	3	0.667	H	3	0	3	1
Q	5	2	7	0.714	I	3	0	3	1
N	3	1	4	0.75	K	3	0	3	1
A	7	2	9	0.778	L	12	0	12	1
C	4	0	4	1	P	21	0	21	1
D	11	0	11	1	R	8	0	8	1

Ordered levels of Pos5 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

GUIDE classification tree for peptide data



Red denotes binder, yellow denotes non-binder

Numbers beneath nodes are misclassified/sample size

CART approach (Breiman et al., 1984)

1. Choose X and S *simultaneously* to find split $X \in S$ that maximizes decrease in node impurity (Gini index for classification, sum of squared errors for piecewise constant regression)

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

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

3. Prune tree using test sample or cross-validation

4. Use surrogates splits to deal with missing values

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 function

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

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

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

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

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

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

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 test-sample estimation

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

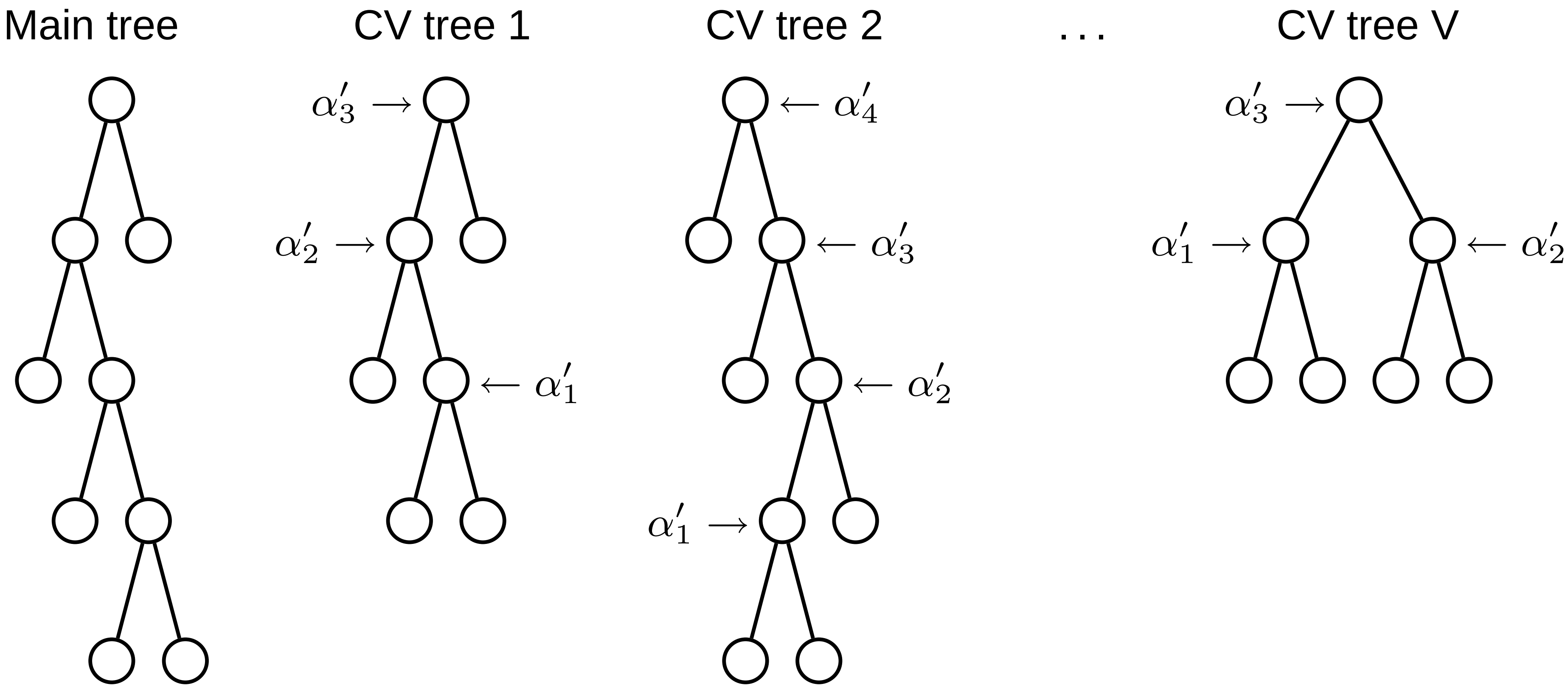
Subtree selection by V-fold cross-validation

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

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

6. Select the subtree with the smallest CV cost

V-fold cross-validation



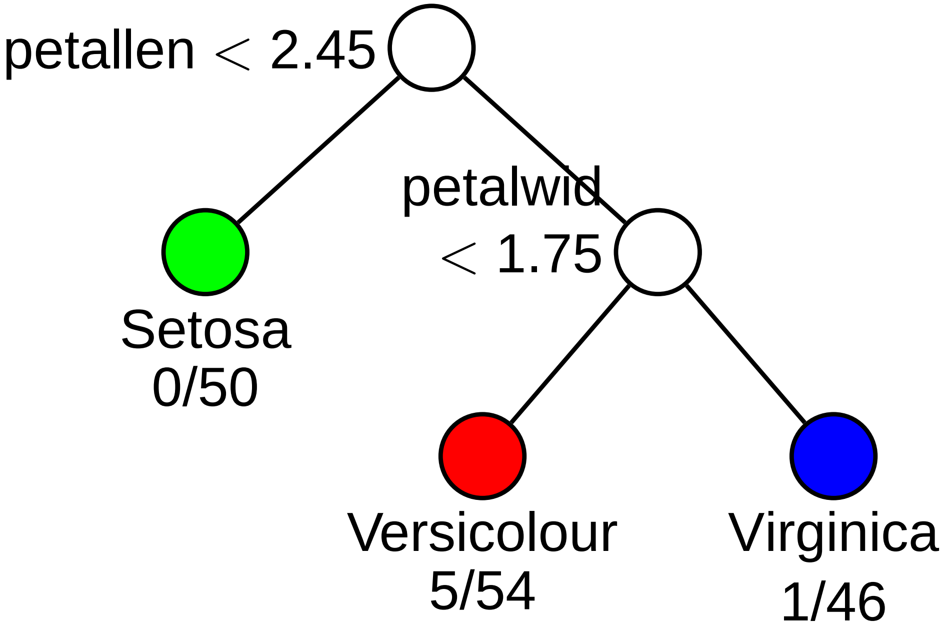
- Main tree is grown using all the data
- Each CV tree is grown using $(V-1)$ subsets

k-SE rule

1. Let $\hat{R}(T)$ be the estimated misclassification cost of T and let $\widehat{\mathrm{SE}}[\hat{R}(T)]$ be the standard deviation of the cross-validation estimates for T
2. Let subtree T^* minimize $\hat{R}(T_k)$
3. The k-SE tree T^{**} is the smallest subtree such that

$$\hat{R}(T^{**}) \leq \hat{R}(T^*) + k\widehat{\mathrm{SE}}[\hat{R}(T^*)]$$

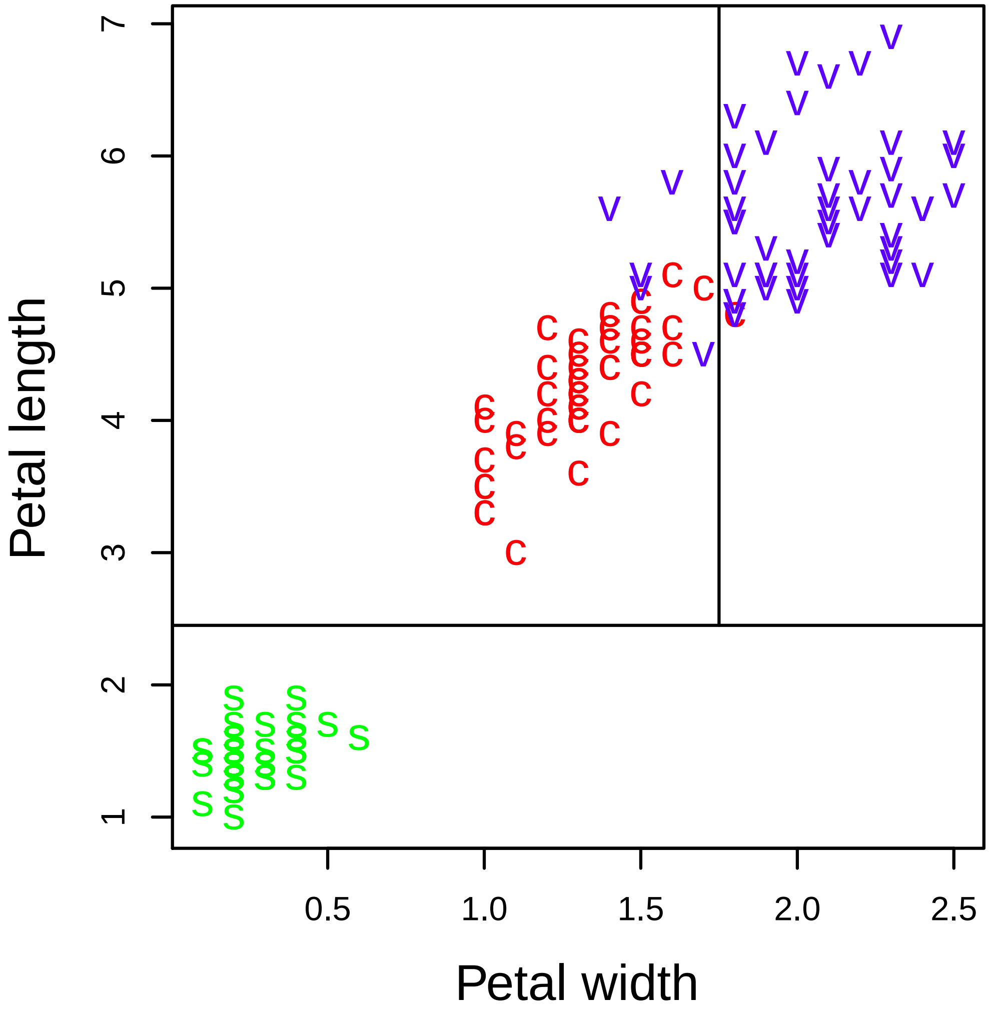
RPART (Therneau and Atkinson, 2012) tree for iris data



Number of errors divided by number cases given beneath each leaf node.

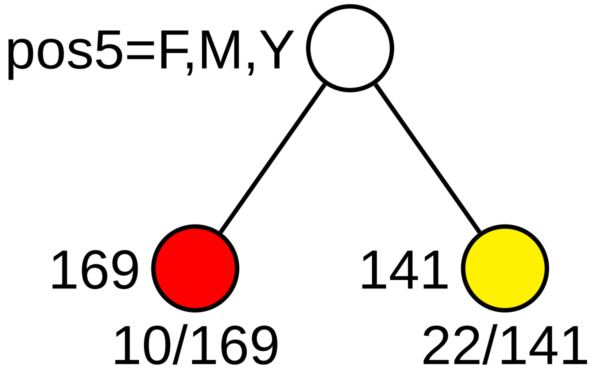
Same tree with 0 or 1-SE pruning.

RPART partitions for iris data



s = Setosa, c = Versicolour, v = Virginica

RPART tree for peptide data



Red denotes binder, yellow denotes non-binder

Numbers beneath nodes are misclassified/sample size

0-SE and 1-SE trees are the same

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
- The 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 $\pi'(j)$

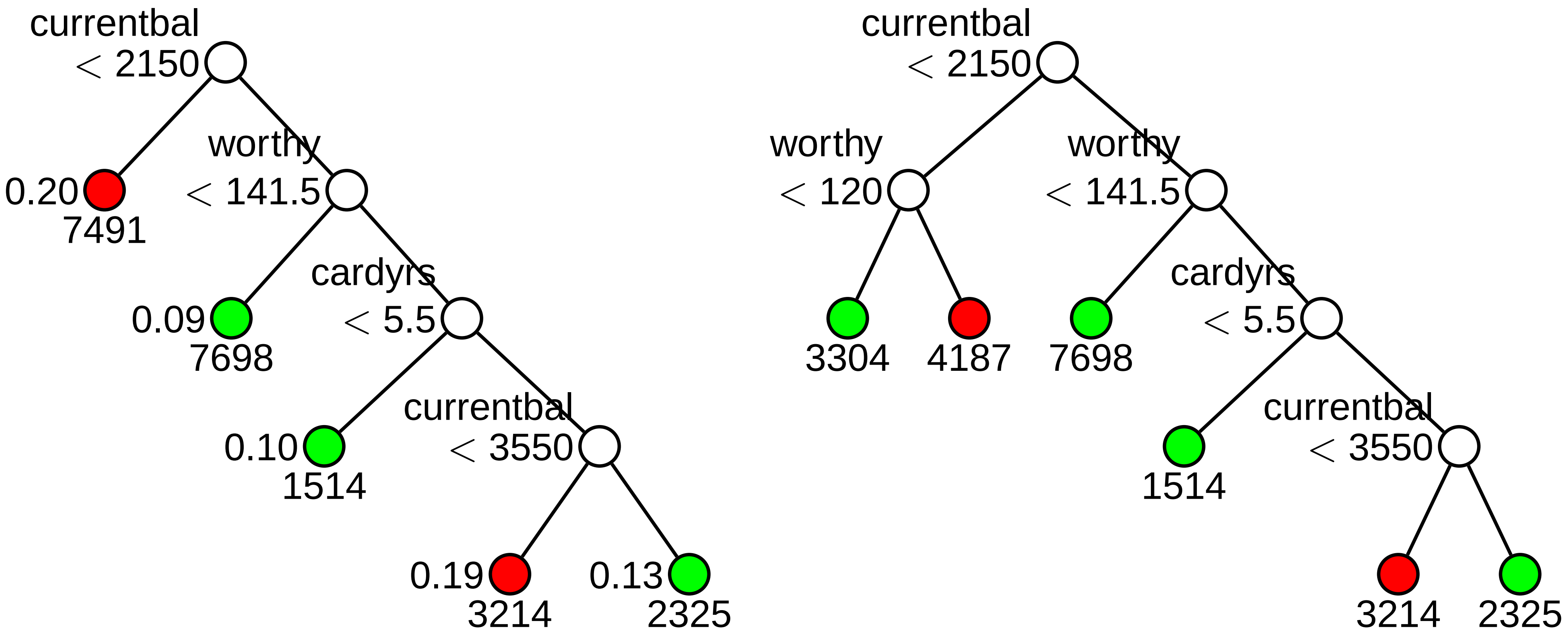
Altered priors (cont'd)

- 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 $\sum_i C(i|j)$
- GUIDE uses altered priors

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

CART surrogate splits for classification

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 $p(t_L \cap t'_L) = \sum_j \pi(j)N_j(LL)/N_j$

Note: If there are missing values, let A_k be the subset of the learning sample with non-missing values in x_k and the variable involved in s^*, and redefine N_j and $N_j(t)$ to be the numbers of class j cases in A_k and $A_k \cap t$, resp.

CART surrogate splits (cont'd.)

- 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 s^* as the split s_k^* that maximizes $\Delta i(s_k^*, t)$. Note: Since $\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R)$ and $i(t)$ is a function of $p(\cdot|t)$, $\Delta i(s_k^*, t)$ depends only on $p(\cdot|t)$. Thus s^* does not account for the number of missing values in x_k.

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 $\tilde{s}_k$

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: An M-valued ordered variable has $(M-1)$ splits; an M-category variable has $(2^{M-1}-1)$ 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.
- Computationally impractical: When there are three or more classes and categorical variables with many categories. Neither RPART nor commercial CART can handle categorical variables with more than 32 categories.
- Prediction accuracy: Often no better than linear discriminant analysis.

Key developments in the evolution of GUIDE

Classification. FACT (Loh and Vanichsetakul, 1988)

Proportional hazards. Loh (1991); Ahn and Loh (1994)

Linear regression. Chaudhuri et al. (1994)

Poisson & logistic regression. Chaudhuri et al. (1995)

Classification. QUEST (Loh and Shih, 1997)

Comparison of QUEST with other methods. Lim et al. (2000)

Classification. CRUISE (Kim and Loh, 2001, 2003)

Quantile regression. Chaudhuri and Loh (2002)

Generalized regression forests. GUIDE (Loh, 2002)

— least squares, least median of squares, quantile, Poisson, & proportional hazards

Logistic regression. LOTUS (Chan and Loh, 2004)

Classification trees and forests. GUIDE (Loh, 2009)

Multivariate and longitudinal responses. GUIDE (Loh and Zheng, 2013)

Proportional hazards and subgroup identification. GUIDE (Loh et al., 2012)

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

Procedure:

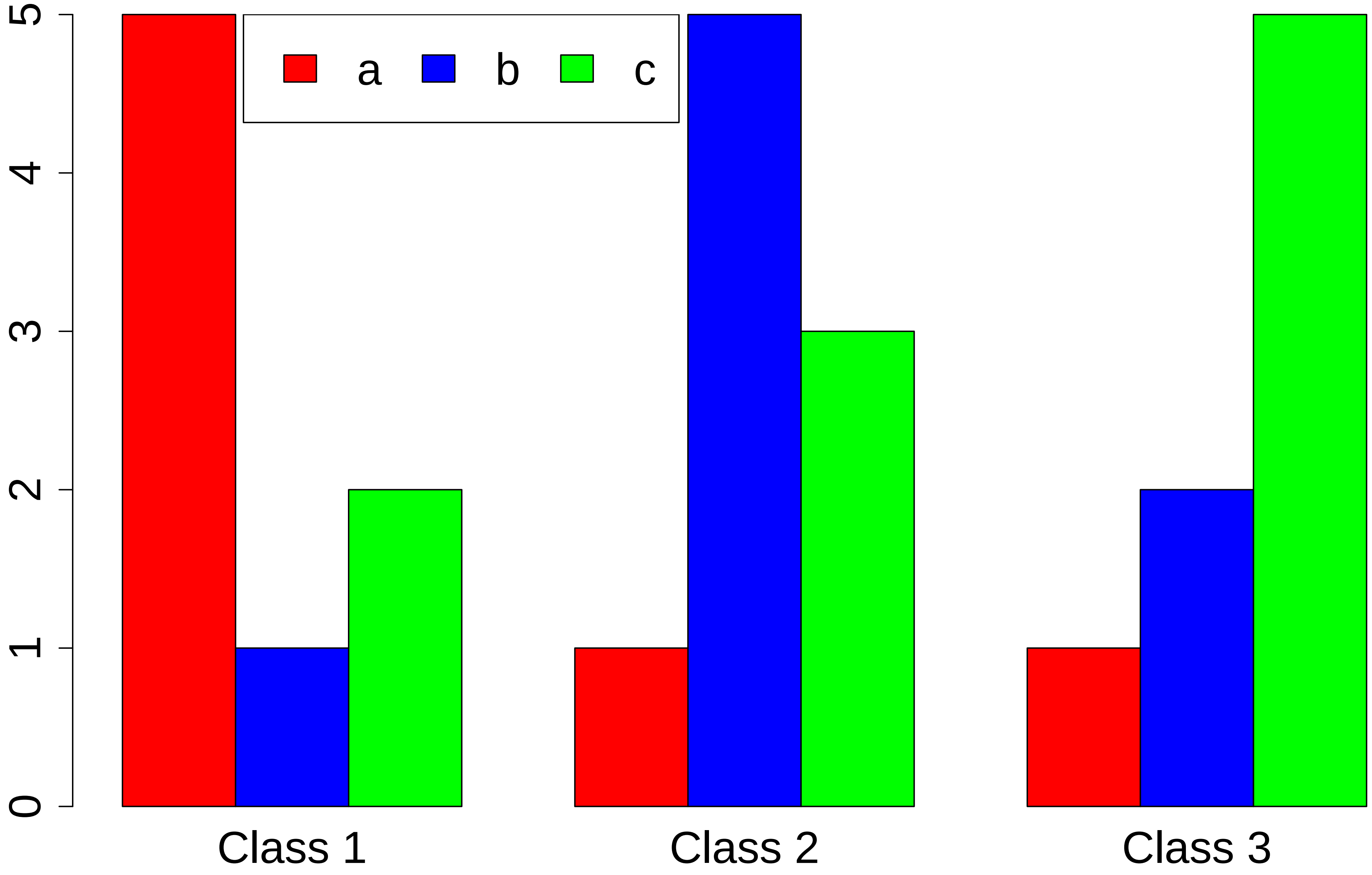
1. Replace missing values by means and modes at each node
2. Convert each categorical variable to a dummy vector and then transform to largest discriminant variable (crimcoord)
3. For linear combination splits, use recursive LDA
4. For univariate splits:
   (a) Use 1-way ANOVA to choose split variable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

Result: Fast and accurate classification tree with J splits at each node

FACT algorithm for univariate and linear combination splits on categorical variables

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. Substitute U for X in the univariate and linear combination split algorithms
5. A split of the form '$U \leq c$' can be expressed in the form '$X \in A$'

Example: A 3-category X variable with 3 Classes



Crimcoord transformation

X space

Y	X	rep
1	a	5
1	b	1
1	c	2
2	a	1
2	b	5
2	c	3
3	a	1
3	b	2
3	c	5

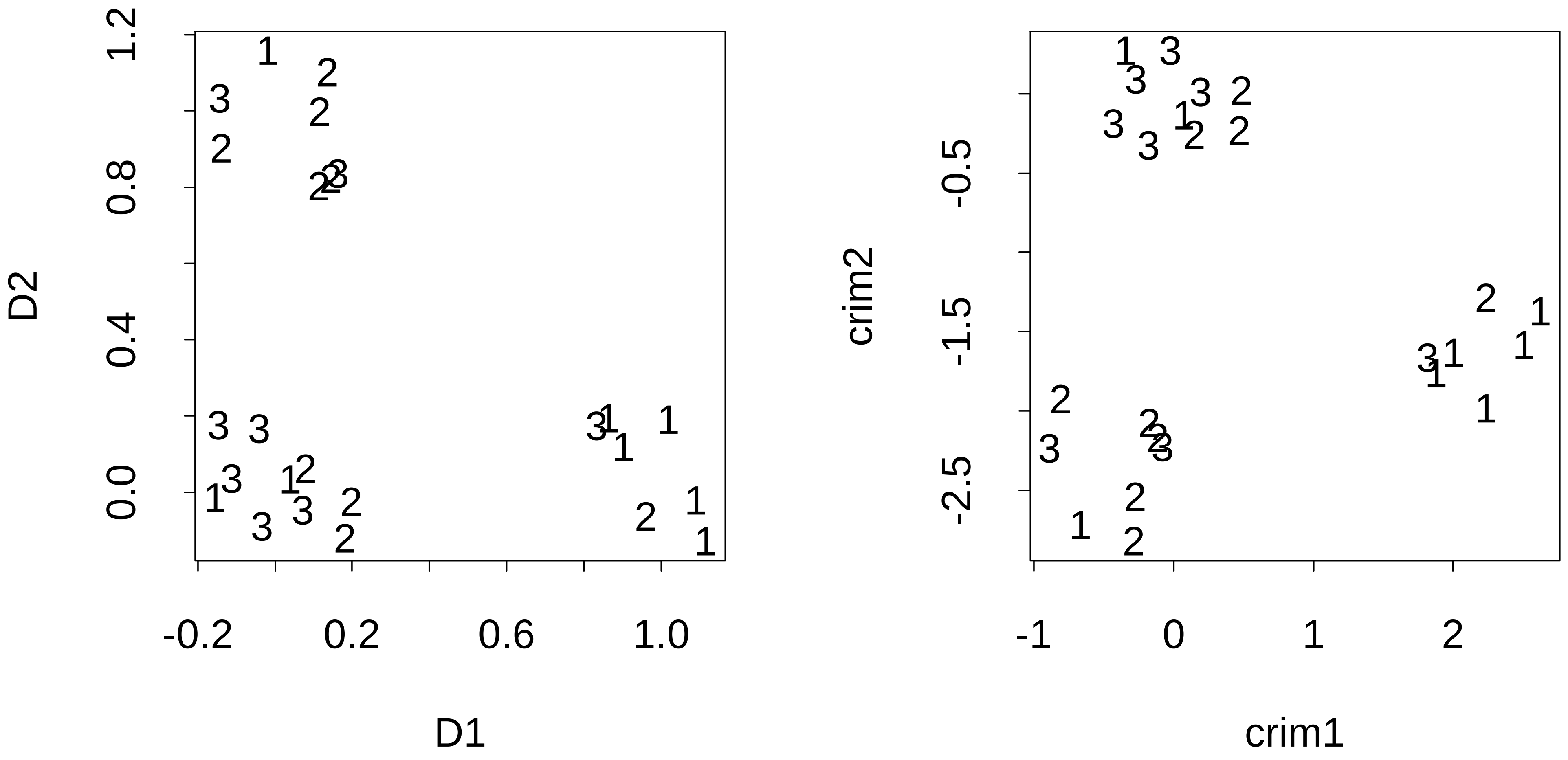
D space

Y	D_1	D_2
1	1	0
1	0	1
1	0	0
2	1	0
2	0	1
2	0	0
3	1	0
3	0	1
3	0	0

crimcoord space

Y	C_1	C_2
1	2.29	-1.51
1	-0.54	-2.38
1	0	0
2	2.29	-1.51
2	-0.54	-2.38
2	0	0
3	2.29	-1.51
3	-0.54	-2.38
3	0	0

Jittered plots of dummy and crimcoord variables



QUEST (Loh and Shih, 1997)
Classification trees with binary splits

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 (recursive LDA on ordinal and crimcoord variables)
4. Use mean/mode imputation for missing values at each node

CRUISE (Kim and Loh, 2001, 2003)

Classification trees with two or more splits/node

1. For split variable selection, first change each ordinal variable to a categorical variable by discretizing it at sample quartiles
2. Marginal test: Find p-value of chi^2 test of Y vs. each variable
3. Interaction test: Find p-value of χ^2 test of Y vs. each pair of variables
4. Select the variable(s) with smallest p-value
5. If smallest p-value is from an interaction test, select the one with smaller marginal p-value
6. If selected variable is categorical, transform it to a crimcoord
7. Use Box-Cox transform. and LDA to split on selected variable or crimcoord
8. For linear combination splits, use recursive LDA

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

P(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 $X_2, \ldots, X_5$ 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

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

GUIDE marginal tests for non-categorical 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 contingency table with class values as rows and intervals as columns.

