STAT 443
Classification and Regression Trees
9:30–10:45AM TR

Instructor: Wei-Yin Loh

TA: Siyu Wang

Class email list: stat443-1-s21@g-groups.wisc.edu

TA Zoom office hours: 7–9PM Mon

https://uwmadison.zoom.us/j/3486510847

Syllabus topics

1. Quick review of linear regression
2. Basic ideas of classification and regression tree methods
   (a) recursive partitioning
   (b) cross-validation for tree pruning and error estimation
   (c) unbiased variable selection
3. Missing values
4. Quantile, Poisson, logistic, and proportional hazards regression trees
5. Variable importance scores

Requisites

- Knowledge of linear and logistic regression
- R programming and ability to install and learn to use R packages
- Ability to clean and manage data

Learning objectives

1. Learn to use GUIDE to solve hard problems involving
   (a) subgroup identification (COVID-19, precision medicine)
   (b) missing data (socio-economic surveys)
   (c) periodic variables (vehicle crash tests)
   (d) large numbers of variables (genetic data)
   (e) multivariate and longitudinal response variables
   (f) importance ranking of variables
   (g) post-selection inference
2. Understand weaknesses and limitations of other methods
3. Know differences between GUIDE and other tree methods

Grading

5%	class attendance
10%	class participation
10%	quizzes
15%	homework
60%	term project

- Weights applied to standardized scores (z-scores)
- Lowest homework z-score is dropped (you can skip 1 homework)
- Late homeworks will not be graded
- Final grade distribution: 25% A, 25% AB, 25% B, 25% BC and below

UW policy on disabilities & religious observances

- Inform me of your need for instructional accommodations by the end of the third week of the semester, or as soon as possible after a disability has been incurred or recognized (`https://mcburney.wisc.edu`)
- Notify me within the first two weeks of class of the specific days or dates on which you wish to request relief for religious observances

Machine learning methods

1. Regression (linear models, LASSO, support vector machines, trees and forests, etc.)
2. Classification (discriminant analysis, logistic regression, support vector machines, trees and forests, etc.)
3. Clustering
4. Dimensionality reduction (principal components, LASSO, etc.)
5. Boosting and ensemble methods
6. Neural nets and deep learning
7. Reinforcement learning

- With many machine learning methods (e.g., deep learning), machines "learn", but humans do not
- Classification and regression trees help humans learn

European Union
General Data Protection Regulation (GDPR)

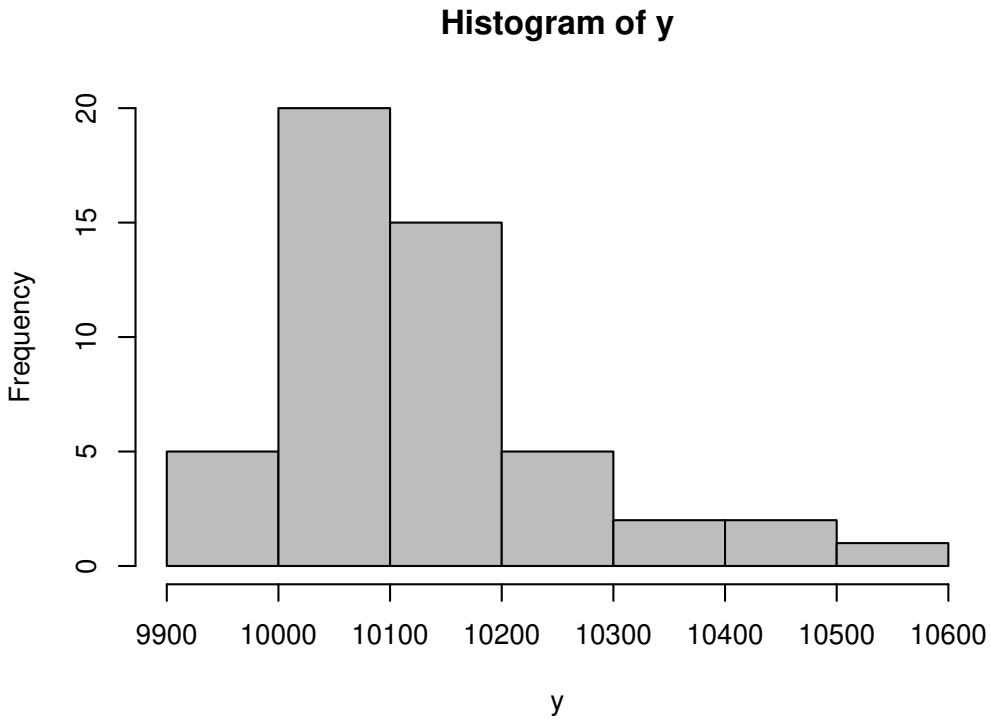
- Most important change in data privacy regulation in 20 years
- Approved by EU Parliament on 14 April 2016; enforced on 25 May 2018
- Organizations not compliant could now face heavy fines
- GDPR designed to
  1. Harmonize data privacy laws across Europe
  2. Protect and empower all EU citizens' data privacy
  3. Reshape the way organizations across the region approach data privacy