5. Let ν be df of the table after deleting rows and columns with no observations. Compute the chi-squared statistic χ^2_ν for testing independence.

6. Use forward and backward Wilson-Hilferty (1931) approximation to convert χ^2_ν to a 1-df chi-squared

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

GUIDE split variable selection: marginal tests for categorical X

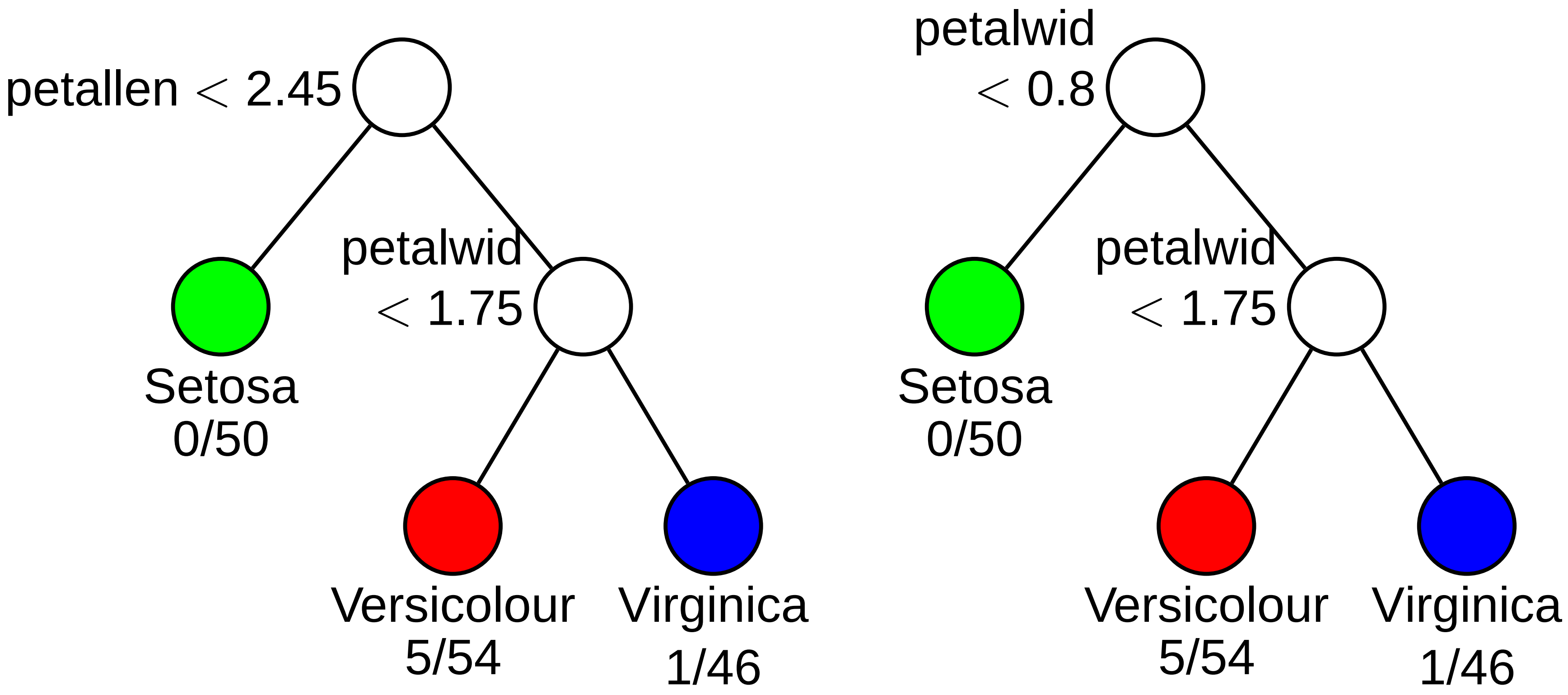
Use X values to form the columns of the contingency table

Chi-squared tests

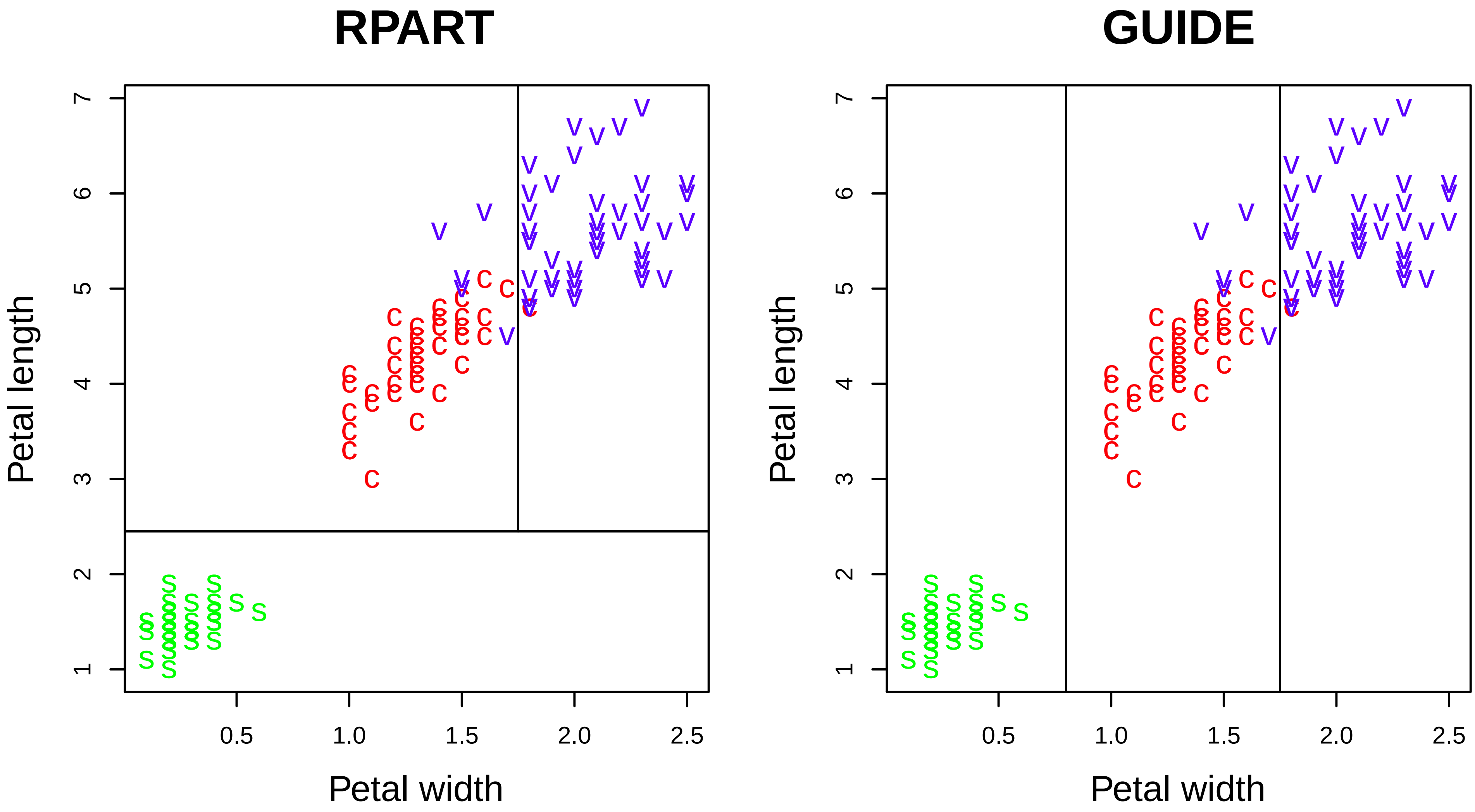
	Petal length (χ^2 = 223.9)				Petal width (χ^2 = 226.0)			
	$\leq$2.2	(2.2, 3.7]	(3.7, 5.2]	$>$5.2	$\leq$0.5	(0.5, 1.1]	(1.1, 1.8]	$>$1.8
Setosa	50	0	0	0	49	1	0	0
Versicol	0	7	43	0	0	10	40	0
Virginica	0	0	18	32	0	0	16	34

	Sepal length (χ^2 = 109.2)				Sepal width (χ^2 = 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 (left) and GUIDE (right) trees for iris data



Number of errors divided by number cases given beneath each leaf node.


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

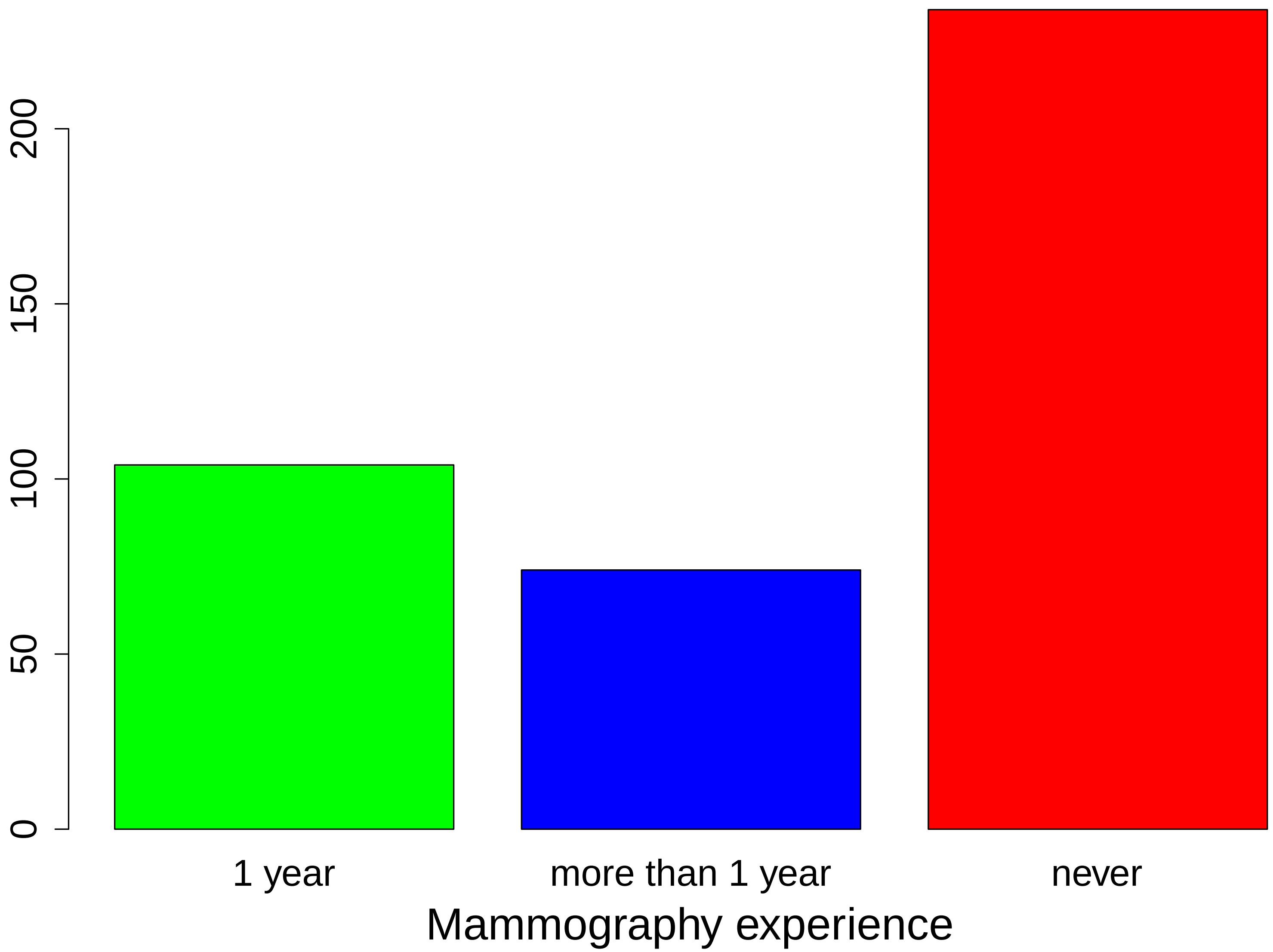
Example: Women’s knowledge, attitude, and behavior toward mammography (Hosmer & Lemeshow, 2nd ed.)

- 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 non-categorical

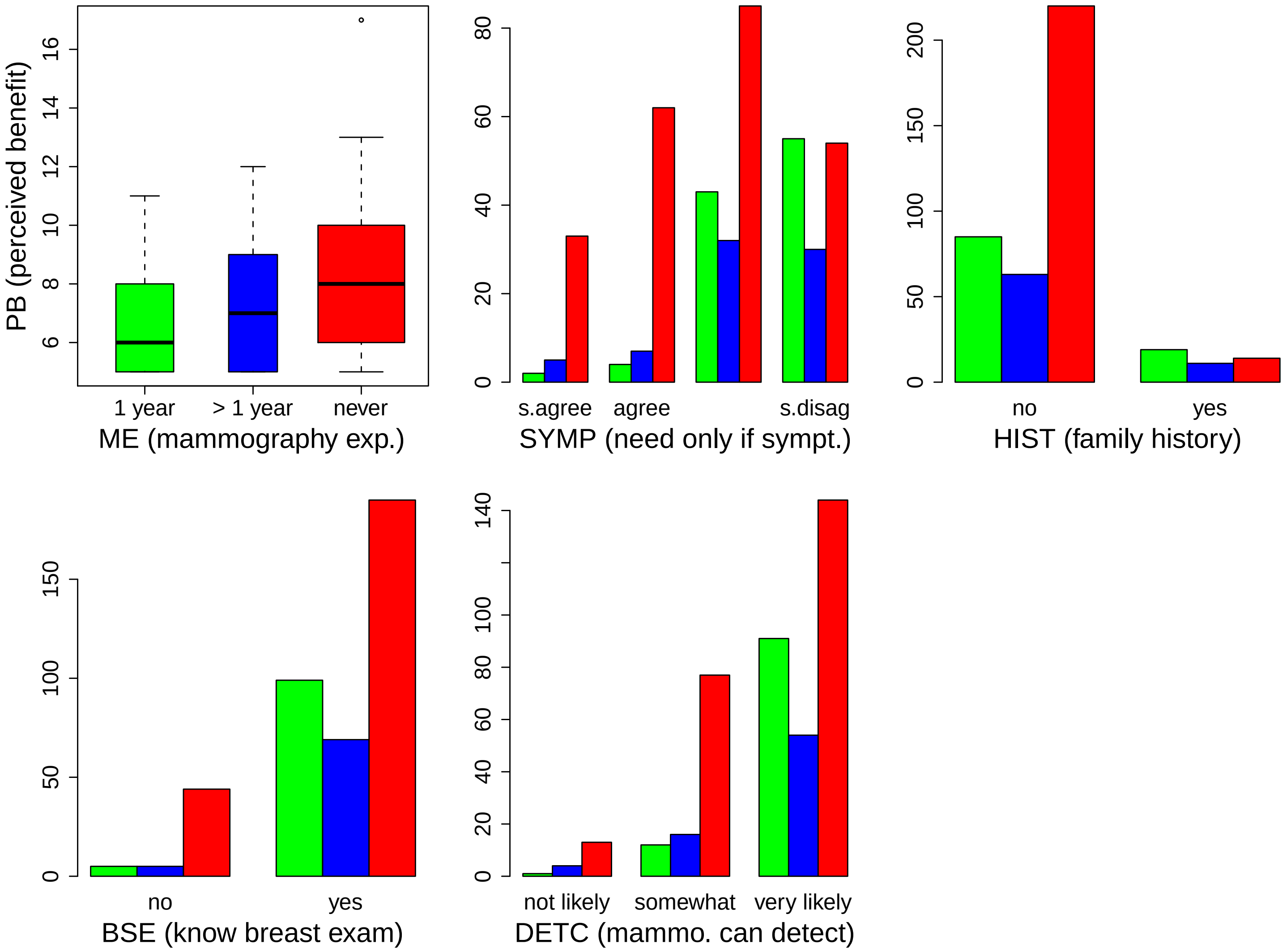
Mammography variables

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	Has 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)

Distribution of classes



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

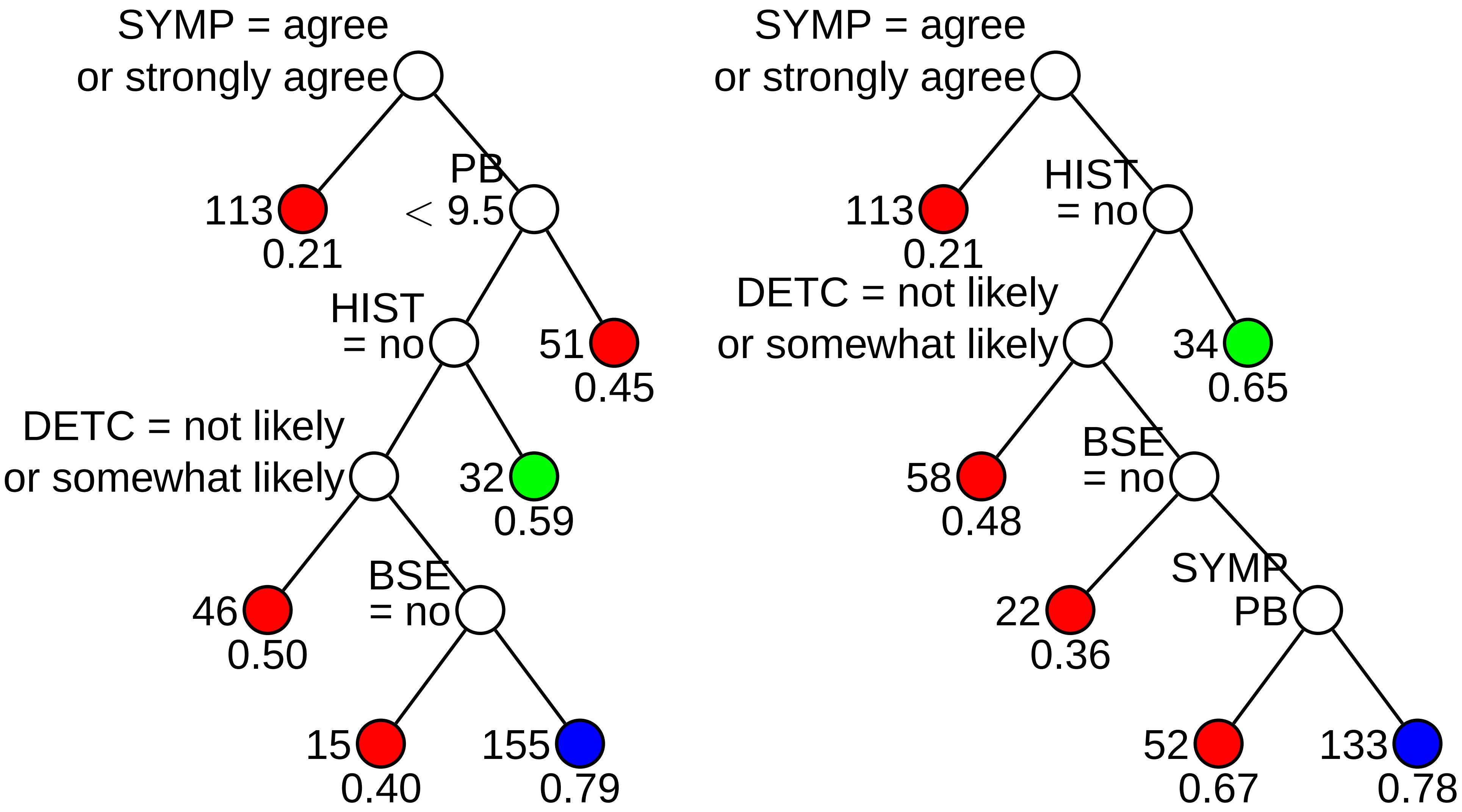
** DETCD = 1 if DETC = "very likely", DETCD = 0 otherwise

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

C4.5 does not allow unequal misclassification costs

RPART (left) and GUIDE (right) trees



- Mean misclassification cost below and sample size on left of each node
- Within 1 year in green, more than one year in blue, never in red

Chi-squared tests

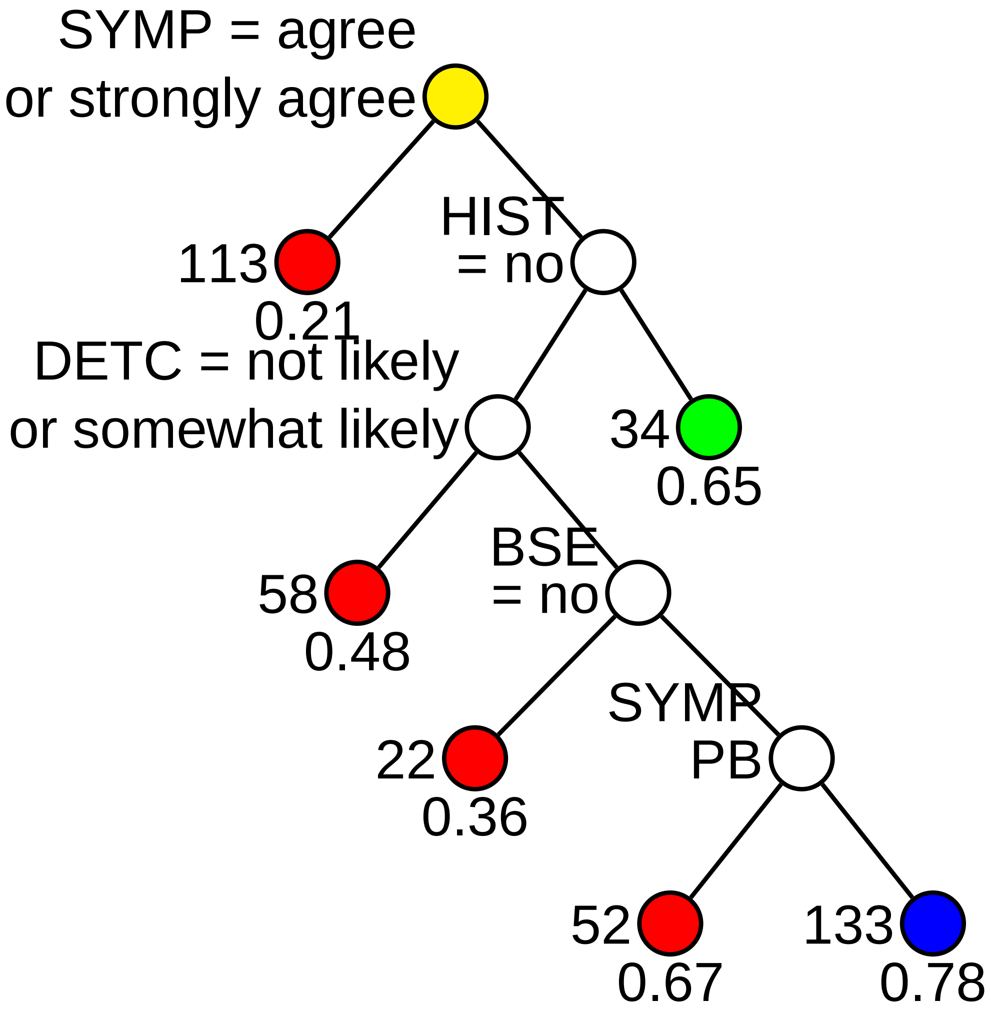
	SYMP (χ^2_6 = 57.2; $\chi^2_1 \approx$ 47)			
	strongly			strongly
ME	agree	agree	disagree	disagree
Never	33	62	85	54
1 year	2	4	43	55
$>$ 1 yr	5	7	32	30

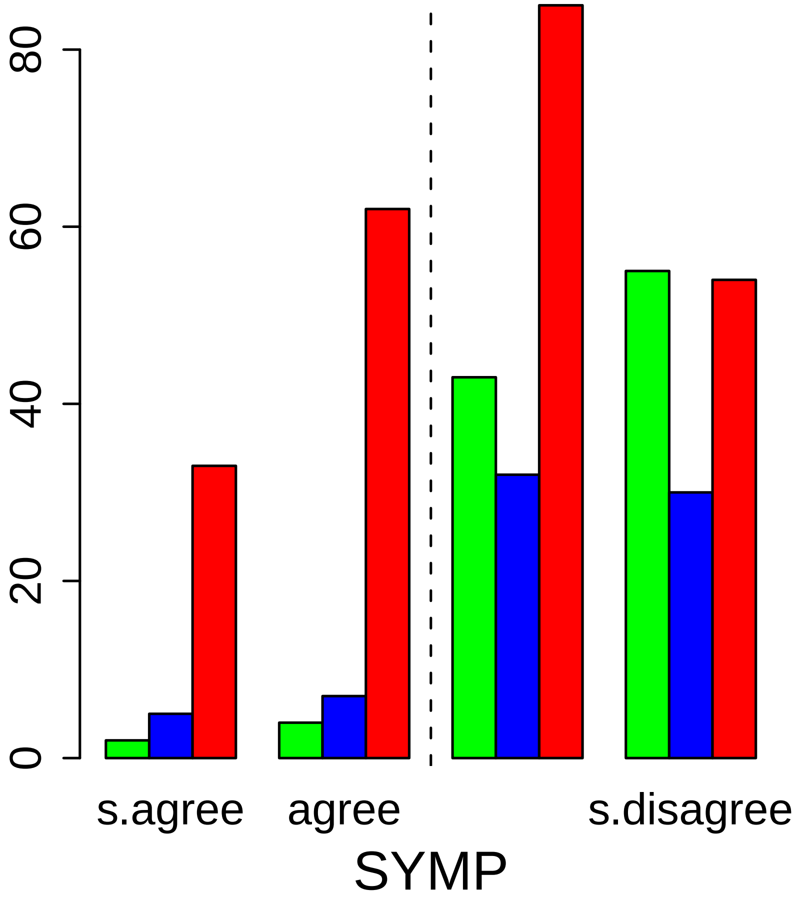
	PB (χ^2_6 = 31.3; $\chi^2_1 \approx$ 19)			
ME	$\leq$ 5.7	(5.7, 7.6]	(7.6, 9.4]	$>$ 9.4
Never	33	68	65	68
1 year	31	43	22	8
$>$ 1 yr	19	25	18	12

	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

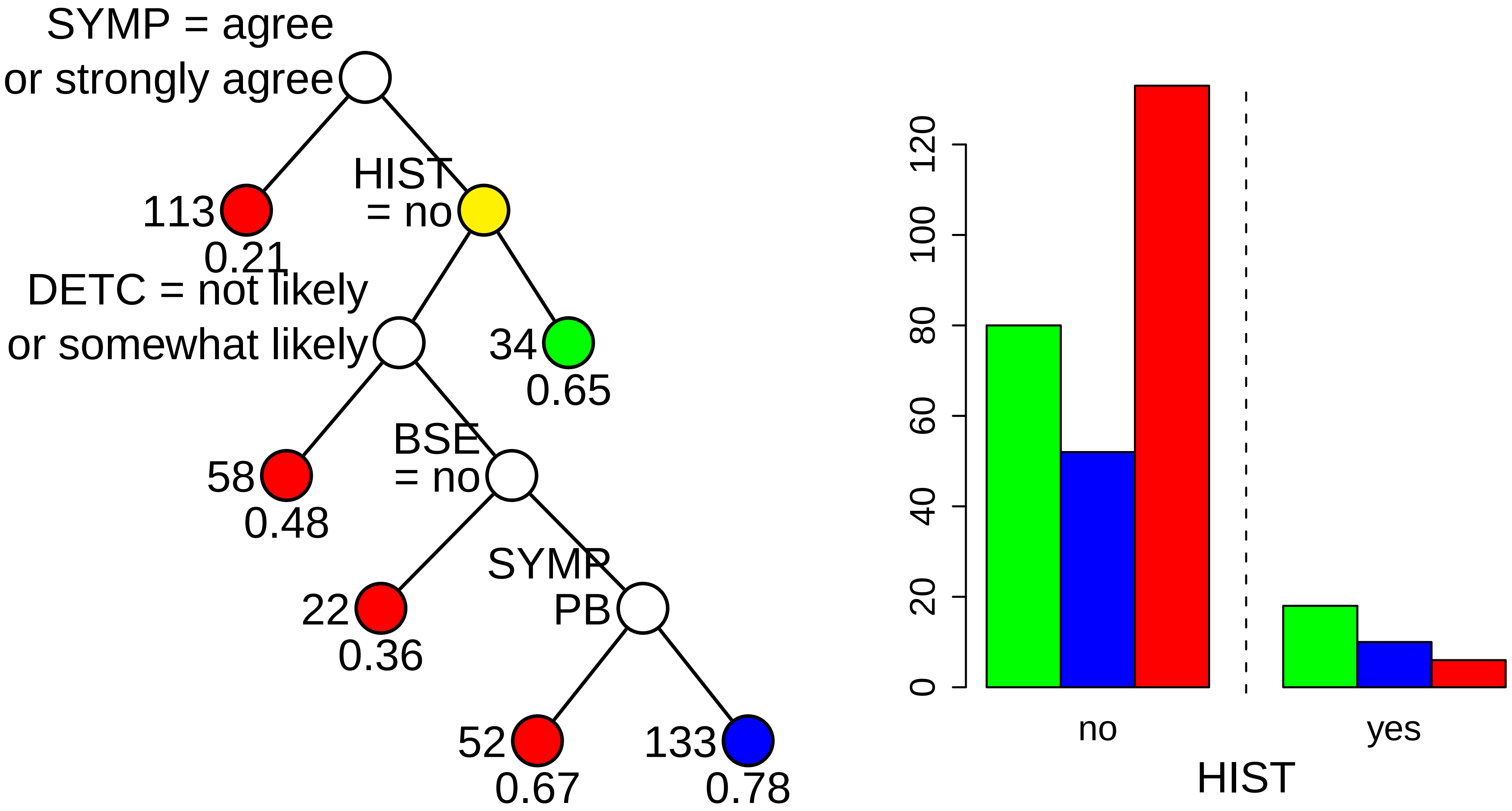
	BSE (χ^2_2 = 15.6, $\chi^2_1 \approx 13$)		HIST (χ^2_2 = 13.1, $\chi^2_1 \approx 10$)	
ME	no	yes	no	yes
Never	44	190	220	14
1 year	5	99	85	19
$>$ 1 yr	5	69	63	11

1st split

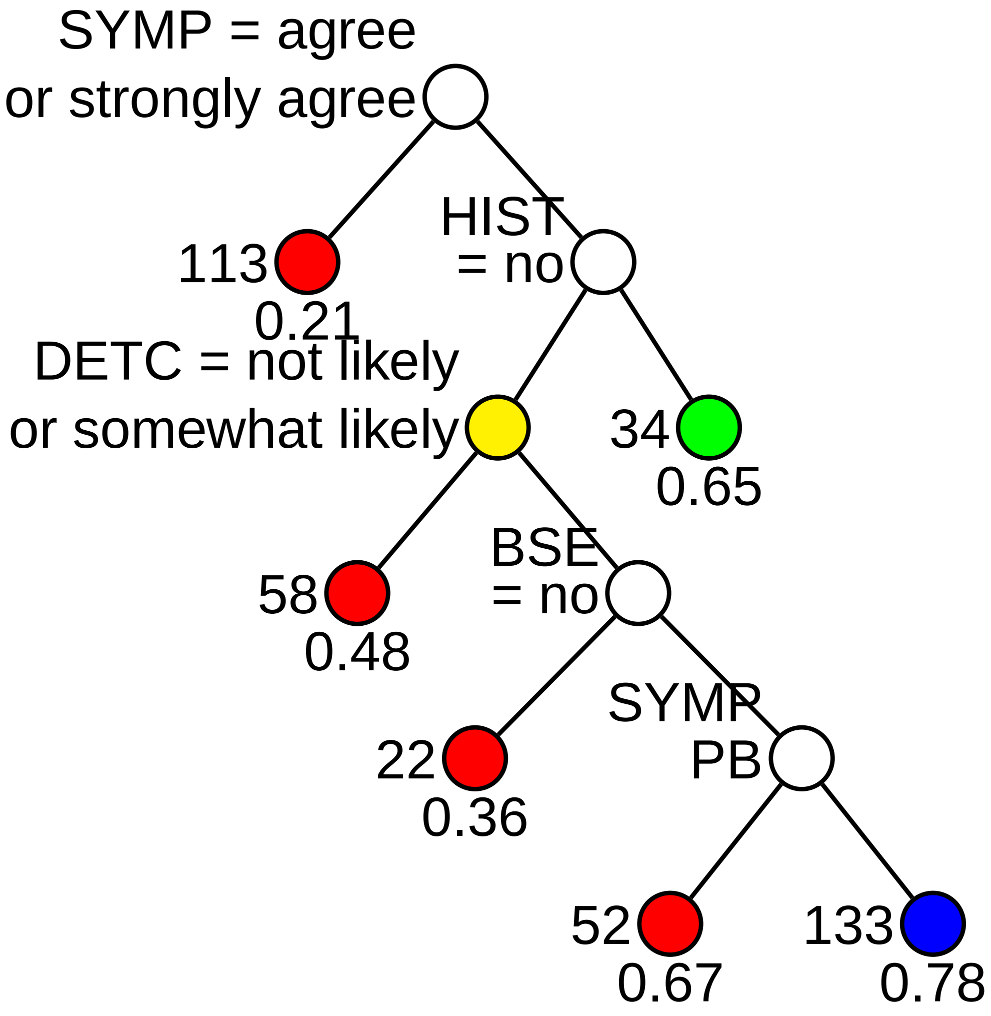

SYMP = agree
or strongly agree
113
0.21
HIST
= no
DETC = not likely
or somewhat likely
34
0.65
58
0.48
BSE
= no
22
0.36
SYMP
PB
52
0.67
133
0.78

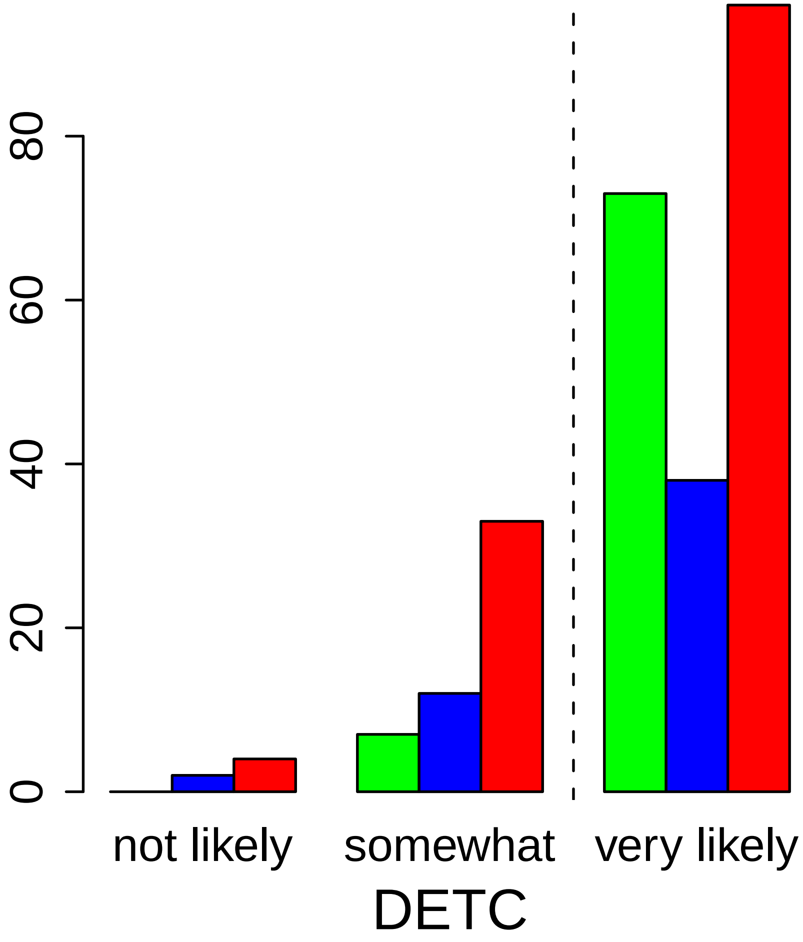

0
20
40
60
80
s.agree
agree
s.disagree
SYMP

2nd split

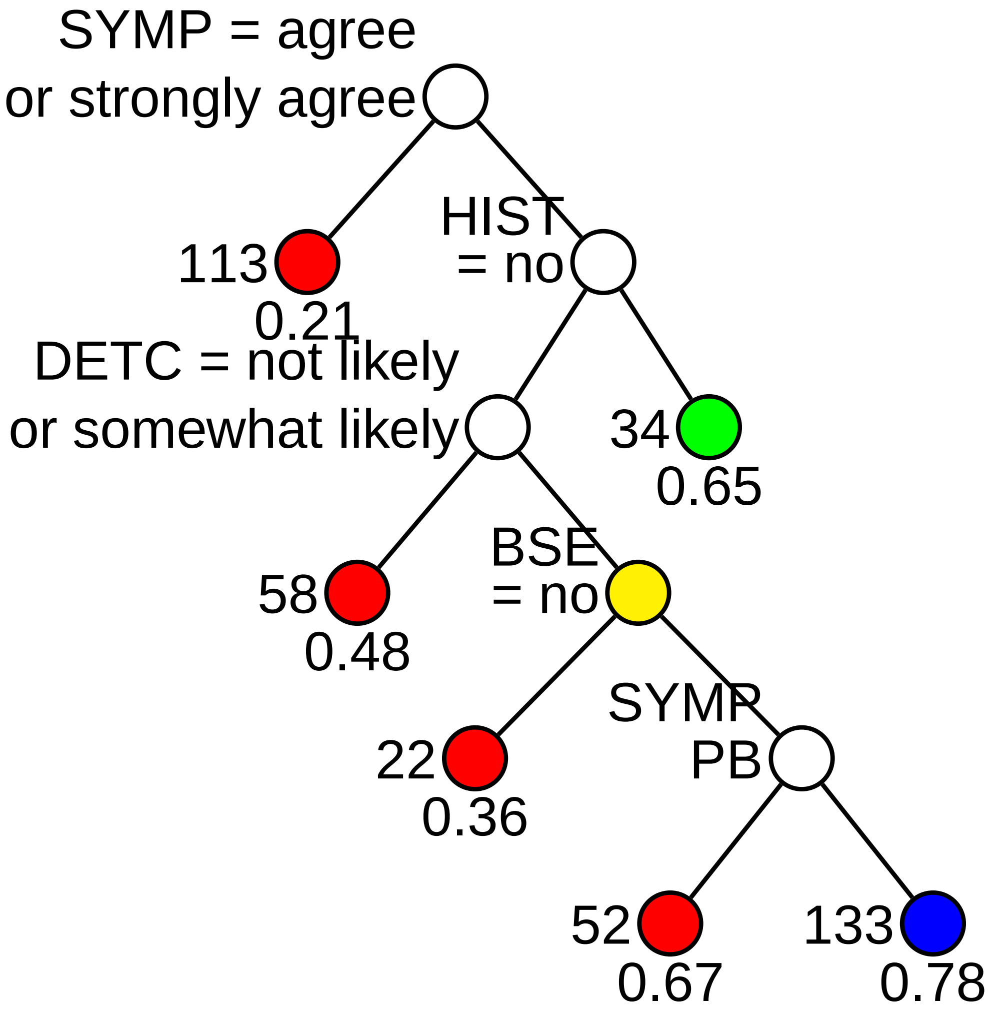

SYMP = agree
or strongly agree
113
0.21
HIST
= no
DETC = not likely
or somewhat likely
34
0.65
58
0.48
BSE
= no
22
0.36
SYMP
PB
52
0.67
133
0.78
0
20
40
60
80
100
120
no
yes
HIST

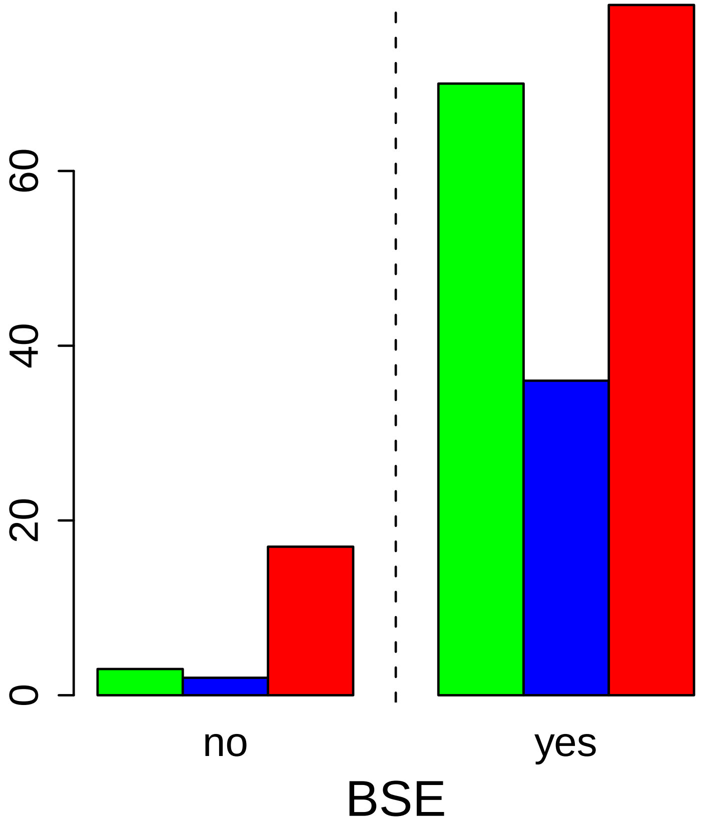
3rd split


SYMP = agree
or strongly agree
113
0.21
HIST
= no
DETC = not likely
or somewhat likely
34
0.65
58
0.48
BSE
= no
22
0.36
SYMP
PB
52
0.67
133
0.78

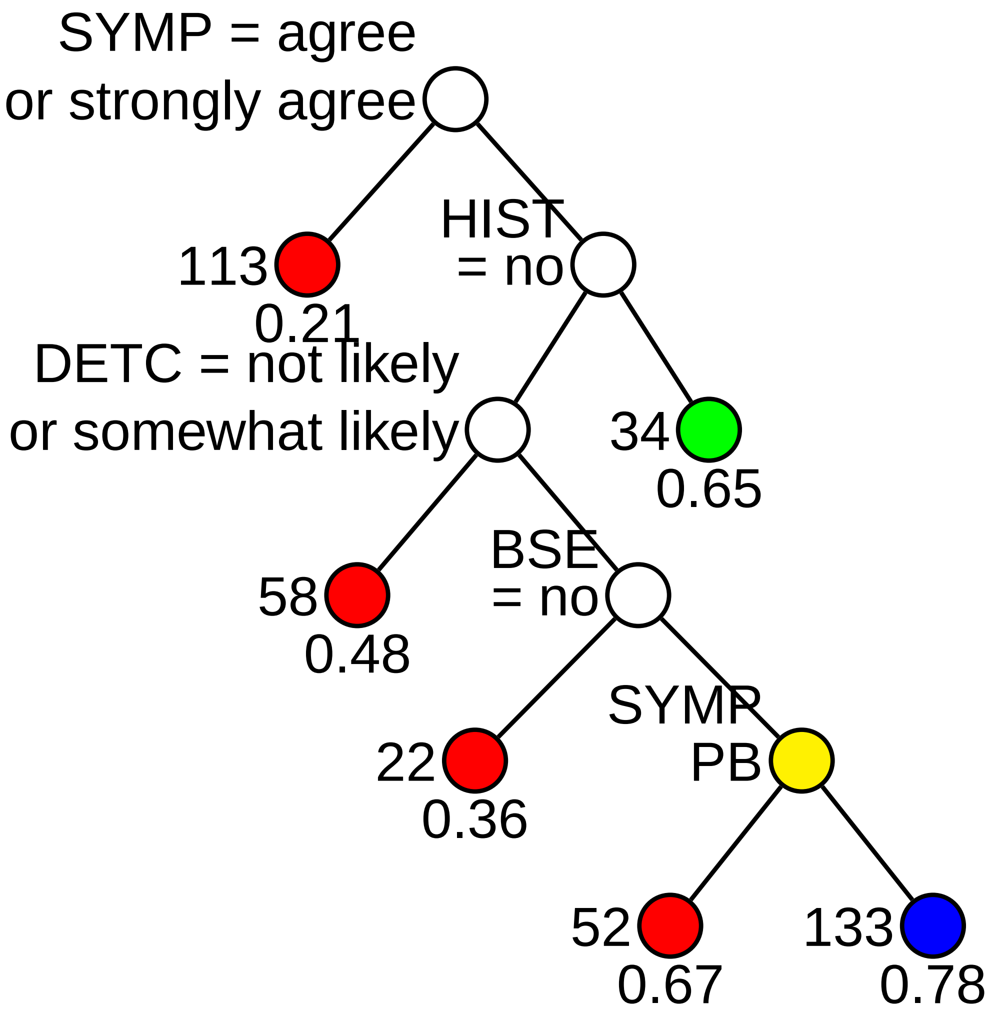

0
20
40
60
80
not likely
somewhat
very likely
DETC

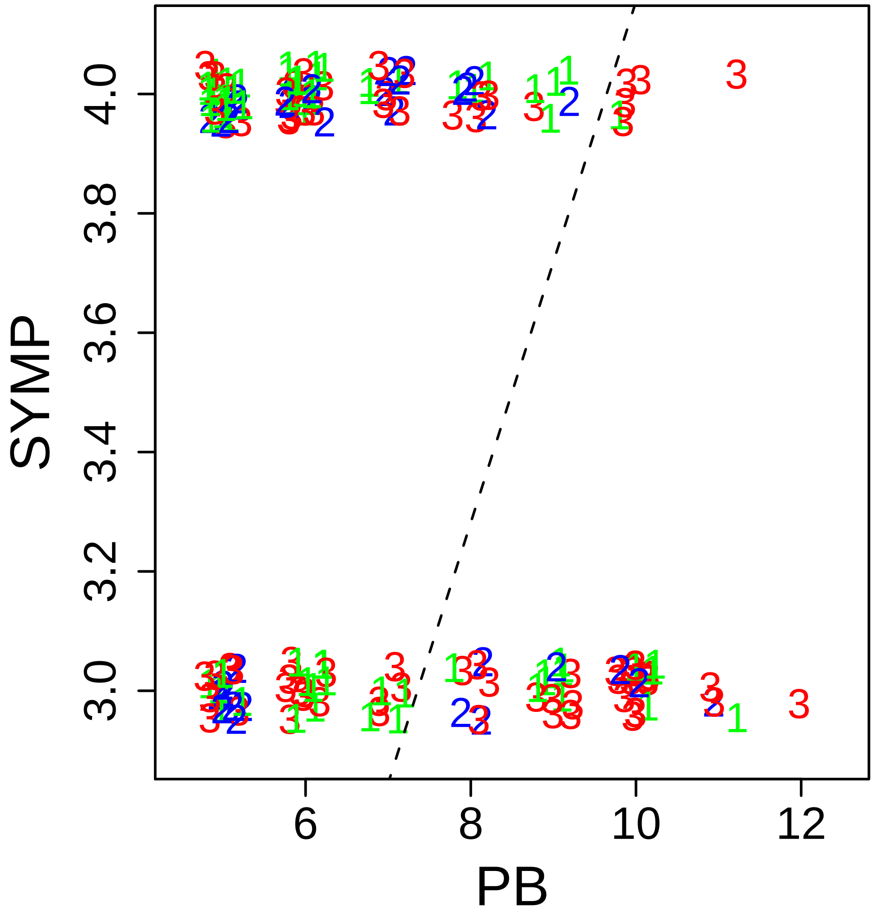
4th split


SYMP = agree
or strongly agree
113
0.21
HIST
= no
DETC = not likely
or somewhat likely
34
0.65
58
0.48
BSE
= no
22
0.36
SYMP
PB
52
0.67
133
0.78