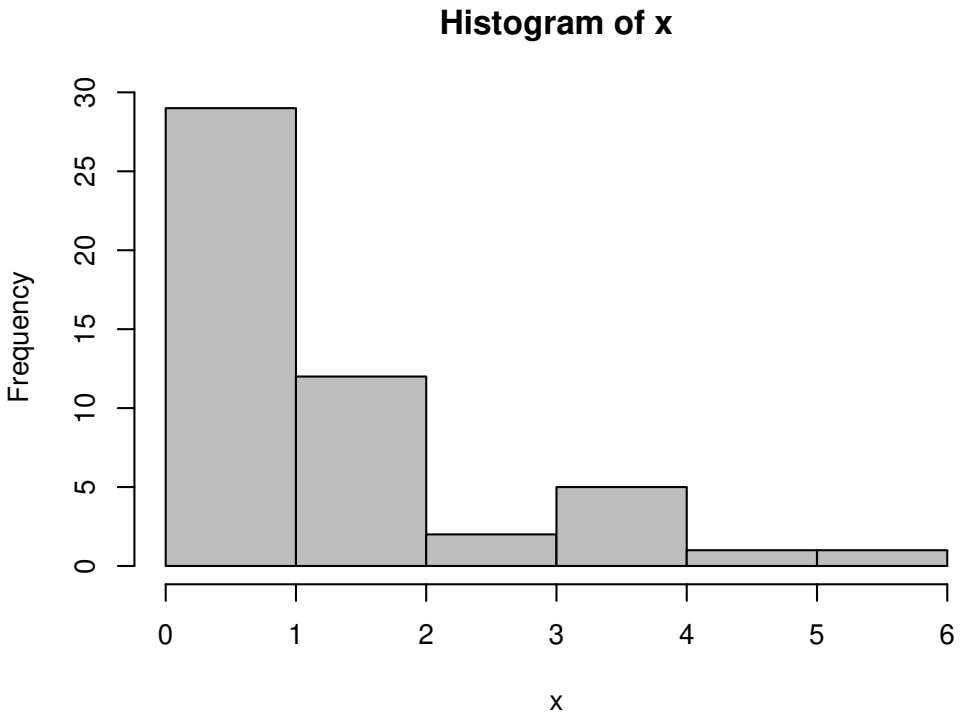
GDPR Article 13 (right to explanation)

Provide the data subject with the following information necessary to ensure fair and transparent processing:

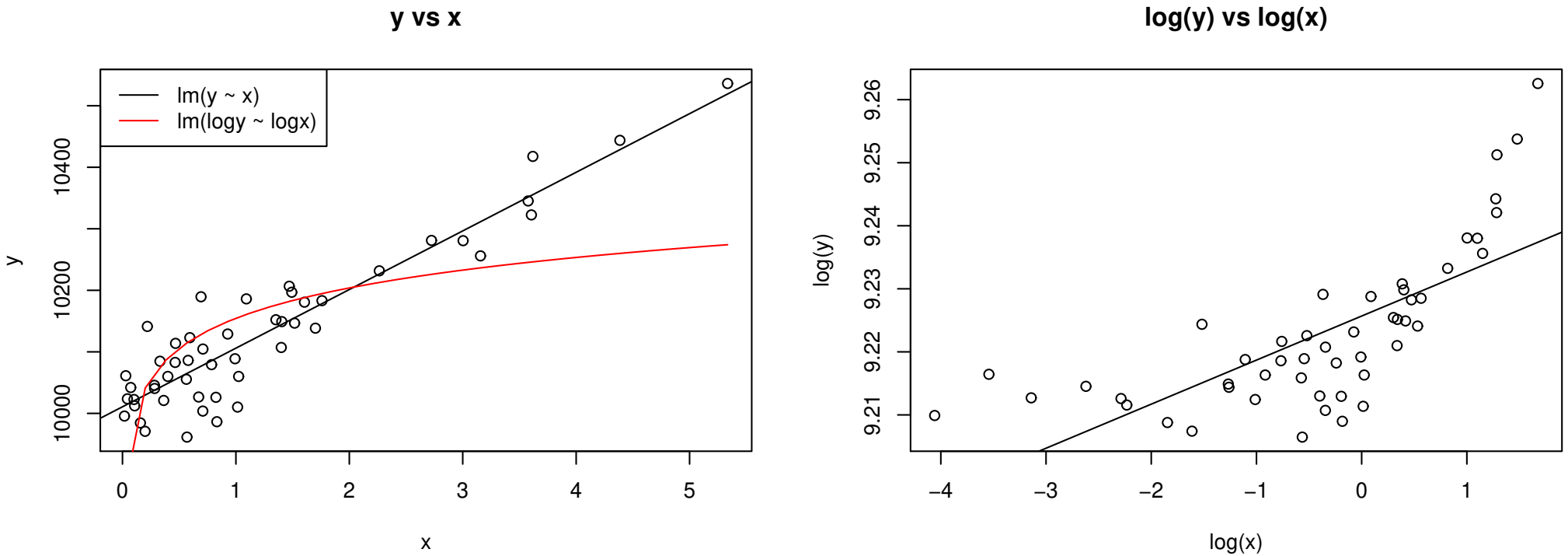
. . . existence of automated decision-making, including profiling, and meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject