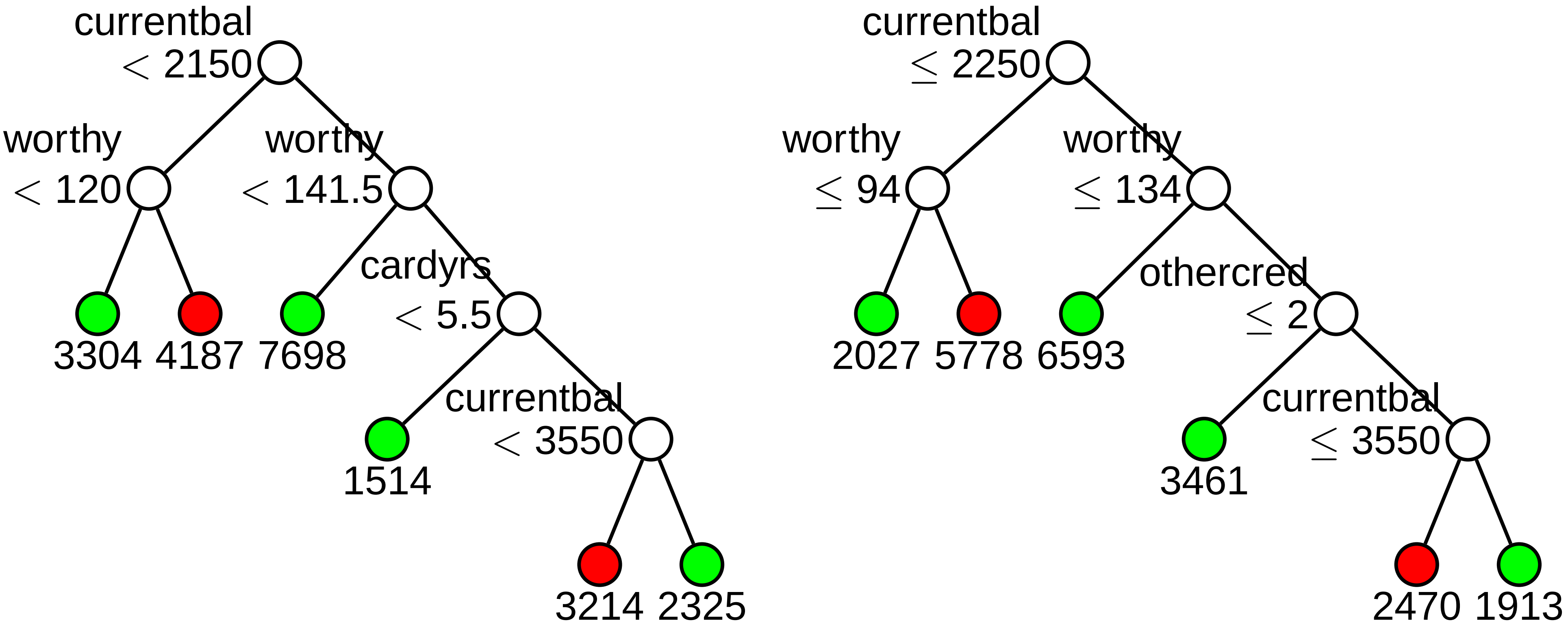

0
20
40
60
no
yes
BSE

5th split


SYMP = agree
or strongly agree
113
0.21
HIST
= no
DETC = not likely
or somewhat likely
34
0.65
58
0.48
BSE
= no
22
0.36
SYMP
PB
52
0.67
133
0.78


SYMP
3.0
3.2
3.4
3.6
3.8
4.0
PB
6
8
10
12

RPART (left) and GUIDE (right) unequal cost trees for credit data



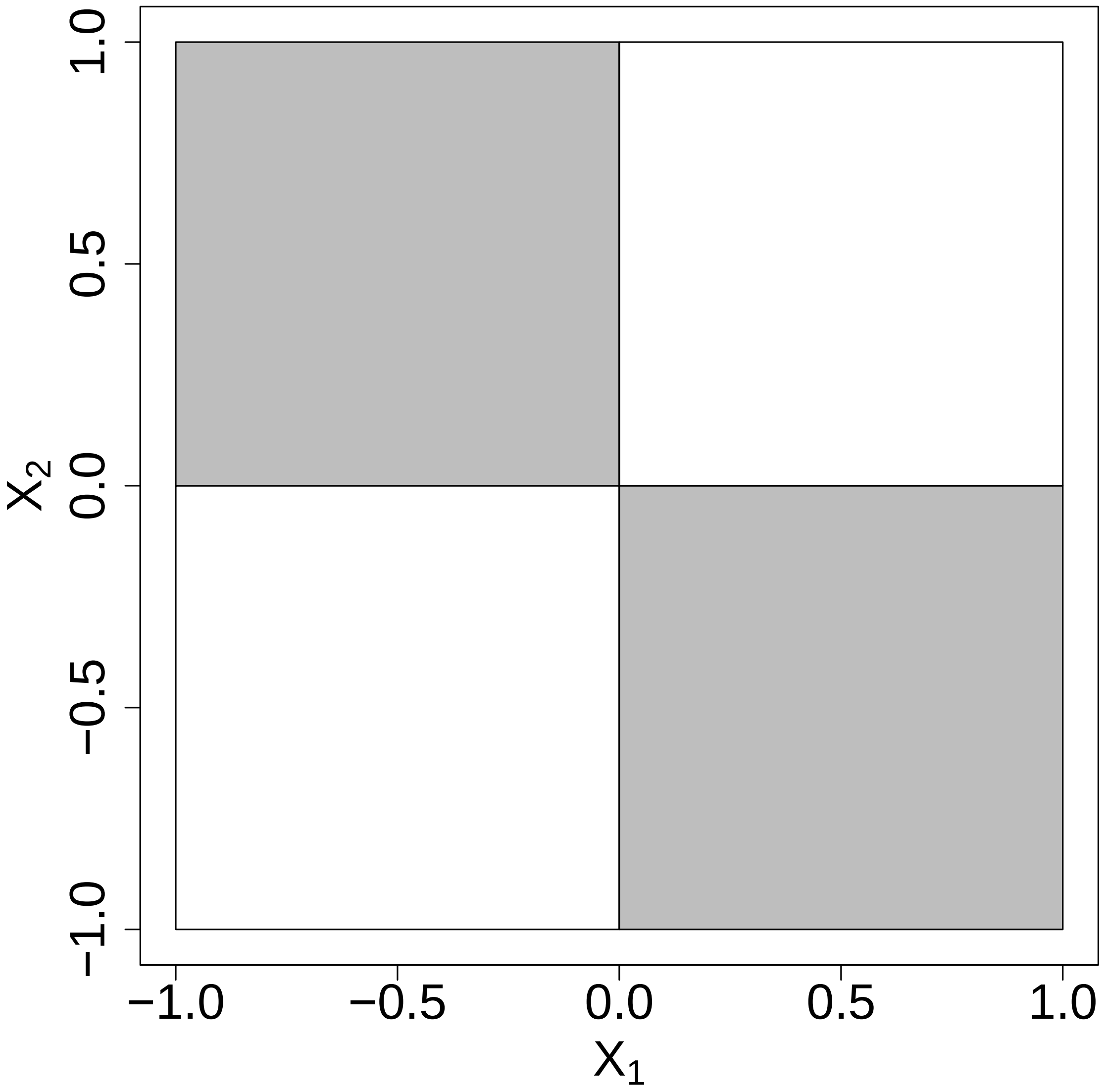
Dissatisfied and satisfied nodes in red and green; sample sizes beneath nodes

Conclusions from GUIDE trees

Dissatisfied customers tend to have:

1. Low current balances
2. High credit worthiness
3. Held credit card for many years
4. $2500 or more credit from other banks

Two-class problem with interaction



GUIDE split variable selection: interaction tests for X_1, X_2

1. If X_i is non-categorical, then:
   (a) If there are no missing values in X_i, split its range into two intervals (A_{i1}, A_{i2}) at $\bar{x}$ if $N(t) < 45J_t$, or three intervals (A_{i1}, A_{i2}, A_{i3}) at $\bar{x} \pm s\sqrt{3}/3$ if $N(t) \geq 45J_t$

   (b) Otherwise, if there are missing values in X_i, split its range into two intervals (A_{i1}, A_{i2}) at $\bar{x}$ and create a third "interval" for missing values
2. If X_i is categorical, let A_{ik} denote the singleton set containing its kth value
3. Divide the (X_1, X_2)-space into sets

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

4. Form a contingency table with class labels as rows and $\{B_{k,m}\}$ as columns
5. 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. Define
$$\alpha = \frac{0.05}{K}, \quad \beta = \frac{0.1}{K(K-1)}$$
and let $\chi^2_{\nu,\alpha}$ be the upper-α quantile of the chi-squared distribution with ν df.
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 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 the 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:

   (a) Convert X into a vector of dummy variables $(u_1, u_2, \ldots)$, where $u_i = 1$ if $X = a_i$ takes the ith value, and $u_i = 0$ otherwise

   (b) Obtain the covariance matrix of the u-vectors and find the eigenvectors associated with the positive eigenvalues. Project the u-vectors onto the space spanned by these eigenvectors (principal components).

   (c) Apply linear discriminant analysis to the transformed u-vectors to find the largest discriminant coordinate $v = \sum_i c_i u_i$

   (d) Let $v_{(1)} < v_{(2)} < \ldots$ denote the sorted v-values. There are at most n.

   (e) Find the split $t_L = \{v \leq v_{(m)}\}$ that minimizes the impurity

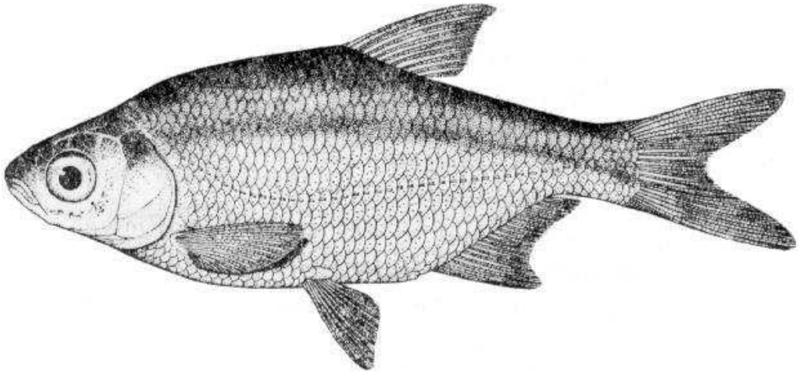
   (f) Re-express the split as $t_L = \{X \in S\}$

Fish classification

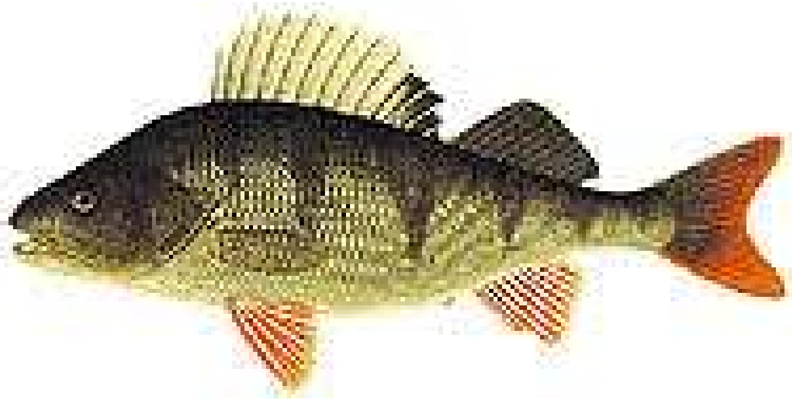
- 159 fish caught from the same lake near Tampere, Finland
- The fish are from 7 species: (1) 35 Bream, (2) 11 Parkki, (3) 56 Perch, (4) 17 Pike, (5) 20 Roach, (6) 14 Smelt, (7) 6 Whitefish

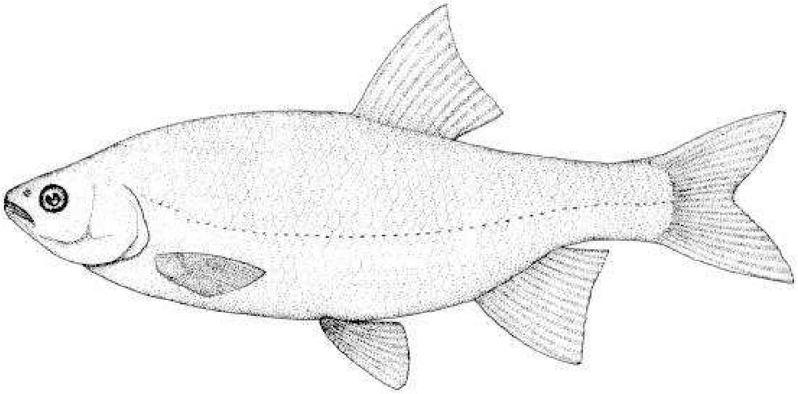
Predictor	Definition
Weight	Weight of the fish (in grams); one missing value
Length1	Length from the nose to the beginning of the tail (in cm)
Length2	Length from the nose to the notch of the tail (in cm)
Length3	Length from the nose to the end of the tail (in cm)
Height	Maximal height as % of Length3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

Bream (left) and Parkki (right)

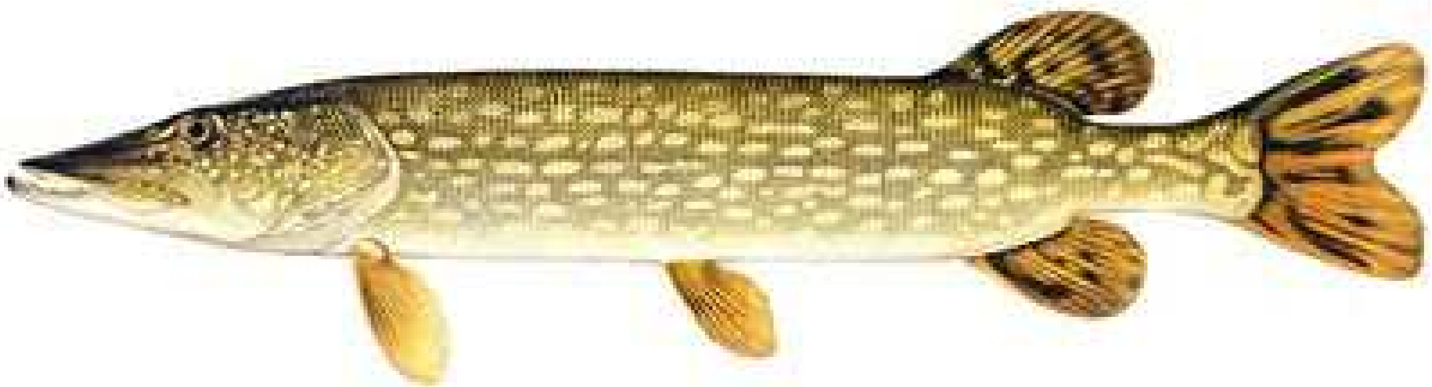


Perch (left) and Whitefish (right)

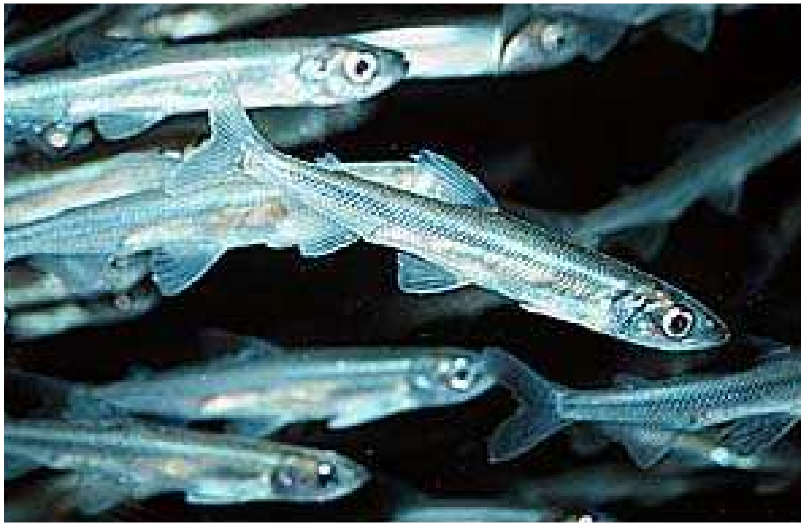




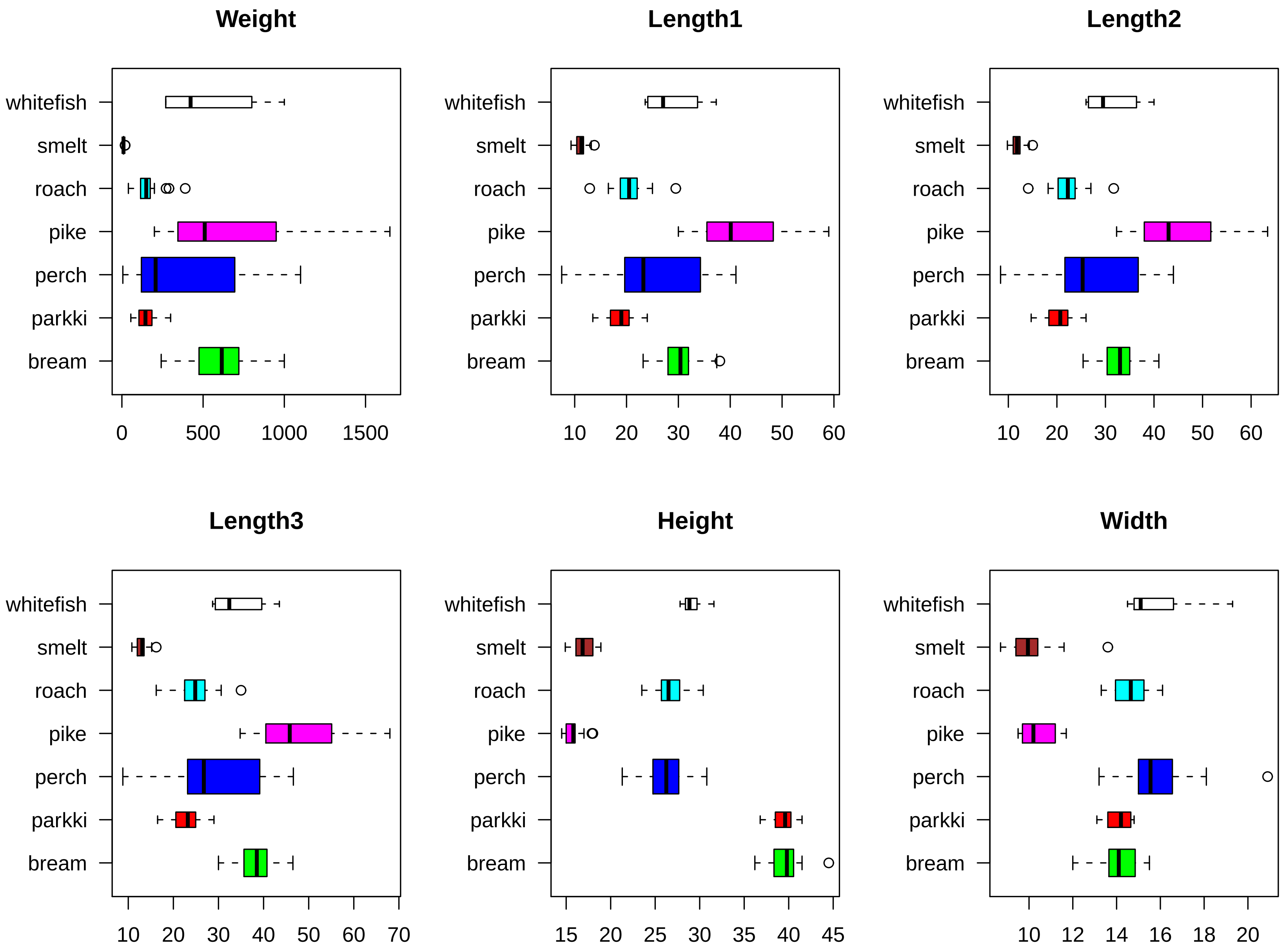
Pike



Roach (left) and Smelt (right)



Boxplots of continuous variables



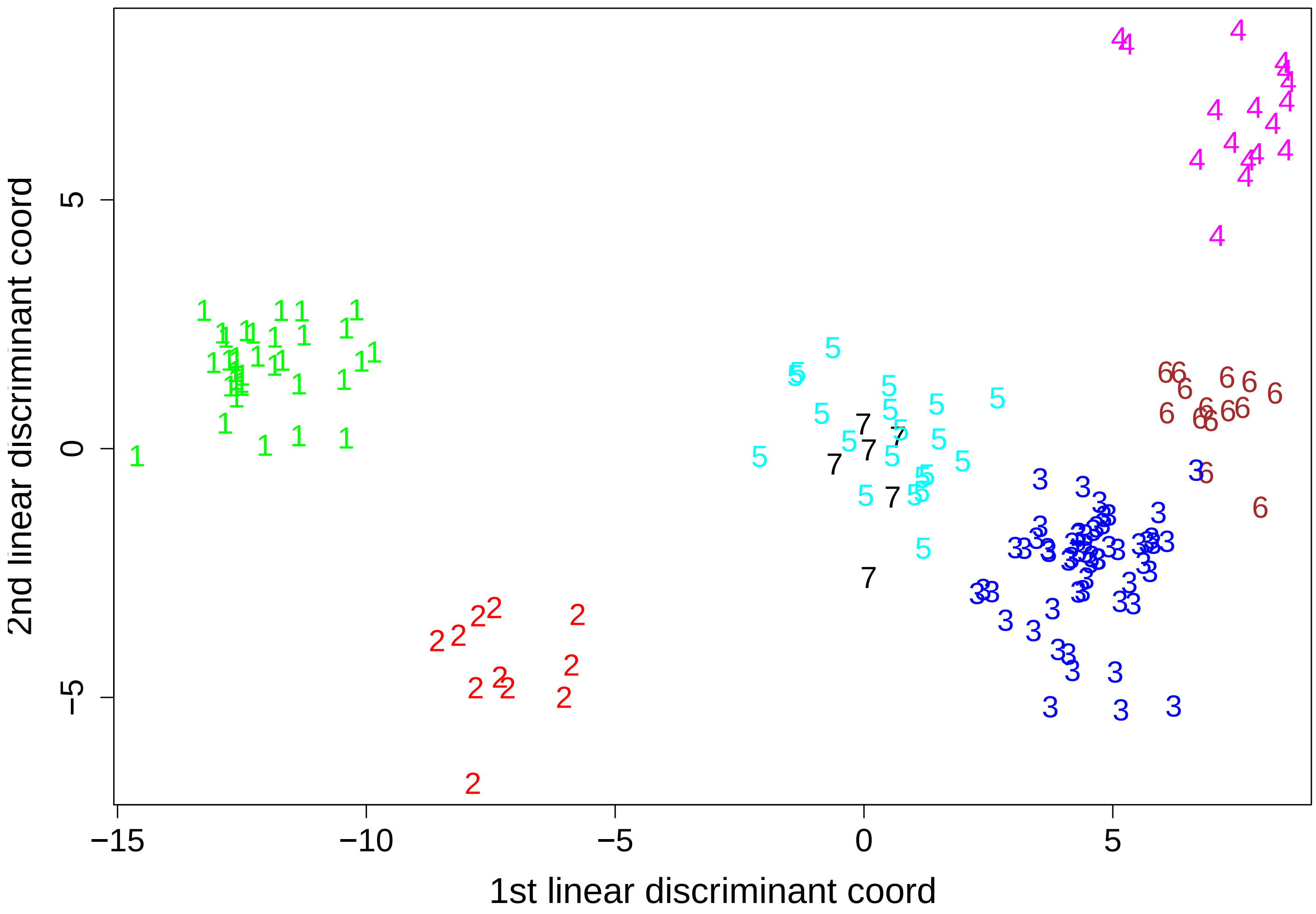
Sex by species

	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

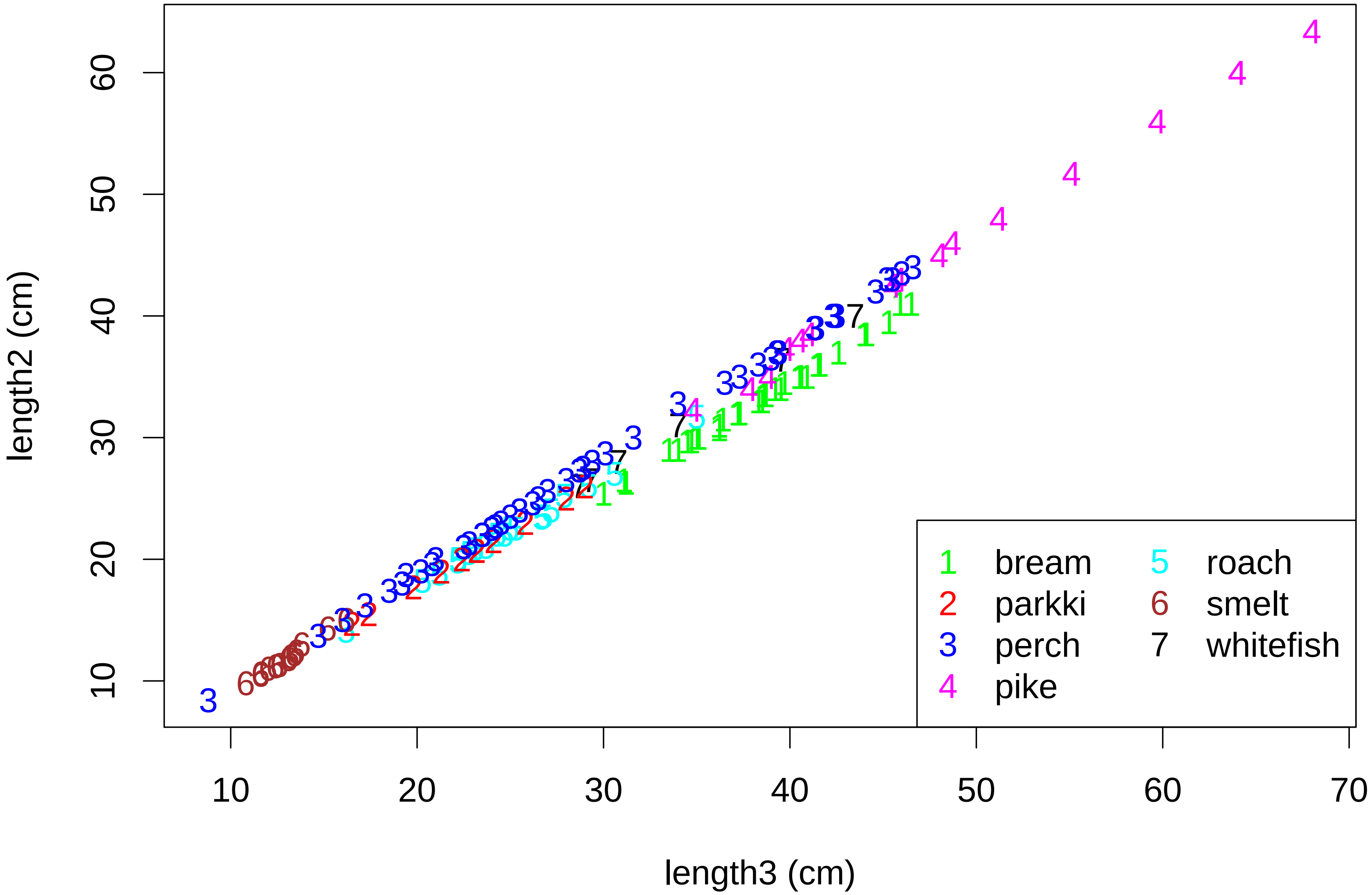
With Sex:	0 errors out of 71 complete cases
Without Sex:	1 error out of 158 complete cases

Plot of 1st two discriminant coords without Sex



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

Plot of Length2 vs. Length3



Linear discriminant split

Let (X_1, X_2) be a pair of non-categorical predictor variables

1. For the jth class and each X_i, compute the class mean $\bar{x}_{i,j}$ and class standard deviation $s_{i,j}$ of the samples in node t
2. Find the trimmed set S_j of class j samples in t such that $|X_i - \bar{x}_{i,j}| \leq 2s_{i,j}$ for $i = 1, 2$
3. Find the larger linear discriminant coordinate Z from the observations in $S_1 \cup \ldots \cup S_J$
4. Project the data in t onto the Z-axis to get their Z-values
5. Compute the Wilson-Hilferty 1-df chi-squared $W_L(X_1, X_2)$ from the Z's

Split variable selection with linear splits

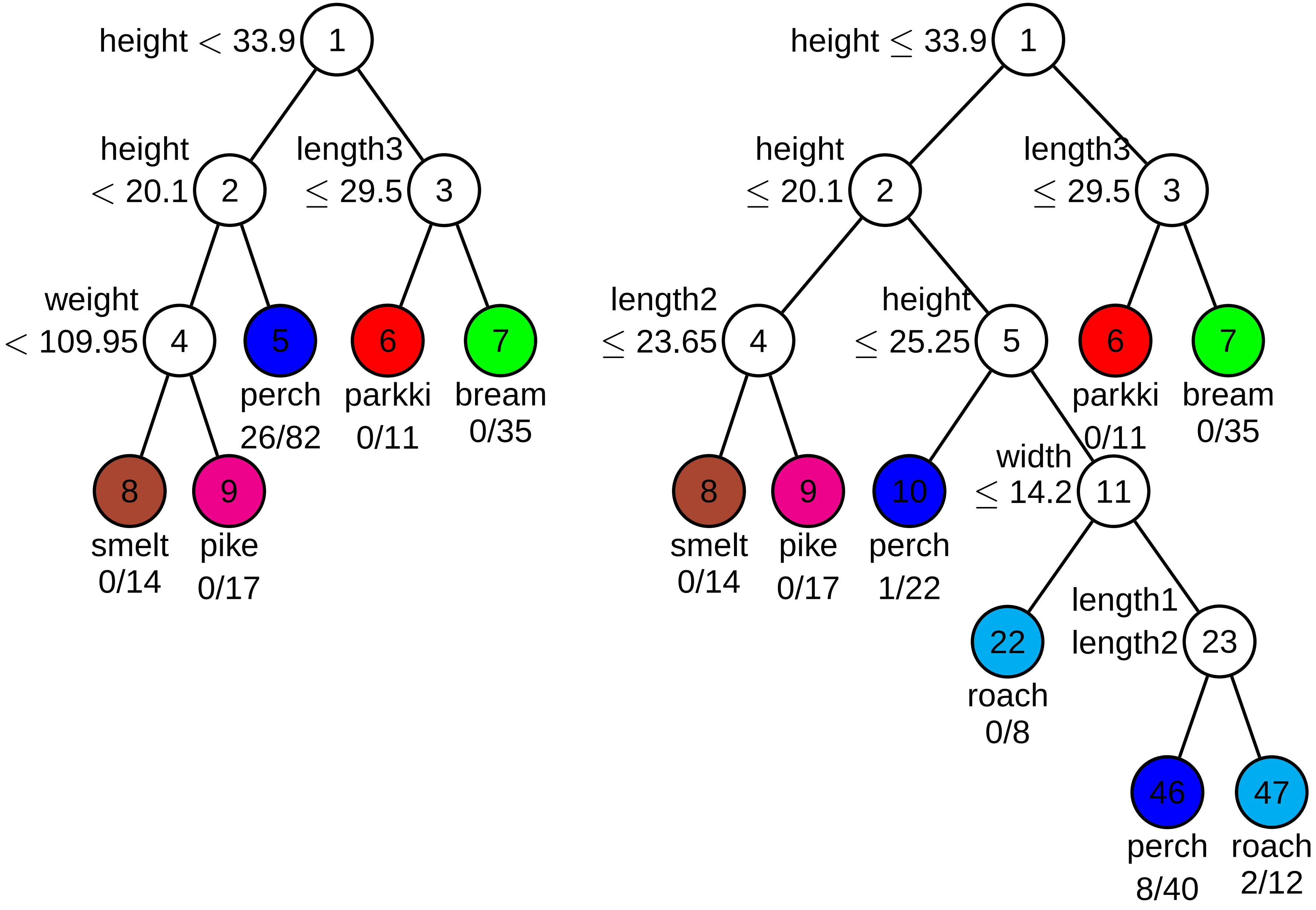
Let K be the number of non-constant predictor variables and let K_1 ($< K$) be the number that are non-categorical. Define

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

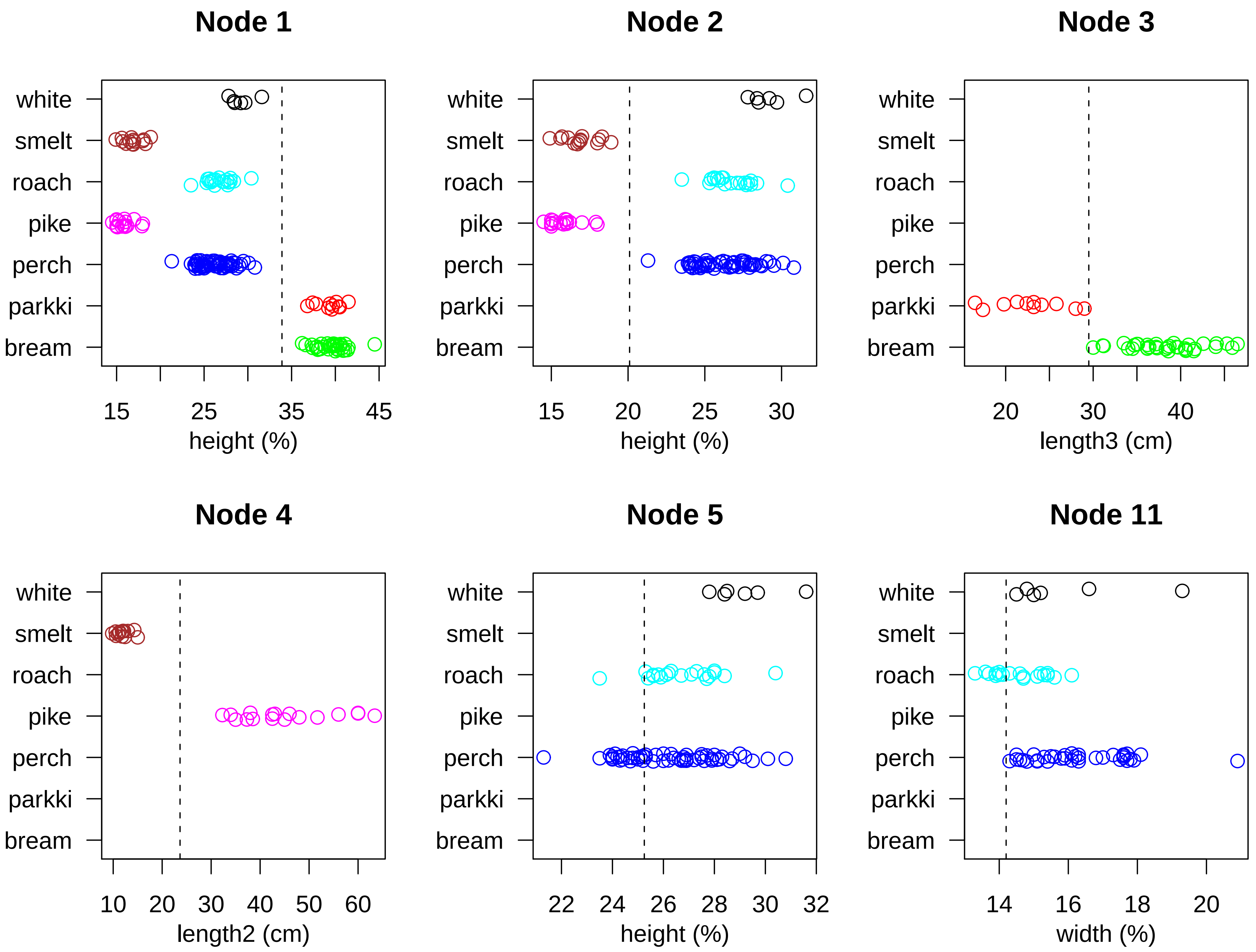
Let $\chi^2_{\nu,\alpha}$ denote the upper-α quantile of the chi-squared distribution with ν df

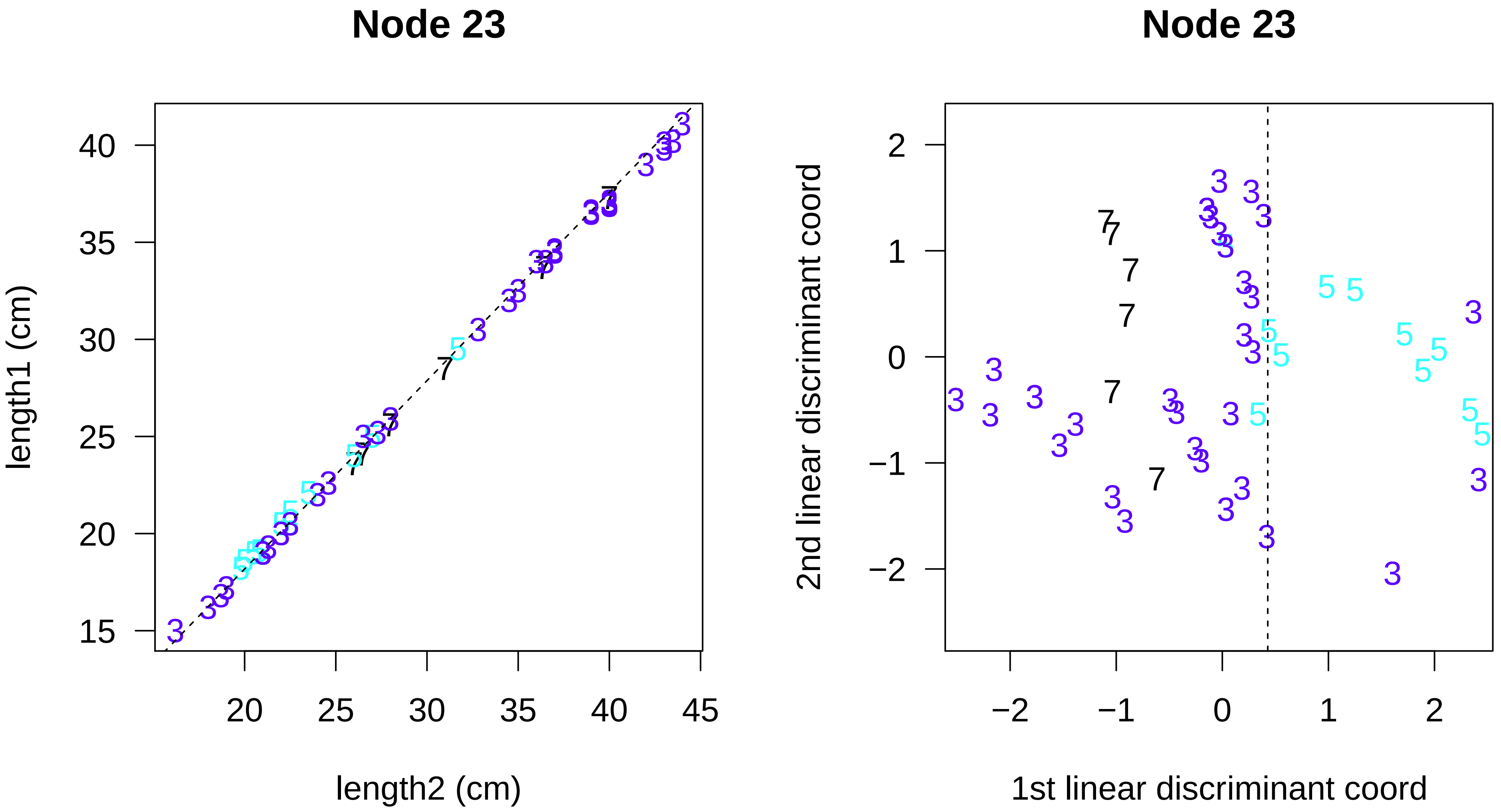
1. Compute $W_M(X_i)$ for each X_i
2. If $\max_i W_M(X_i) > \chi^2_{1,\alpha}$, split with the X_i having the largest $W_M(X_i)$
3. If $\max_i W_M(X_i) \leq \chi^2_{1,\alpha}$, find $W_I(X_i, X_j)$ for each pair of predictors
   (a) If $\max_{i \neq j} W_I(X_i, X_j) > \chi^2_{1,\beta}$, select the pair with the largest value of $W_I(X_i, X_j)$ and split on one of them
   (b) Else, compute $W_L(X_i, X_j)$ for each pair of non-categorical variables
      i. If $\max_{i \neq j} W_L(X_i, X_j) > \chi^2_{1,\gamma}$, select the pair with the largest $W_L(X_i, X_j)$ and split with their larger discriminant coordinate
      ii. Else, split with the X_i having the largest $W_M(X_i)$

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



RPART and GUIDE misclassify 26 and 11, respectively


Node 1
white
smelt
roach
pike
perch
parkki
bream
15
25
35
45
height (%)
Node 2
white
smelt
roach
pike
perch
parkki
bream
15
20
25
30
height (%)
Node 3
white
smelt
roach
pike
perch
parkki
bream
20
30
40
length3 (cm)
Node 4
white
smelt
roach
pike
perch
parkki
bream
10
20
30
40
50
60
length2 (cm)
Node 5
white
smelt
roach
pike
perch
parkki
bream
22
24
26
28
30
32
height (%)
Node 11
white
smelt
roach
pike
perch
parkki
bream
14
16
18
20
width (%)


Node 23
length1 (cm)
15
20
25
30
35
40
length2 (cm)
20
25
30
35
40
45
Node 23
2nd linear discriminant coord
−2
−1
0
1
2
1st linear discriminant coord
−2
−1
0
1
2

Importance ranking of variables

Importance score of X_i is

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

- $W_M(t, i)$ 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

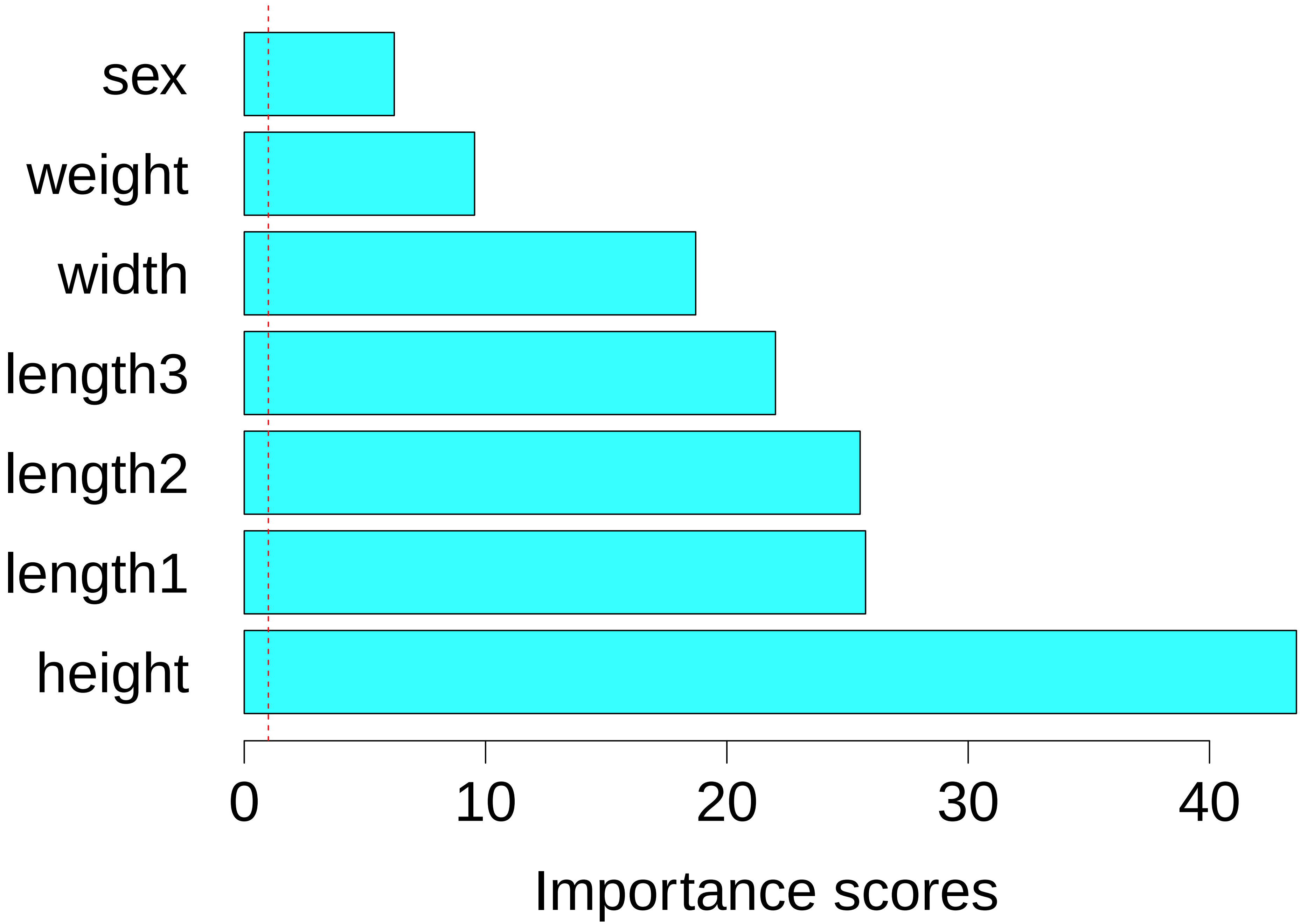
If X_i is constant at t, set $W_M(t, i) = 1$

Null distribution of importance scores

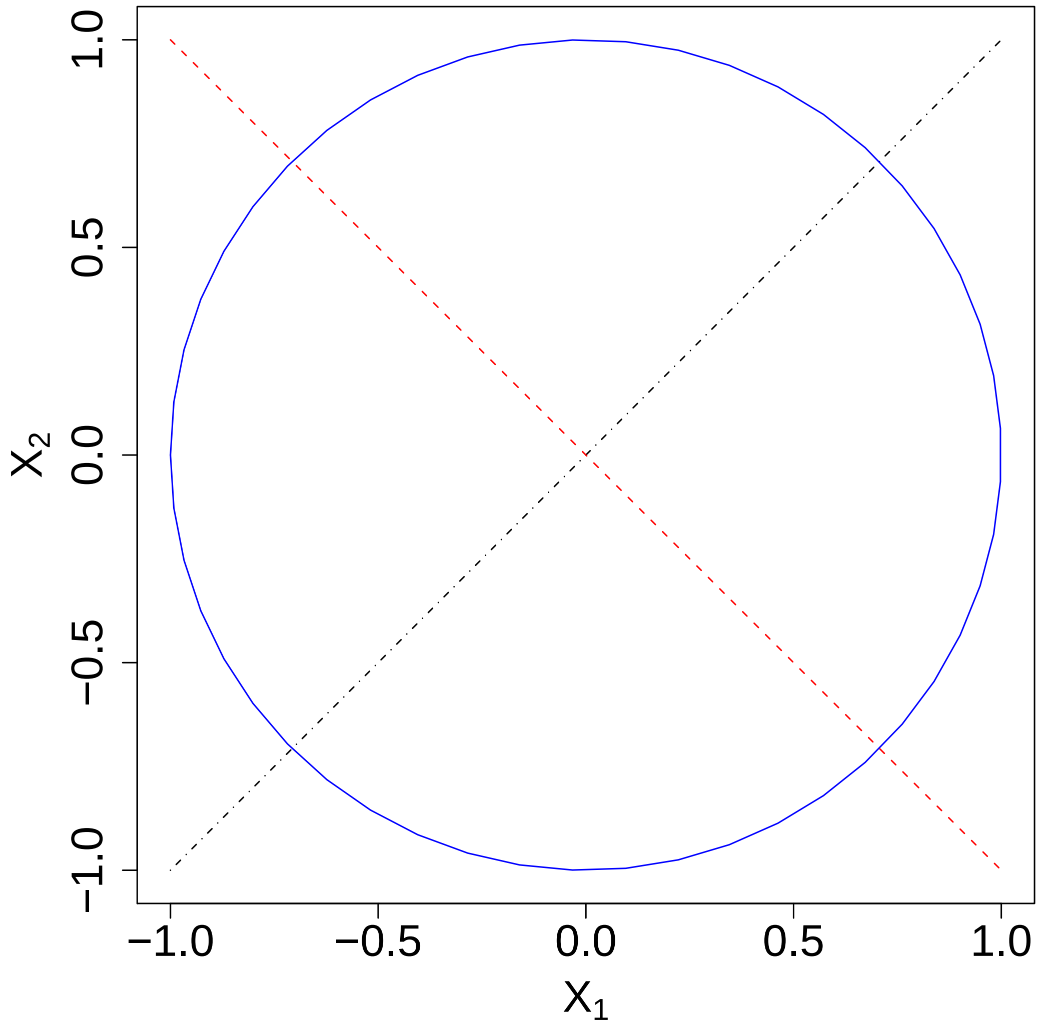
- If X_i is independent of Y, then
  - IMP(i) is a linear combination of independent chi-squared variables
  - Use Satterthwaite (1946) method to approximate distribution of IMP(i)
- Cut-off score for separating important from unimportant variables is the upper-α quantile of the corresponding chi-squared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


sex
weight
width
length3
length2
length1
height
0
10
20
30
40
Importance scores

Three-class problem with 8 predictors



GUIDE trees with kernel and k-NN node models have no splits

Importance scores for three-class problem based on 100 observations for each class

Scaled	Unscaled	Variable	Rank
100.0	2.35	x2	1
53.9	1.27	x1	2
48.4	1.14	x7	3
35.1	0.83	x6	4
29.8	0.70	x5	5
27.1	0.64	x4	6
19.6	0.46	x3	7
13.0	0.30	x8	8

Variables with unscaled scores greater than 1 are important

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 non-categorical, 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 non-categorical, 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 non-categorical, 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 non-categorical, $\hat{Y}$ 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 non-categorical, 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 non-categorical 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) If a split is on a nonmissing value:

      i. missing values in training cases are temporarily replaced by node class means for passing through the split

      ii. missing values in test cases are temporarily replaced by node mean (over all classes) for passing through split

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 $p_w(j|t) = \frac{\text{sum of class } j \text{ weights in } t}{\text{total weight in } t}$ 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)