Linear regression: histograms of X and Y





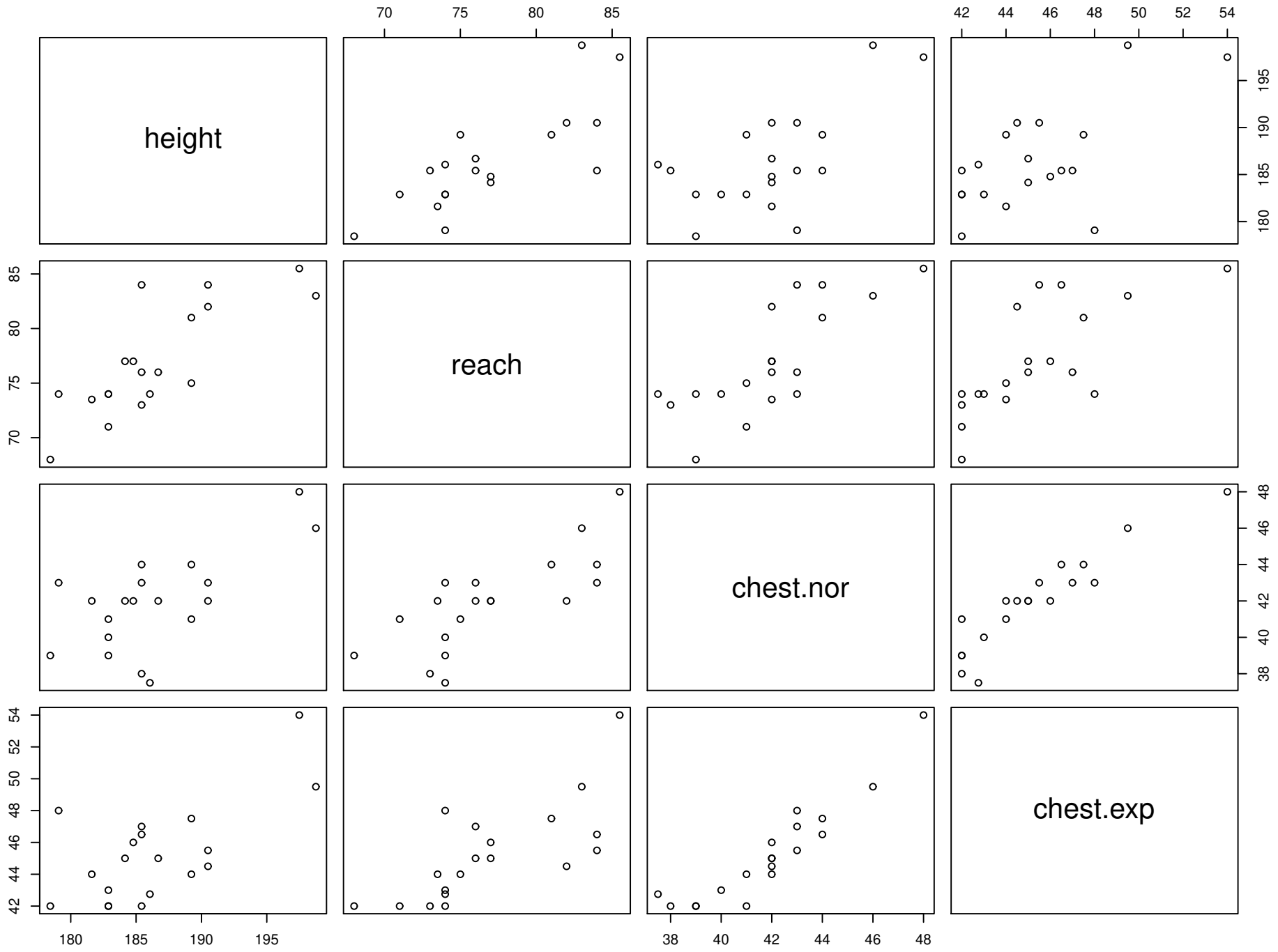
Plots and fits of simulated data



Height, reach and chest (normal & expanded) measurements of 19 heavyweight boxing champions

https://www.boxingscene.com/forums/showthread.php?t=286210

Name	height (cm)	reach (in)	chest.nor (in)	chest.exp (in)
Ali	190.5	84	43	45.5
Baer	189.23	81	44	47.5
Braddock	189.23	75	41	44
Carnera	197.485	85.5	48	54
Charles	182.88	74	39	42
Corbett	185.42	73	38	42
Dempsey	184.785	77	42	46
Foreman	190.5	82	42	44.5
Frazier	181.61	73.5	42	44
Johnson	186.055	74	37.5	42.75
Liston	185.42	84	44	46.5
Louis	186.69	76	42	45
Marciano	178.435	68	39	42
Patterson	182.88	71	41	42
Schmeling	185.42	76	43	47
Sullivan	179.07	74	43	48
Tunney	184.15	77	42	45
Walcott	182.88	74	40	43
Willard	198.755	83	46	49.5


height
reach
chest.nor
chest.exp
70
75
80
85
42
44
46
48
50
52
54
180
185
190
195
38
40

Simple linear regressions

```
lm(formula = height ~ chest.nor)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 130.8738    15.8267   8.269 2.32e-07 ***
chest.nor     1.3243     0.3768   3.514  0.00266 **

lm(formula = height ~ chest.exp)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 134.8947    14.4654   9.325 4.27e-08 ***
chest.exp     1.1373     0.3188   3.568  0.00237 **

lm(formula = height ~ reach)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 120.0299    12.3684   9.705 2.40e-08 ***
reach         0.8624     0.1604   5.376 5.03e-05 ***
```

Predicting with chest.nor and chest.exp

```
lm(formula = height ~ chest.nor)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 130.8738    15.8267   8.269 2.32e-07 ***
chest.nor     1.3243     0.3768   3.514  0.00266 **

lm(formula = height ~ chest.exp)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 134.8947    14.4654   9.325 4.27e-08 ***
chest.exp     1.1373     0.3188   3.568  0.00237 **

lm(formula = height ~ chest.nor + chest.exp)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 131.1085    16.0325   8.178 4.17e-07 ***
chest.nor     0.6114     1.0165   0.601    0.556
chest.exp     0.6549     0.8654   0.757    0.460
```