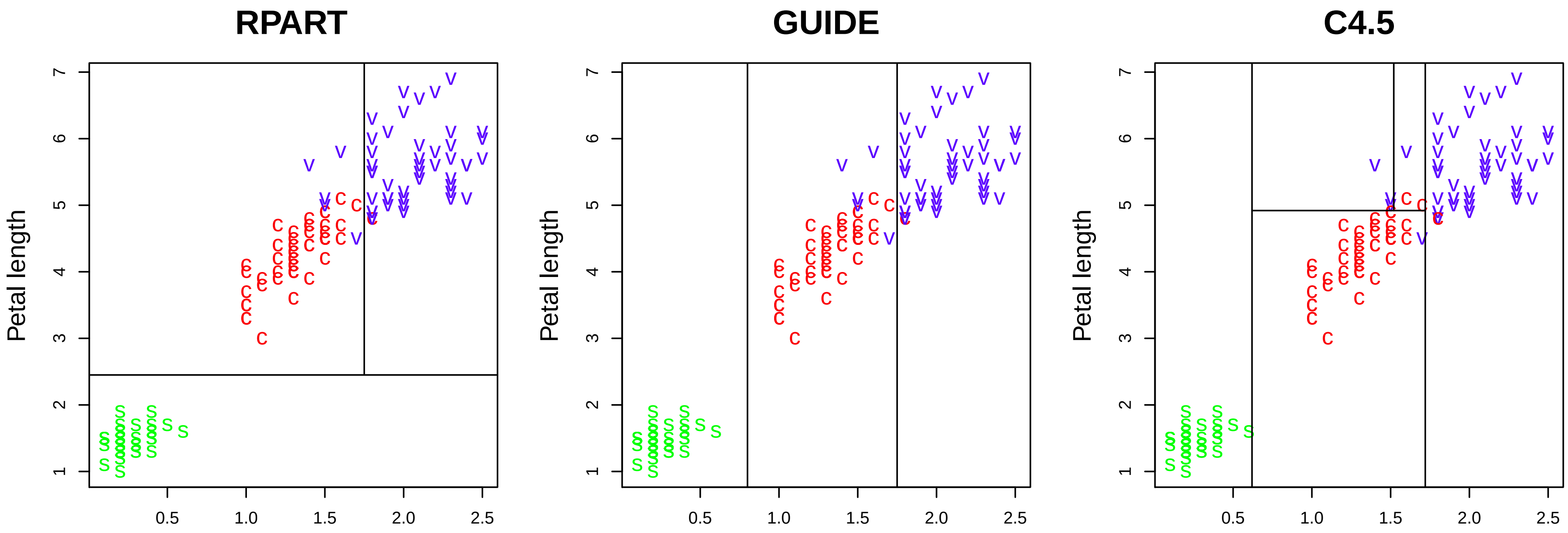
$$p = 1 - \frac{N_E(t)}{N_E(t) + (N(t) - N_E(t) + 1)F_{\nu_1,\nu_2;0.75}}$$

- 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

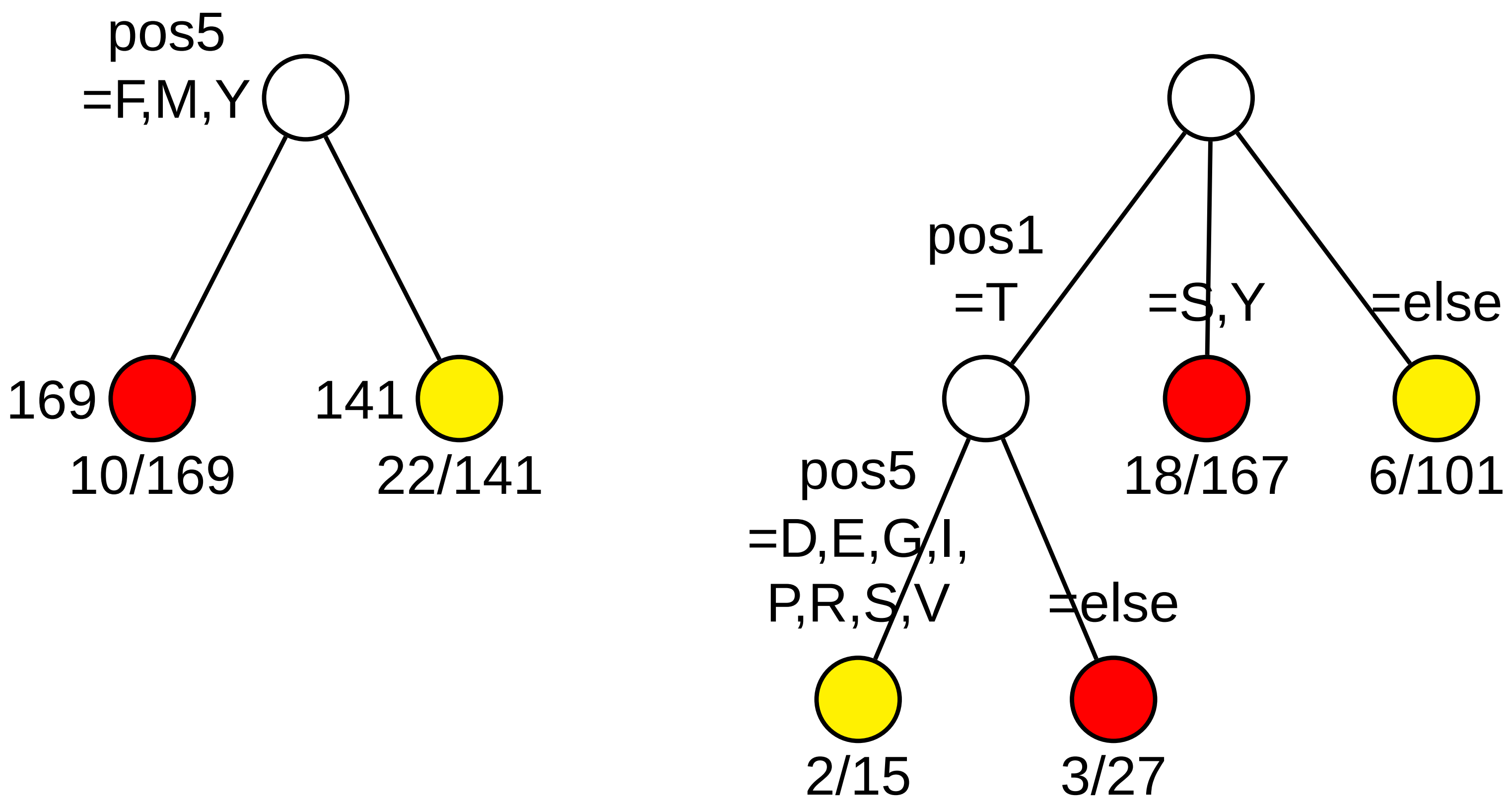
C4.5 computer program

- The original C source for C4.5 is available at `http://www.rulequest.com/Personal/c4.5r8.tar.gz`
- A tutorial is at `http://www2.cs.uregina.ca/~dbd/cs831/notes/ml/dtrees/c4.5/tutorial.html`
- A java implementation (called J48) is included in the WEKA package `http://www.cs.waikato.ac.nz/ml/weka/`

RPART, GUIDE and C4.5 trees for iris data



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

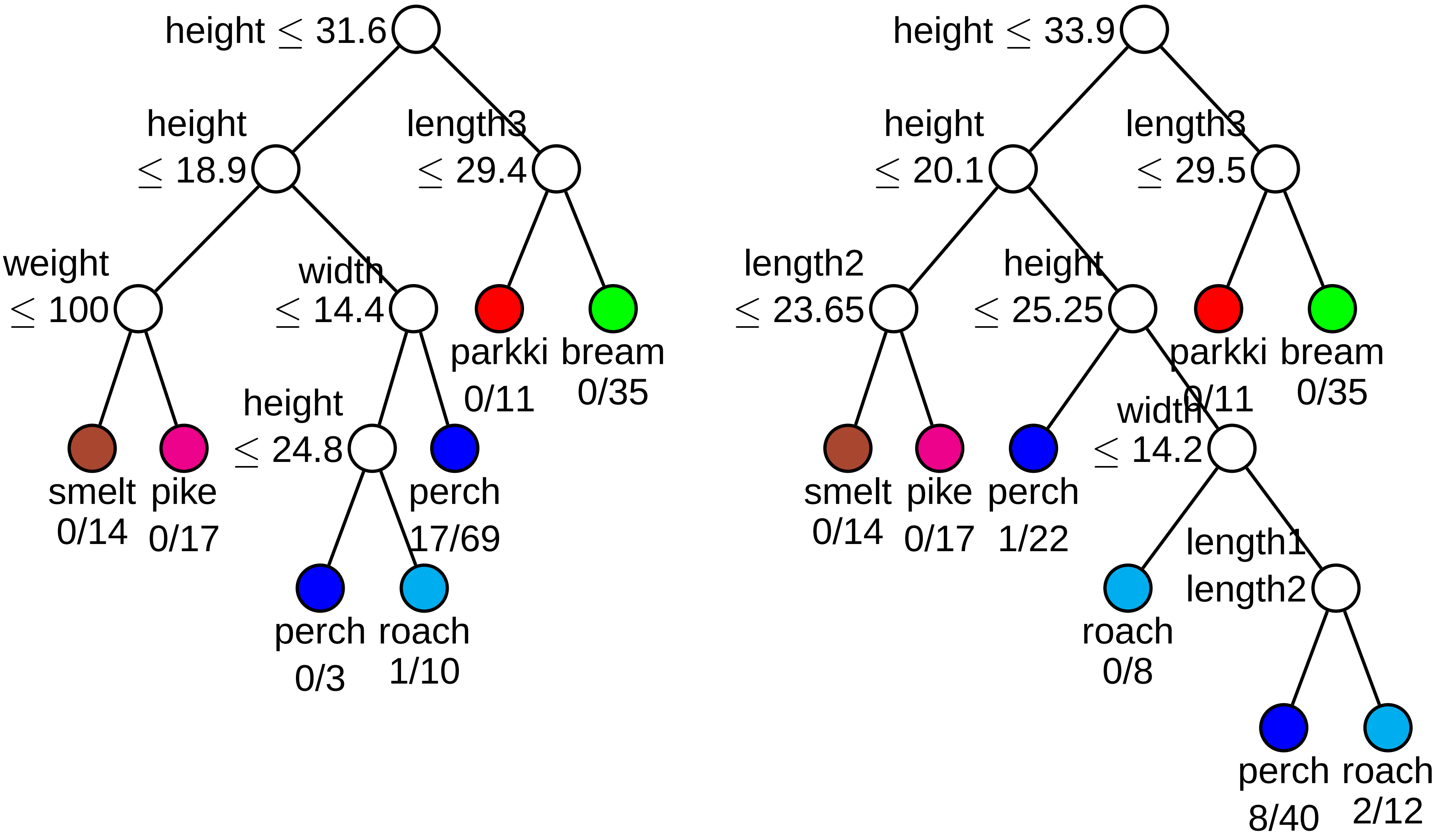


Red denotes binder, yellow denotes non-binder

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 (18 and 11 misclassified)



How to use C4.5

- C4.5 uses a “stem” file name structure
- For the fish example, we can use `fish` as the stem
- C4.5 requires two files: `fish.names` and `fish.data`
- These two files can be produced by GUIDE using option 3
- After these two files are constructed, the program is executed by the command:

```
c4.5 -f fish
```

Creating C4.5 names and data files with GUIDE

```
Choose one of the following options:
1. Read the warranty disclaimer
2. Fit a model
3. Convert data to other formats
4. Create a batch input file
5. Rank and select regressor variables
Input your choice: 3
Input name of log file: log
Input 1 if D variable is categorical, 2 if real, 0 if none
([0:2], <cr>=1):
Input name of data description file (max 100 chars;
enclose within quotes if it contains spaces): fish.dsc
Reading data description file ...
Training sample file: fish.dat
Missing value code: NA
Warning: N variables changed to S
Dependent variable is species
Length of longest data entry =  9
Total number of cases =         159
Number of classes =            7
```

```
 Choose one of the following data formats:
               Field  Miss.val.codes
No. Name       Separ  char.   numer. Remarks
------------------------------------------------------------------
 1  R/Splus    space  NA      NA     1 line/case, var names on 1st line
 2  SAS        space  .       .      strings trunc., spaces -> '_'
 3  TEXT       comma  empty   empty  1 line/case, var names on 1st line
 4  STATISTICA comma  empty   empty  1 line/case, commas stripped
                                     var names on 1st line
 5  SYSTAT     comma  space   .      1 line/case, var names on 1st line
                                     strings trunc. to 8 chars
 6  BMDP       space          *      strings trunc. to 8 chars
                                     cat values -> integers (alph. order)
 7  DATADESK   space  ?       *      1 line/case, var names on 1st line
                                     spaces -> '_'
 8  MINITAB    space          *      cat values -> integers (alph. order)
                                     var names trunc. to 8 chars
 9  NUMBERS    comma  NA      NA     1 line/case, var names on 1st line
                                     cat values -> integers (alph. order)
10  C4.5       comma  ?       ?      1 line/case, dependent variable last
11  ARFF       comma  ?       ?      1 line/case
------------------------------------------------------------------
 Input your choice ([0:11], <cr>=3):10
Input stem name of the new data files
(suffices .data and .names will be appended): fish
New data files are fish.data and fish.names
```

fish.names

```
| Classes
bream, parkki, perch, pike, roach, smelt, whitefish

| Attributes
weight: continuous
length1: continuous
length2: continuous
length3: continuous
height: continuous
width: continuous
sex: female,male,unknown
```

fish.data

```
242,23.2,25.4,30,38.4,13.4,unknown,bream
290,24,26.3,31.2,40,13.8,unknown,bream
340,23.9,26.5,31.1,39.8,15.1,unknown,bream
363,26.3,29,33.5,38,13.3,unknown,bream
430,26.5,29,34,36.6,15.1,unknown,bream
450,26.8,29.7,34.7,39.2,14.2,unknown,bream
500,26.8,29.7,34.5,41.1,15.3,unknown,bream
390,27.6,30,35,36.2,13.4,unknown,bream
450,27.6,30,35.1,39.9,13.8,unknown,bream
...
```

Results from “c4.5 -f fish”

```
Read 159 cases (7 attributes) from fish.data
Decision Tree:
height <= 31.6 :
|   height <= 18.9 :
|   |   weight <= 100 : smelt (14.0)
|   |   weight > 100 : pike (17.0)
|   height > 18.9 :
|   |   width <= 14.3 :
|   |   |   height <= 24.8 : perch (3.0)
|   |   |   height > 24.8 : roach (10.0/1.0)
|   |   width > 14.3 :
|   |   |   height <= 27.7 :
|   |   |   |   sex = male: perch (2.0)
|   |   |   |   sex = unknown: perch (22.0/1.0)
|   |   |   |   sex = female:
|   |   |   |   |   length2 > 25.4 : perch (13.0)
|   |   |   |   |   length2 <= 25.4 :
|   |   |   |   |   |   length3 <= 24.1 : perch (4.0/1.0)
|   |   |   |   |   |   length3 > 24.1 : roach (4.0)
|   |   |   height > 27.7 :
|   |   |   |   length2 <= 27 :
```

```
|   |   |   |   |   length2 <= 25 : perch (7.0/3.0)
|   |   |   |   |   length2 > 25 :
|   |   |   |   |   |   weight <= 270 : whitefish (2.0)
|   |   |   |   |   |   weight > 270 : roach (2.0)
|   |   |   |   length2 > 27 :
|   |   |   |   |   sex = female: perch (7.0/1.0)
|   |   |   |   |   sex = male: perch (0.0)
|   |   |   |   |   sex = unknown:
|   |   |   |   |   |   length1 <= 29.5 : whitefish (2.0)
|   |   |   |   |   |   length1 > 29.5 : perch (4.0/1.0)
height > 31.6 :
|   length3 <= 29.4 : parkki (11.0)
|   length3 > 29.4 : bream (35.0)
```

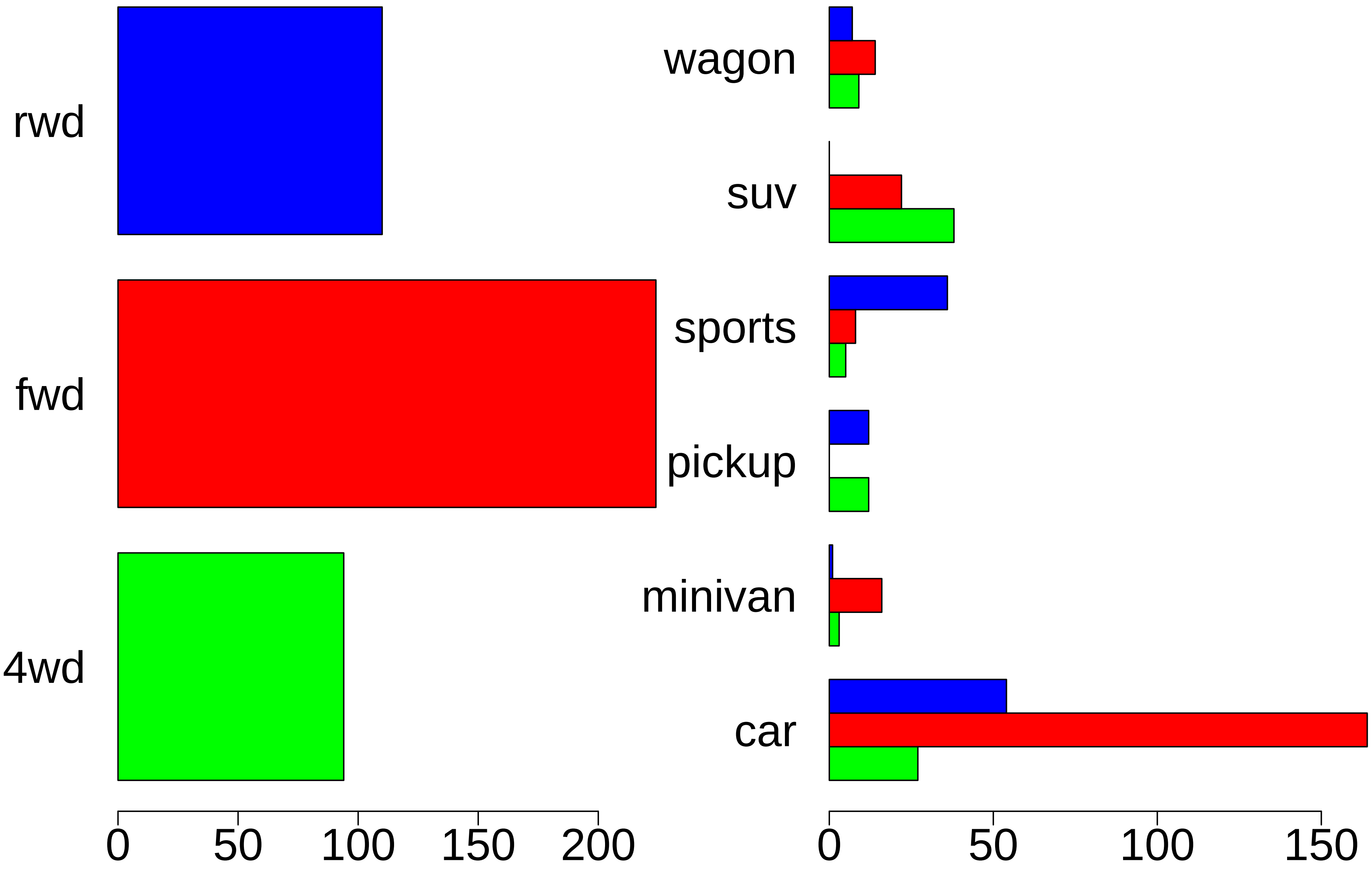
Results (cont'd.)

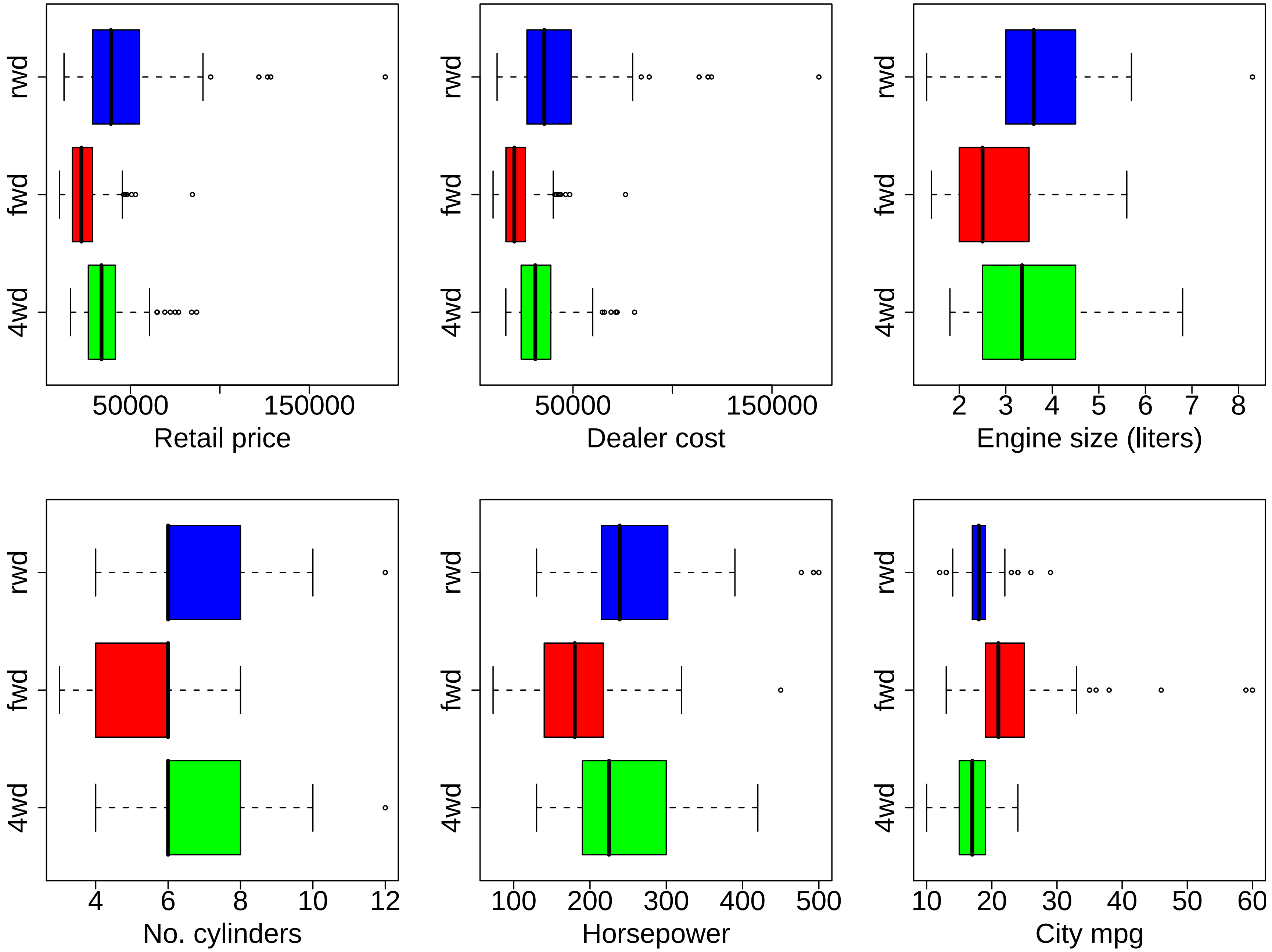
```
Simplified Decision Tree:
height <= 31.6 :
|   height <= 18.9 :
|   |   weight <= 100 : smelt (14.0/1.3)
|   |   weight > 100 : pike (17.0/1.3)
|   height > 18.9 :
|   |   width > 14.3 : perch (69.0/20.1)
|   |   width <= 14.3 :
|   |   |   height <= 24.8 : perch (3.0/1.1)
|   |   |   height > 24.8 : roach (10.0/2.4)
height > 31.6 :
|   length3 <= 29.4 : parkki (11.0/1.3)
|   length3 > 29.4 : bream (35.0/1.4)

Evaluation on training data (159 items):
         Before Pruning           After Pruning
        ----------------   ---------------------------
        Size      Errors   Size      Errors   Estimate
          33    8( 5.0%)     13   18(11.3%)    (18.2%)   <<
```

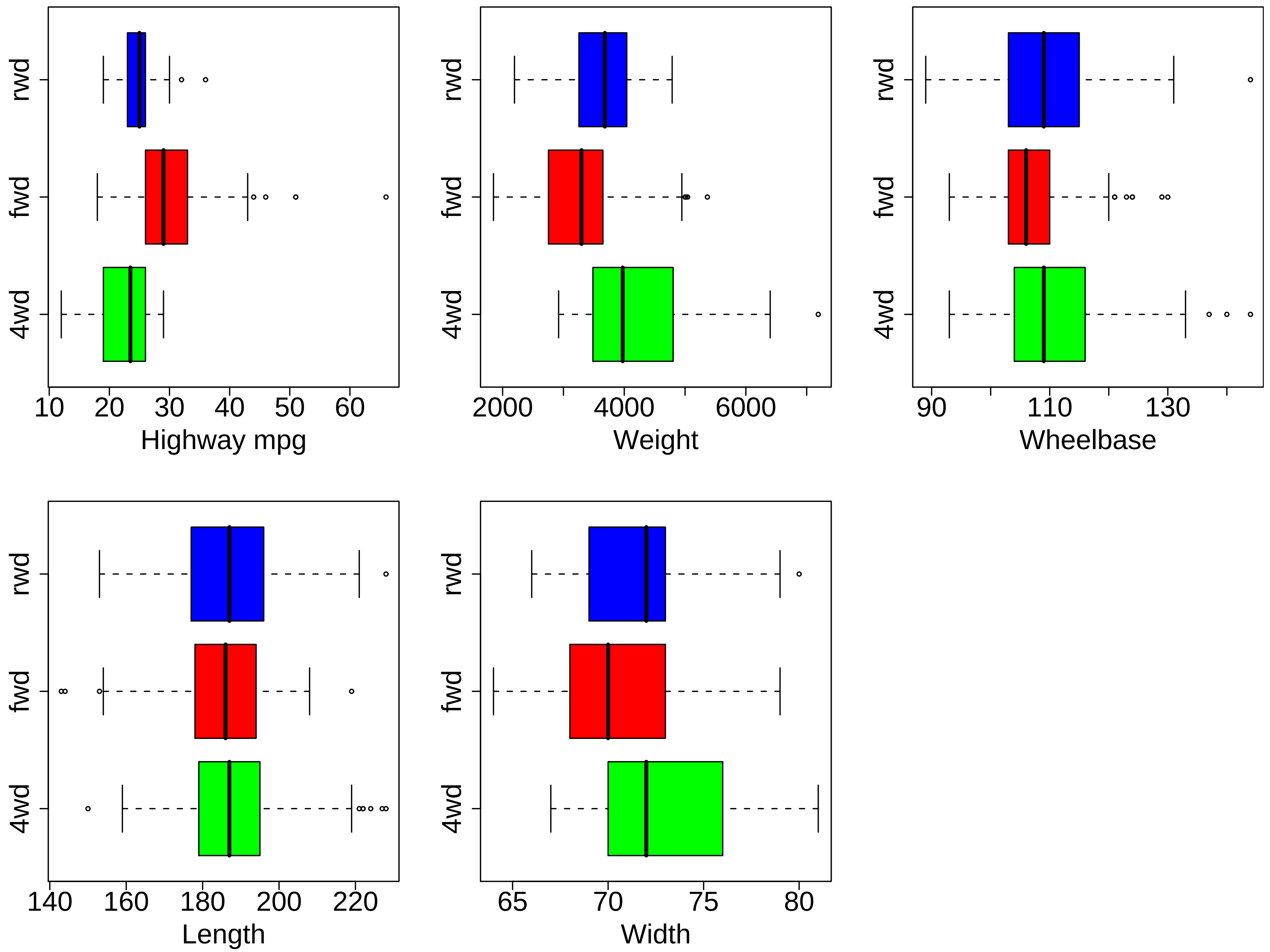
Predicting drive type of vehicles

- 428 new vehicles from the 2004 model year
- 14 predictor variables
- Ref: *Journal of Statistics Education*