Predicting with chest.nor and reach

```
lm(formula = height ~ chest.nor)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 130.8738    15.8267   8.269 2.32e-07 ***
chest.nor     1.3243     0.3768   3.514  0.00266 **

lm(formula = height ~ reach)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 120.0299    12.3684   9.705 2.40e-08 ***
reach         0.8624     0.1604   5.376 5.03e-05 ***

lm(formula = height ~ chest.nor + reach)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 118.3070    13.6408   8.673 1.91e-07 ***
chest.nor     0.1699     0.4904   0.346  0.73357
reach         0.7922     0.2611   3.035  0.00789 **
```

Predicting with chest.exp and reach

```
lm(formula = height ~ chest.exp)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 134.8947    14.4654   9.325 4.27e-08 ***
chest.exp     1.1373     0.3188   3.568  0.00237 **

lm(formula = height ~ reach)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 120.0299    12.3684   9.705 2.40e-08 ***
reach         0.8624     0.1604   5.376 5.03e-05 ***

lm(formula = height ~ chest.exp + reach)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 117.2207    13.1218   8.933 1.29e-07 ***
chest.exp     0.2777     0.3812   0.729  0.47674
reach         0.7355     0.2383   3.086  0.00709 **
```

Predicting with all 3 variables

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 130.8738    15.8267   8.269 2.32e-07 ***
chest.nor     1.3243     0.3768   3.514  0.00266 **

(Intercept) 134.8947    14.4654   9.325 4.27e-08 ***
chest.exp     1.1373     0.3188   3.568  0.00237 **

(Intercept) 120.0299    12.3684   9.705 2.40e-08 ***
reach         0.8624     0.1604   5.376 5.03e-05 ***

(Intercept) 118.6661    13.8234   8.584 3.58e-07 ***
chest.nor    -0.4140     0.9040  -0.458  0.65352
chest.exp     0.5501     0.7116   0.773  0.45153
reach         0.7820     0.2648   2.954  0.00986 **
```

Partial residuals

```
lm(formula = height ~ chest.nor + chest.exp + reach)
(Intercept)    chest.nor    chest.exp        reach
   118.6661      -0.4140       0.5501       0.7820

model1 <- lm(height ~ chest.exp + reach)
model2 <- lm(chest.nor ~ chest.exp + reach)
lm(formula = model1$res ~ model2$res)
(Intercept)   model2$res
 -3.759e-16      -0.4140
```

COVID-19 data

- 31,461 patients with COVID-19 between Jan 20 and May 26, 2020, in US
- 21 variables:
  - death during hospitalization
  - age group
  - sex
  - race
  - 16 comorbidities
  - Charlson comorbidity index (weighted sum of comorbidities)
- 4.1% died

COVID-19 variables

`died`	Died while hospitalized (0=no, 1=yes)
`agecat`	Age group (0=18–50, 1=50–59, 2=60–69, 3=70–79, 4=80–90 years)
`race`	American Indian or Alaska Native; Asian; Black or African American; Native Hawaiian or other Pacific Islander; White; Unknown
`sex`	Gender (male/female)
`aids`	AIDS/HIV (0=no, 1=yes)
`cancer`	Any malignancy, including lymphoma and leukemia, except malignant neoplasm of skin (0=no, 1=yes)
`cerebro`	Cerebrovascular disease (0=no, 1=yes)
`CHF`	Congestive heart failure (0=no, 1=yes)
`CPD`	Chronic pulmonary disease (0=no, 1=yes)
`dementia`	Dementia (0=no, 1=yes)
`diabetes`	Diabetes mellitus (0=no, 1=yes)

COVID-19 variables (cont'd.)

`hemipara`	Hemiplegia or paraplegia (0=no, 1=yes)
`metastatic`	Metastatic solid tumor (0=no, 1=yes)
`MI`	Myocardial infarction (0=no, 1=yes)
`mildliver`	Mild liver disease (0=no, 1=yes)
`modsevliv`	Moderate/severe liver disease (0=no, 1=yes)
`PUD`	Peptic ulcer disease (0=no, 1=yes)
`PVD`	Peripheral vascular disease (0=no, 1=yes)
`RD`	Rheumatic disease (0=no, 1=yes)
`renal`	Renal disease (0=no, 1=yes)
`charlson`	CHF + CPD + MI + RD + PUD + PVD + cerebro + dementia + diabetes + mildliver + 2$\times$(cancer + hemipara + renal) + 3$\times$modsevliv + 6$\times$(metastatic + aids)