Variable	Definition	Values (#unique values in parentheses)
Region	Manufacturer region	Asia, Europe, U.S. (3)
Make	Make of car	Acura, Audi, etc. (38)
Type	Type of car	Car, minivan, pickup, sports car, suv, wagon (6)
Drive	Drive type	Front, rear, four-wheel drive (3)
Rprice	Suggested retail price	U.S. dollars (410)
Dcost	Dealer cost	U.S. dollars (425)
Engnsz	Size of engine	liters (43)
Cylin	Number of cylinders	-1 for the rotary-engine Mazda RX-8 (8)
Hp	Horsepower	hp (110)
City	City miles/gallon	miles (29)
Hwy	Highway miles/gallon	miles (32)
Weight	Weight of car	pounds (347)
Whlbase	Length of wheel base	inches (40)
Length	Length of car	inches (66)
Width	Width of car	inches (18)


rwd
fwd
4wd
0
50
100
150
200
wagon
suv
sports
pickup
minivan
car
0
50
100
150

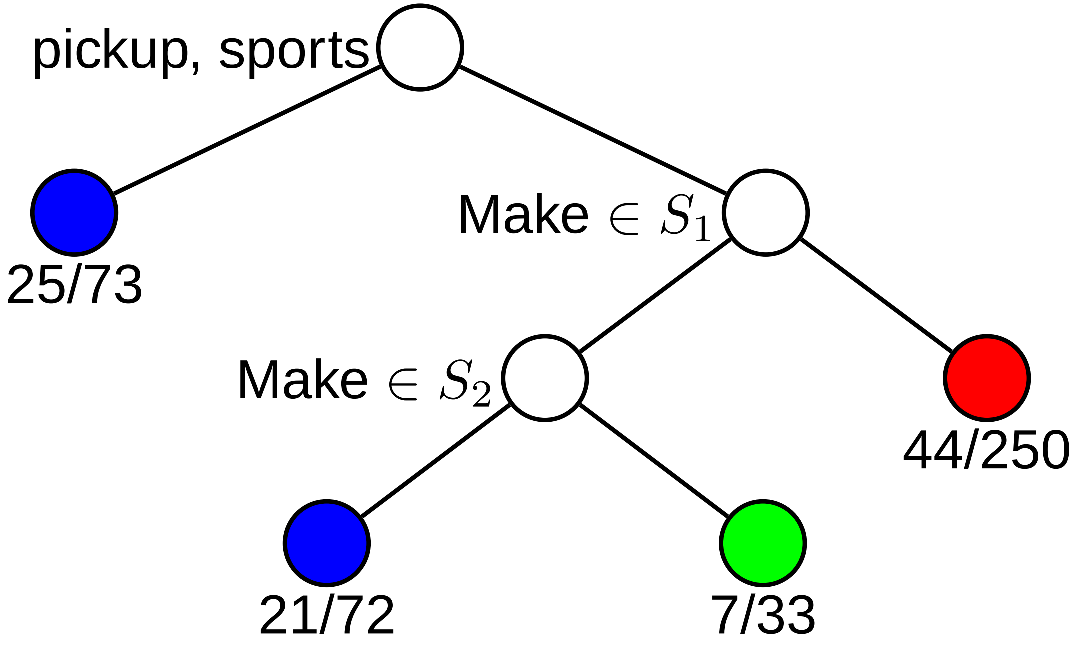

rwd
fwd
4wd
50000
150000
Retail price
rwd
fwd
4wd
50000
150000
Dealer cost
rwd
fwd
4wd
2
3
4
5
6
7
8
Engine size (liters)
rwd
fwd
4wd
4
6
8
10
12
No. cylinders
rwd
fwd
4wd
100
200
300
400
500
Horsepower
rwd
fwd
4wd
10
20
30
40
50
60
City mpg


rwd
fwd
4wd
10
20
30
40
50
60
Highway mpg
2000
4000
6000
Weight
90
110
130
Wheelbase
140
160
180
200
220
Length
65
70
75
80
Width

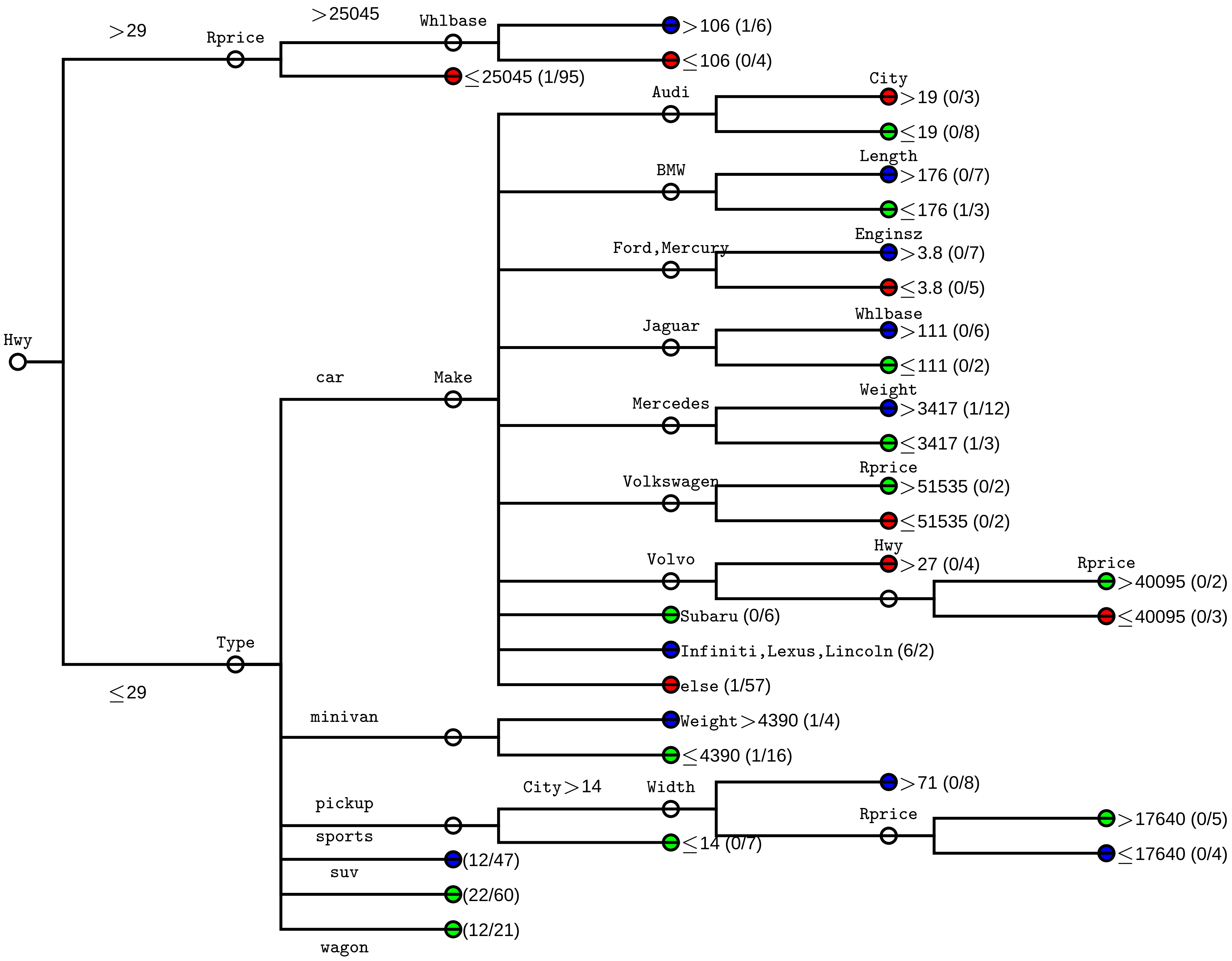
RPART not applicable to this data set

RPART cannot handle categorical variables that take more than 32 values
— **Make** takes 38 values

GUIDE tree for Drive type



- S_1 = {Audi, BMW, Hummer, Infiniti, Isuzu, Jaguar, Jeep, Land-Rover, Lexus, Lincoln, Mercedes, Porsche, Subaru}.
- S_2 = {BMW, Infiniti, Jaguar, Lexus, Lincoln, Mercedes}.
- Red denotes fwd, green denotes 4wd, and blue denotes rwd. Tree misclassifies 64.


Hwy
>29
Rprice
>25045
Whlbase
>106 (1/6)
≤106 (0/4)
≤25045 (1/95)
Audi
City
>19 (0/3)
≤19 (0/8)
BMW
Length
>176 (0/7)
≤176 (1/3)
Ford,Mercury
Enginsz
>3.8 (0/7)
≤3.8 (0/5)
Jaguar
Whlbase
>111 (0/6)
≤111 (0/2)
car
Make
Mercedes
Weight
>3417 (1/12)
≤3417 (1/3)
Volkswagen
Rprice
>51535 (0/2)
≤51535 (0/2)
Volvo
Hwy
>27 (0/4)
Rprice
>40095 (0/2)
≤40095 (0/3)
Subaru (0/6)
Infiniti,Lexus,Lincoln (6/2)
Type
≤29
else (1/57)
minivan
Weight>4390 (1/4)
≤4390 (1/16)
City>14
Width
>71 (0/8)
pickup
Rprice
>17640 (0/5)
sports
≤14 (0/7)
≤17640 (0/4)
(12/47)
suv
(22/60)
(12/21)
wagon

CHAID (Kass 1980 *Applied Statistics*)

- An extension of 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
- Each node can be split into many subnodes

CHAID predictor types

Monotonic: 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. Create categorical predictors out of any ordered predictors. Values of each ordered predictor are grouped into 10 intervals. For categorical predictors, the groups are the categories.

Merge categories. Do for each predictor variable:

1. For classification, take each pair of categories in turn and compute the p-value of the chi-squared test of independence between the categories and the class variable
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 the 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.

CHAID algorithm (cont'd)

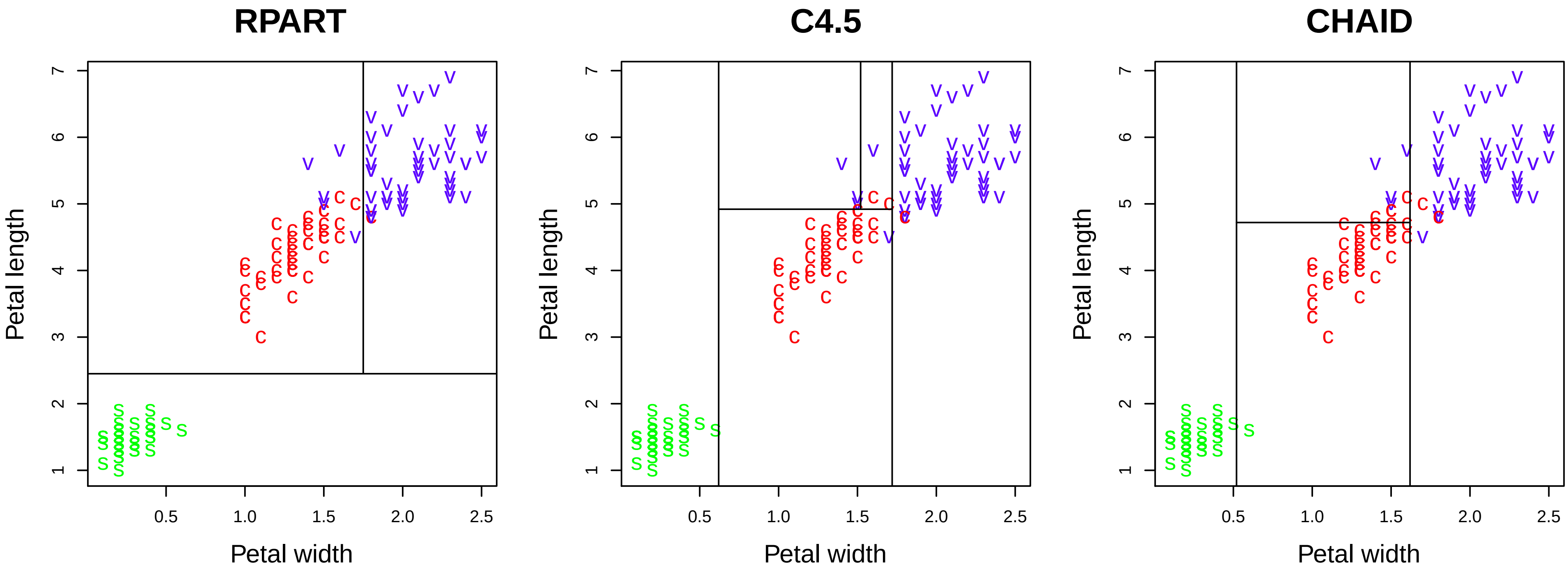
Select split. Compute the Bonferroni-adjusted p-value for each predictor.

If the smallest adjusted $p < \alpha_3$, split the node according to the merged categories of the chosen predictor. Otherwise make the node terminal.

CHAID Bonferroni multipliers

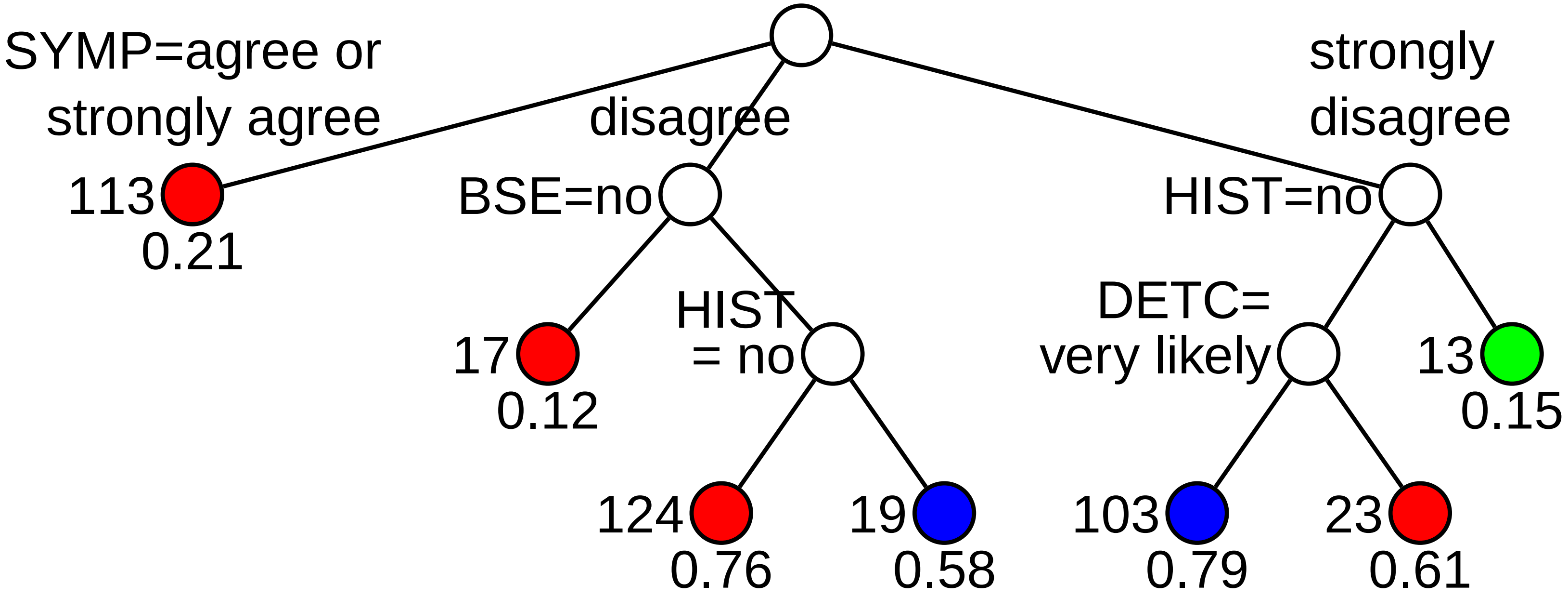
Suppose a predictor with c original categories is merged into r categories. The Bonferroni adjustments to the p-values are:

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



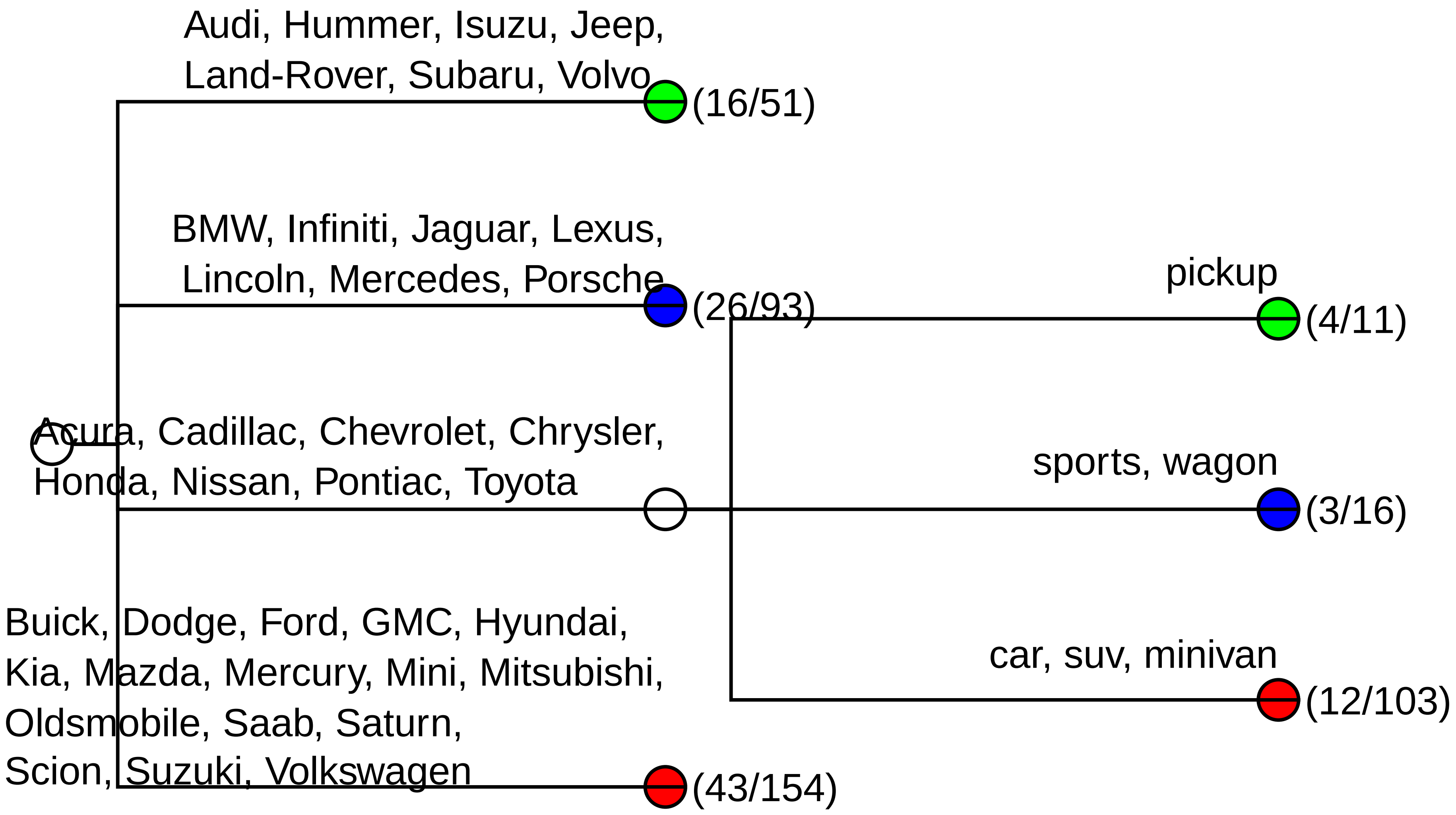
s = Setosa, c = Versicolour, v = Virginica

CHAID tree for mammography data



- Number beneath node is mean misclassification cost, number on left is sample size
- Within 1 year in green, more than one year in blue, never in red
- Total misclassification cost is 228

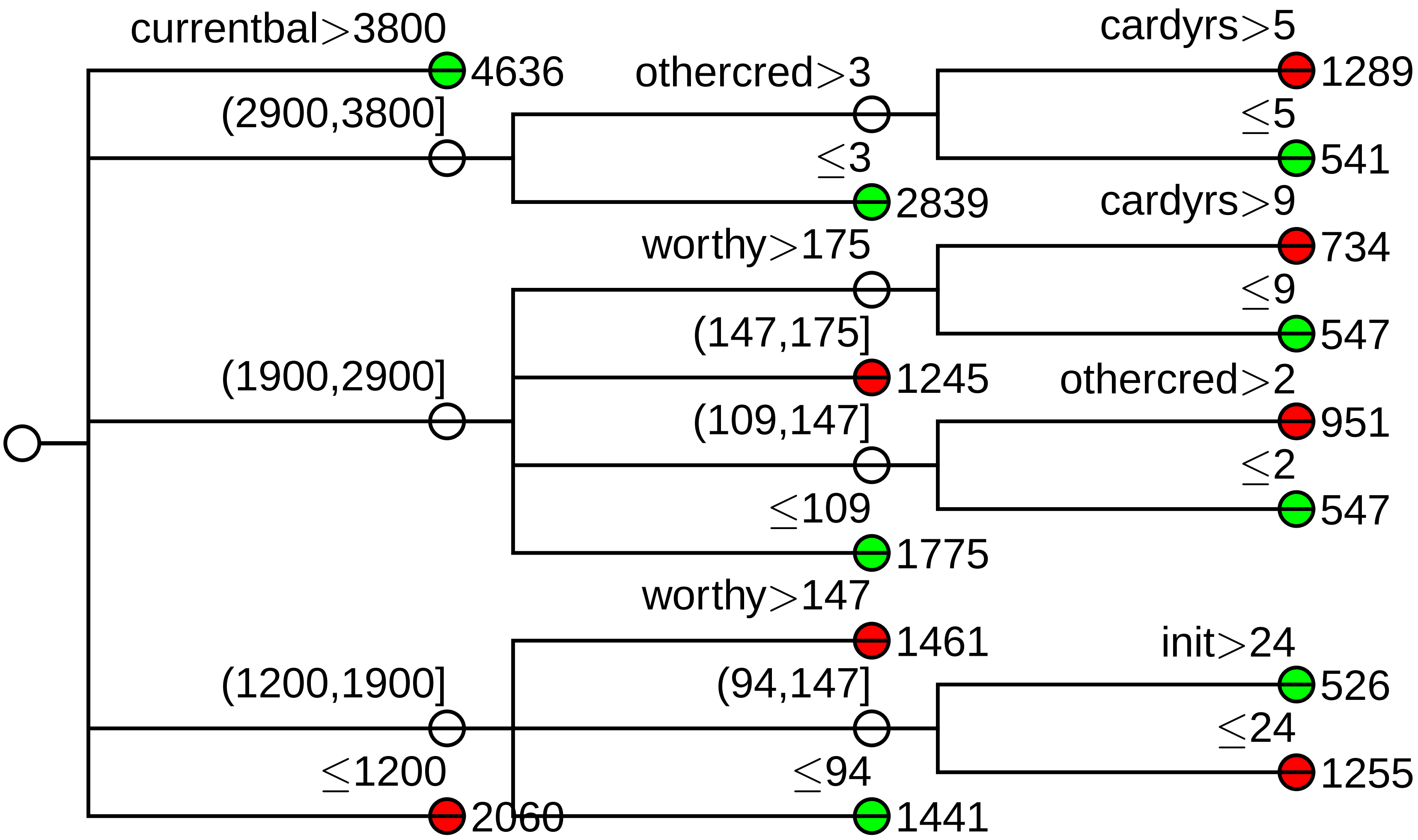
CHAID tree for car data


Audi, Hummer, Isuzu, Jeep,
Land-Rover, Subaru, Volvo
(16/51)
BMW, Infiniti, Jaguar, Lexus,
Lincoln, Mercedes, Porsche
(26/93)
Acura, Cadillac, Chevrolet, Chrysler,
Honda, Nissan, Pontiac, Toyota
Buick, Dodge, Ford, GMC, Hyundai,
Kia, Mazda, Mercury, Mini, Mitsubishi,
Oldsmobile, Saab, Saturn,
Scion, Suzuki, Volkswagen
(43/154)
pickup
(4/11)
sports, wagon
(3/16)
car, suv, minivan
(12/103)

Red denotes fwd, green denotes 4wd, and blue denotes rwd

Tree misclassifies 104 cases

CHAID tree with 11:2 costs for credit card data

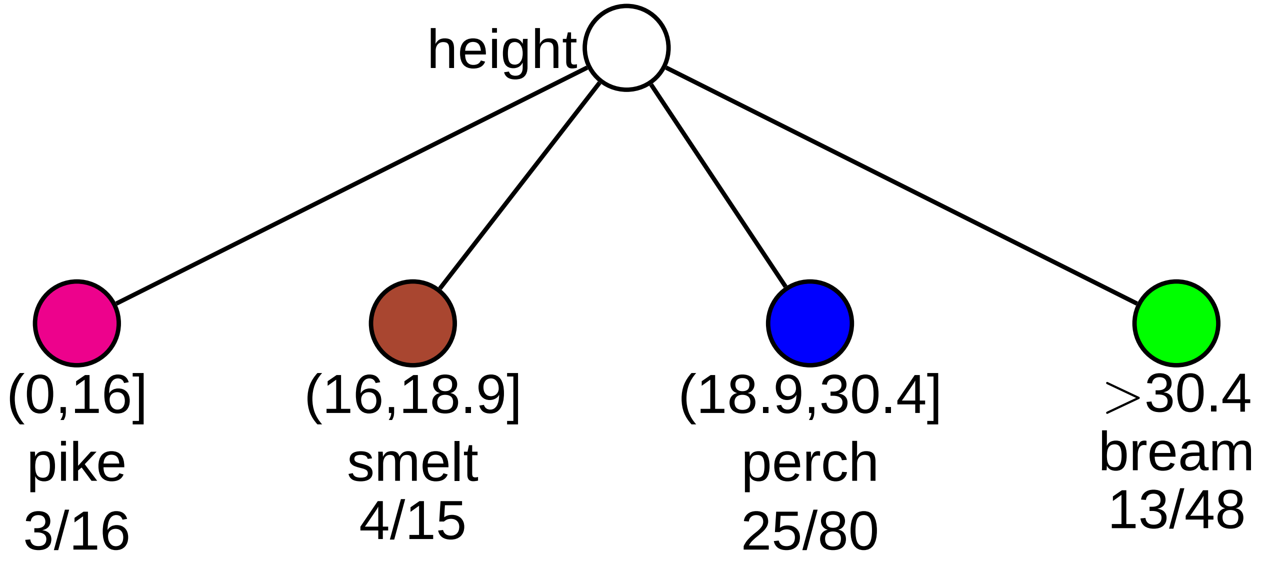


Dissatisfied and satisfied nodes in red and green colors

Total misclassification is 14280.5; sample sizes beside nodes

CHAID does not allow priors to be specified

CHAID tree for fish data (45 misclassified)



Comparisons on 46 datasets using 10-fold CV (Loh, 2009, *Ann. Appl. Statist.*)

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

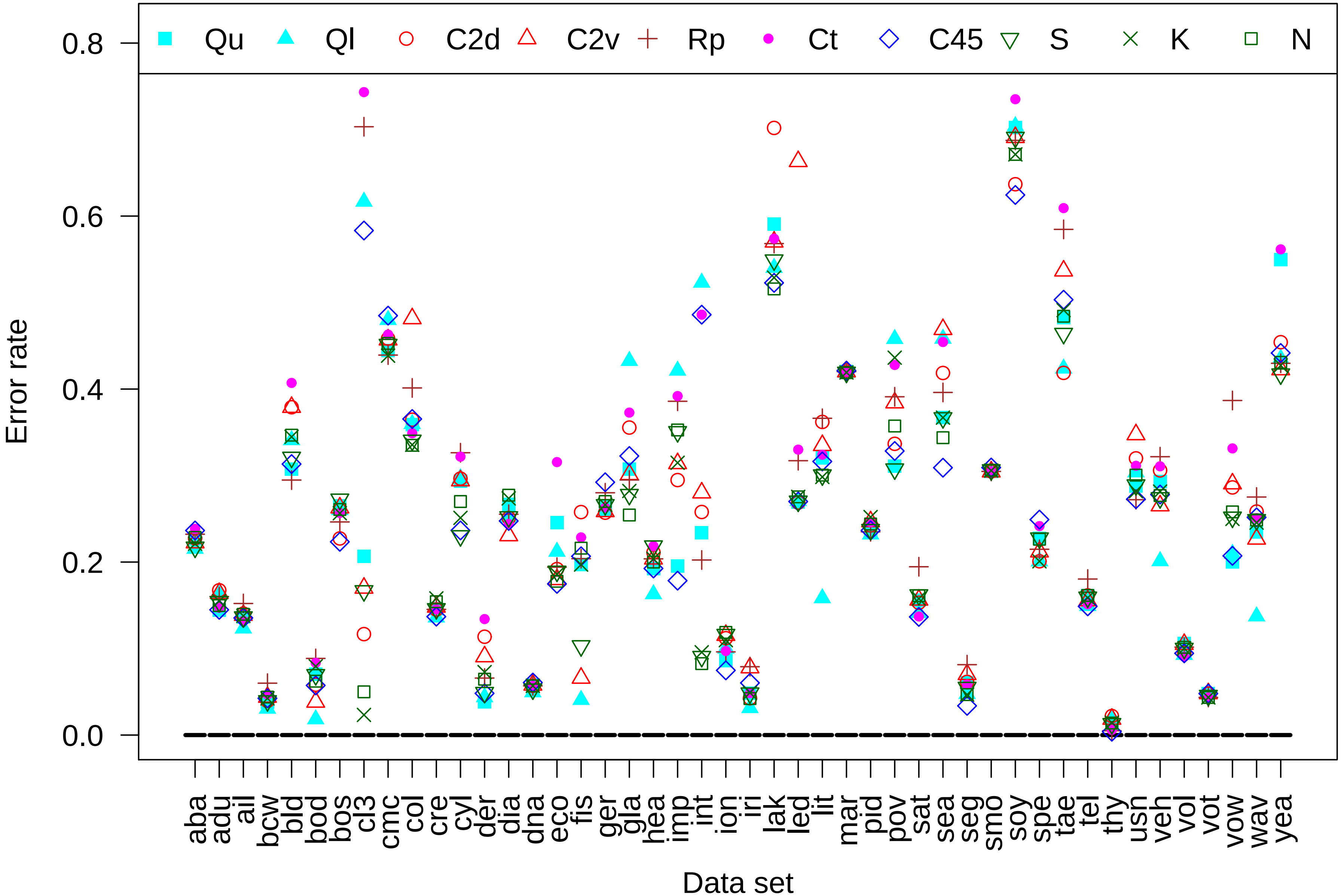
Data	N	J	M	K_1	K_2	C
aba	4177	2	n	7	1	3
adu	45222	2	n	6	8	41
ail	13750	2	n	12	0	0
bcw	683	2	n	9	0	0
bld	345	2	n	6	0	0
bod	507	2	n	24	0	0
bos	506	3	n	12	1	2
cl3	300	3	n	5	3	21
cmc	1473	3	n	5	4	4
col	368	3	y	9	6	9
cre	630	2	y	6	9	15
cyl	540	2	y	19	16	8
der	358	6	n	34	0	0
dia	768	2	n	8	0	0
dna	3186	3	n	0	60	4
eco	336	8	n	7	0	0
fis	159	7	y	6	1	3
ger	1000	2	n	7	13	10
gla	214	6	n	9	0	0
hea	270	2	n	10	3	4
imp	205	6	y	15	10	22
int	1000	2	n	10	0	0

M = #missing values; K_1 = #cont., K_2 = #cat. variables; C = largest #categories

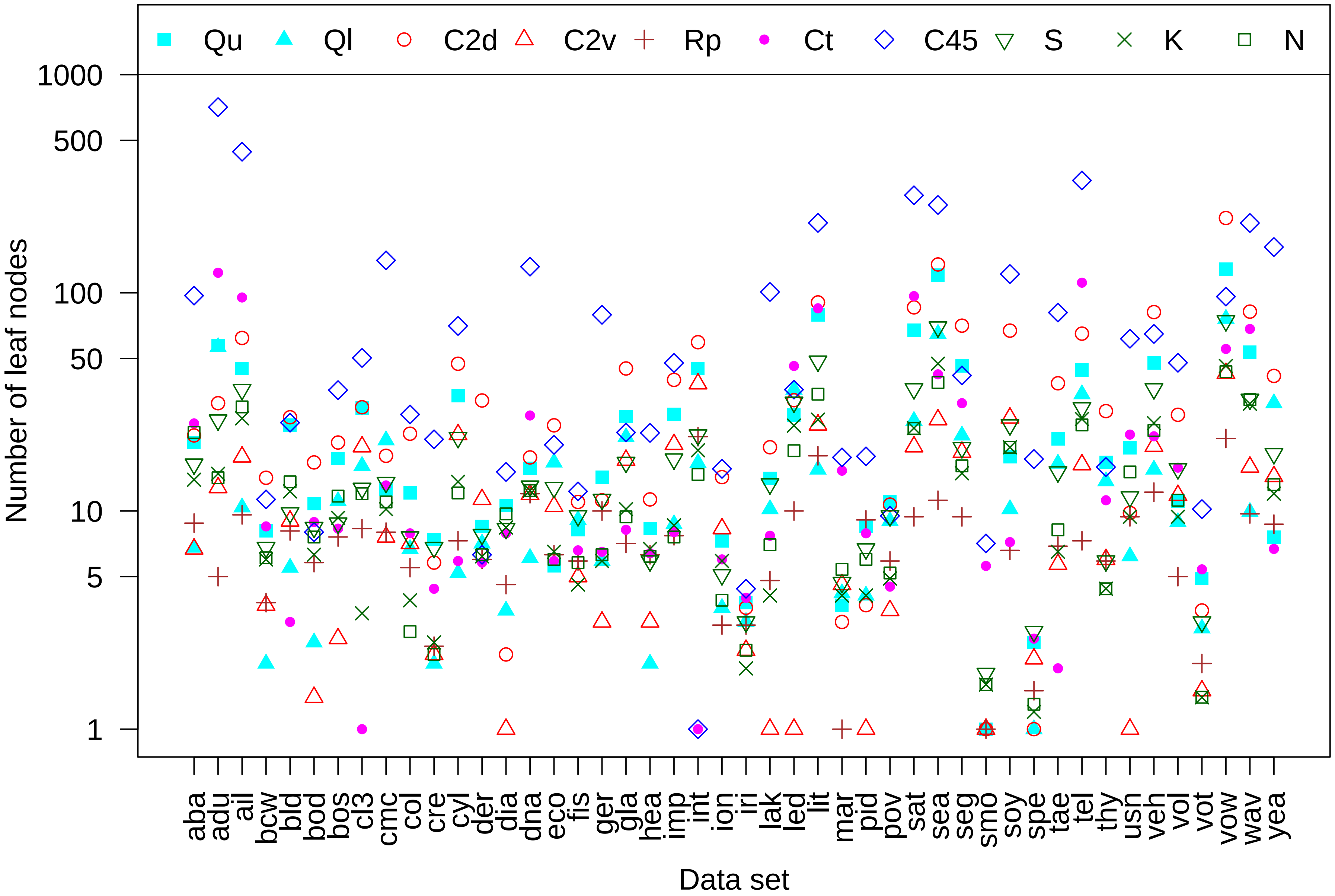
Data	N	J	M	K_1	K_2	C	Data	N	J	M	K_1	K_2	C
ion	351	2	n	34	0	0	soy	307	7	y	0	35	19
iri	150	3	n	4	0	0	spe	267	2	n	44	0	0
lak	259	6	y	13	3	35	tae	151	3	n	3	2	26
led	6000	10	n	7	0	0	tel	19020	2	n	10	0	0
lit	2329	9	n	69	0	0	thy	7200	3	n	21	0	0
mar	8777	10	n	3	1	2	usn	1302	3	y	26	1	2
pid	532	2	n	7	0	0	veh	846	4	n	18	0	0
pov	97	6	y	6	0	0	vol	1521	6	y	4	2	28
sat	6435	6	n	36	0	0	vot	435	2	n	0	16	3
sea	3000	3	y	7	0	0	vow	990	11	n	10	0	0
seg	2310	7	n	19	0	0	wav	3600	3	n	21	0	0
smo	2855	3	n	6	2	5	yea	1484	10	n	8	0	0

M = #missing values; K_1 = #cont., K_2 = #cat. variables; C = largest #categories

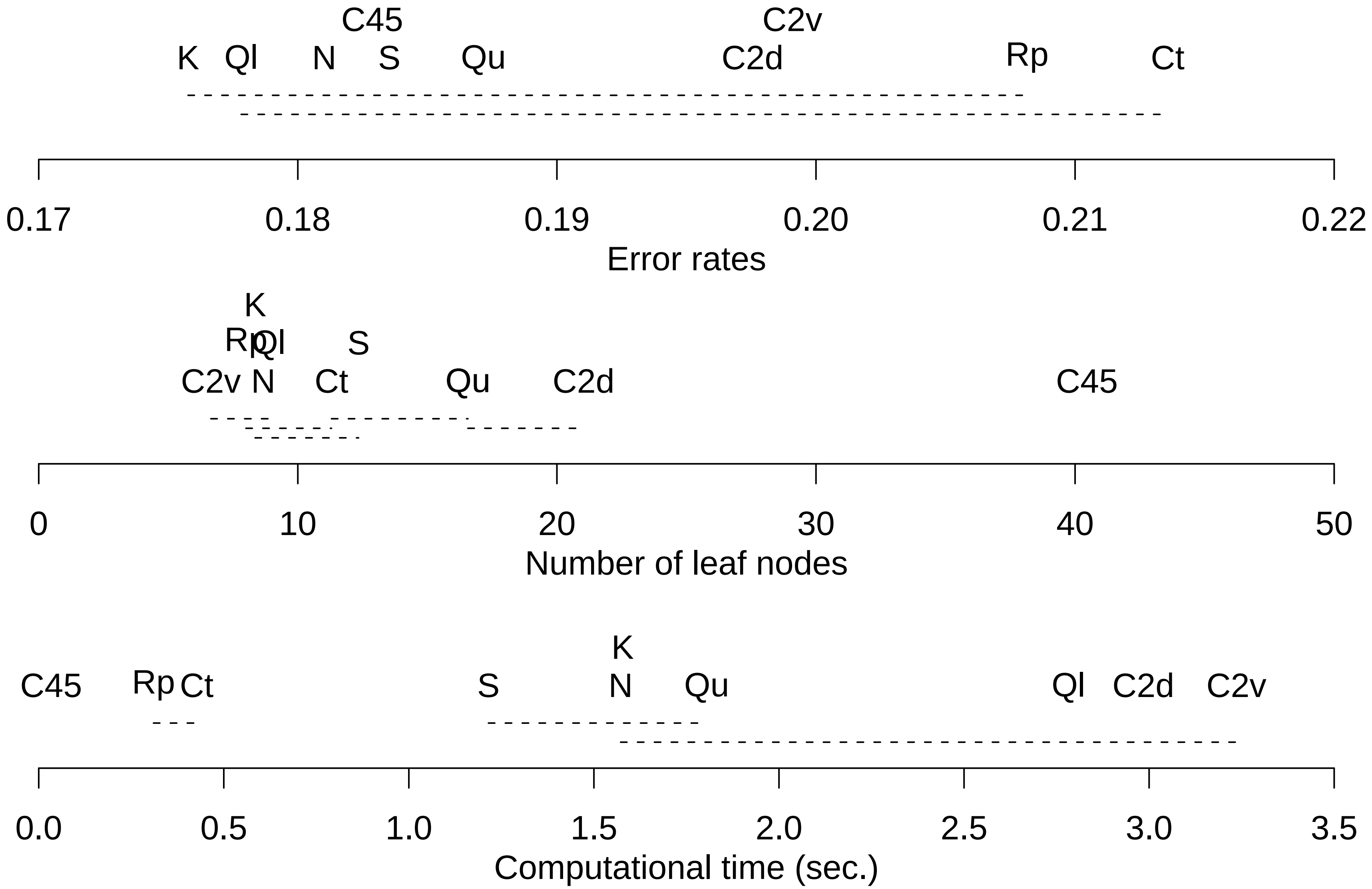
Error rates by dataset



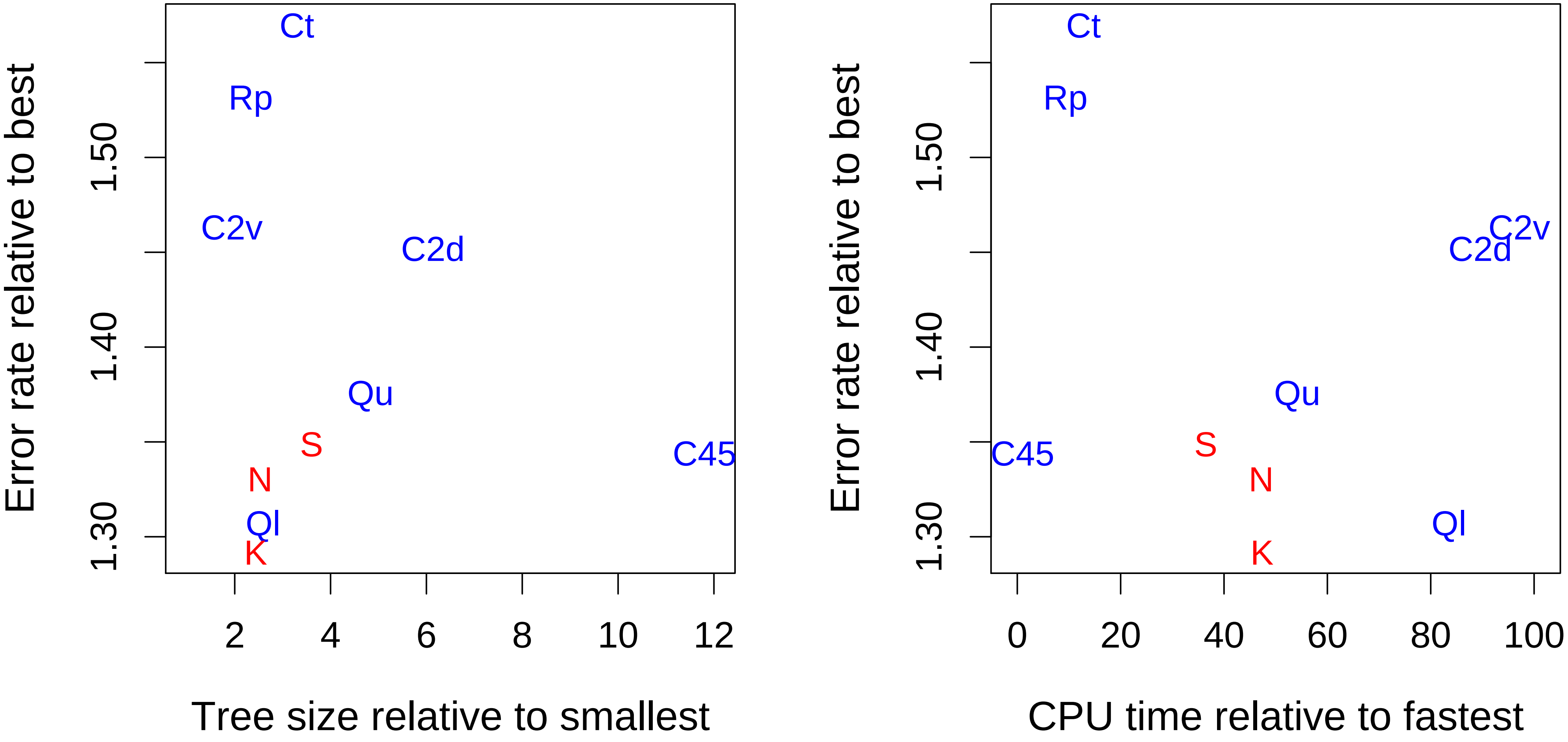
Number of leaf nodes by dataset



Geometric means over 46 datasets



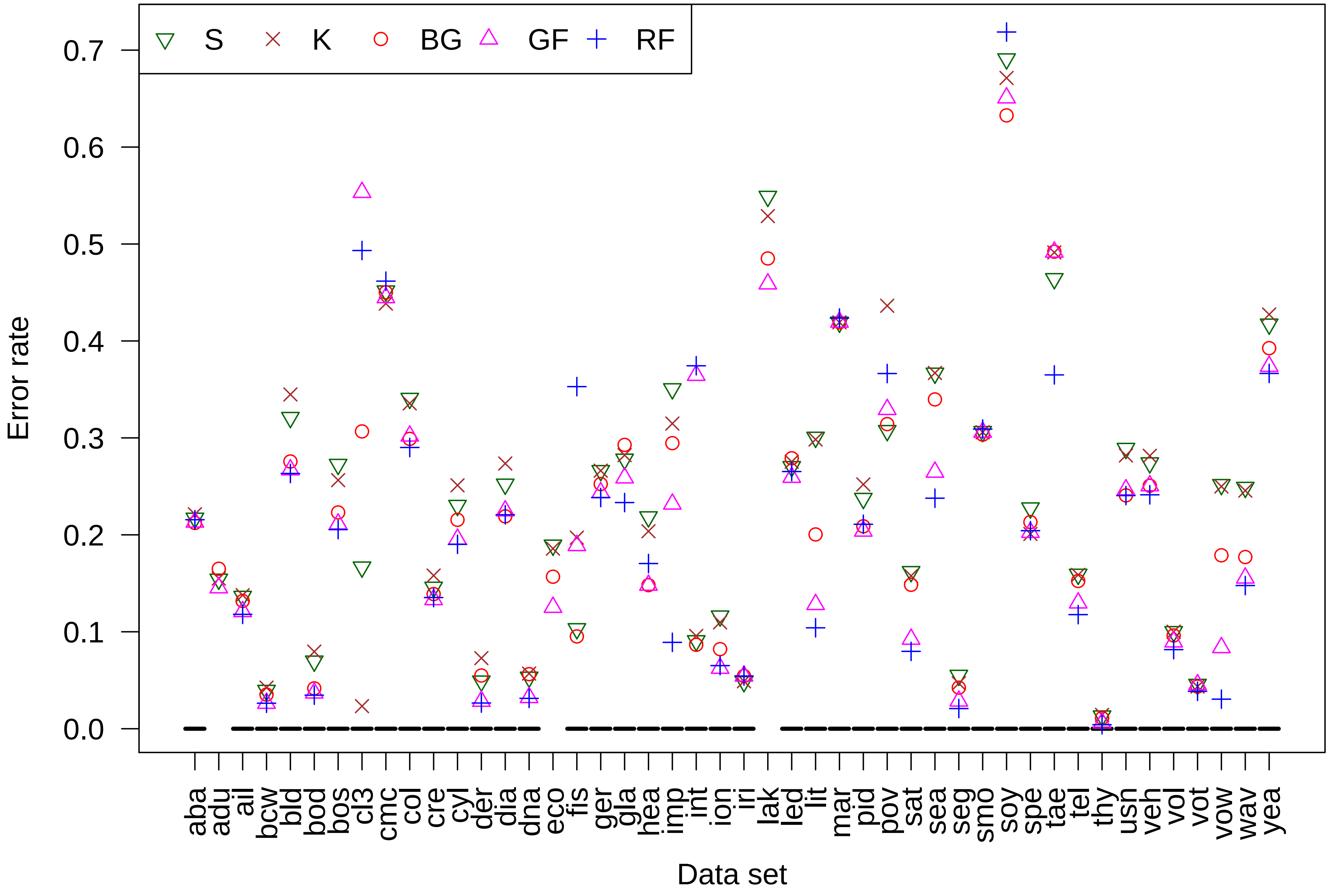
Geometric means relative to best for dataset



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

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 gives incorrect results 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

References

Ahn, H. and Loh, W.-Y. (1994). Tree-structured proportional hazards regression modeling. *Biometrics*, 50:471–485.

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

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.

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

Kahn, M. (2005). An exhalent problem for teaching statistics. *Journal of Statistics Education*, 13(2).

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.

Lim, T.-S., Loh, W.-Y., and Shih, Y.-S. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning Journal*, 40:203–228.

Loh, W.-Y. (1991). Survival modeling through recursive stratification. *Computational Statistics and Data Analysis*, 12:295–313.

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

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

Loh, W.-Y., Man, M., and He, X. (2012). Regression trees for censored data and subgroup identification. Submitted for publication.

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*. In press. arXiv:1209.4690.

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.

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.

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.

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

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