```
sex                                                 race           cancer
F:17155   American Indian or Alaska Native          :   96   Min.   :0.00000
M:14306   Asian                                     :  791   1st Qu.:0.00000
          Black or African American                 : 8758   Median :0.00000
          Native Hawaiian or Other Pacific Islander:  115   Mean   :0.06249
          Unknown                                   : 7476   3rd Qu.:0.00000
          White                                     :14225   Max.   :1.00000
       MI                CHF              cerebro           dementia
Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
Median :0.00000   Median :0.00000   Median :0.00000   Median :0.00000
Mean   :0.04069   Mean   :0.07301   Mean   :0.06109   Mean   :0.03277
3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
      CPD               RD               PUD            mildliver
Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
Median :0.0000   Median :0.00000   Median :0.00000   Median :0.00000
Mean   :0.1752   Mean   :0.02165   Mean   :0.01373   Mean   :0.04758
3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
```

```
diabetes          hemipara          renal          modsevliv
Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.000000
1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000000
Median :0.0000   Median :0.00000   Median :0.00000   Median :0.000000
Mean   :0.1497   Mean   :0.01338   Mean   :0.08693   Mean   :0.004386
3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.000000
Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.000000
metastaticcancer      aids              PVD               died
Min.   :0.00000   Min.   :0.000000   Min.   :0.00000   Min.   :0.00000
1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.00000   1st Qu.:0.00000
Median :0.00000   Median :0.000000   Median :0.00000   Median :0.00000
Mean   :0.01217   Mean   :0.007183   Mean   :0.05089   Mean   :0.04119
3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.00000   3rd Qu.:0.00000
Max.   :1.00000   Max.   :1.000000   Max.   :1.00000   Max.   :1.00000
    agecat
Min.   :0.000
1st Qu.:0.000
Median :1.000
Mean   :1.057
3rd Qu.:2.000
Max.   :4.000
```

Natural questions

1. Who are most susceptible to dying from COVID-19?
2. Can we build a model to predict death from COVID-19?
3. Which variables are most predictive of death?

Cross-tabs of died vs other variables

```
agecat
        0     1     2     3     4
  0 15474  5710  4558  2616  1807
  1   104   145   285   399   363

sex
        F     M
  0 16623 13542
  1   532   764
```

```
race
    American Indian or Alaska Native   Asian   Black or African American
  0                                96     775                       8252
  1                                 0      16                        506

    Native Hawaiian or Other Pacific Islander   Unknown   White
  0                                       106      7340   13596
  1                                         9       136     629
```

```
aids
      0     1
0 29956   209
1  1279    17

diabetes
      0     1
0 25875  4290
1   876   420

metastatic
      0     1
0 29833   332
1  1245    51

MI
      0     1
0 29147  1018
1  1034   262

cancer
      0     1
0 28391  1774
1  1104   192

hemipara
      0     1
0 29783   382
1  1257    39

renal
      0     1
0 27916  2249
1   810   486

PUD
      0     1
0 29770   395
1  1259    37

cerebro
      0     1
0 28497  1668
1  1042   254

mildliver
      0     1
0 28789  1376
1  1175   121

CHF
      0     1
0 28267  1898
1   897   399

PVD
      0     1
0 28800  1365
1  1060   236

dementia
      0     1
0 29333   832
1  1097   199

modsevliv
      0     1
0 30049   116
1  1274    22

CPD
      0     1
0 25043  5122
1   905   391

RD
      0     1
0 29538   627
1  1242    54
```

Ordinary logistic regression

- Given $X_1 = x_1, X_2 = x_2, \ldots,$ Y is Bernoulli with $p = P(Y = 1)$ given by

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$$

- Solving for p gives

$$p = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots)}$$

Two problems for ordinary logistic regression

1. Charlson index is linearly dependent on comorbidities
2. No deaths among American Indian or Alaska Native subjects in data

	Alive	Died
American Indian or Alaska Native	96	0
Asian	775	16
Black or African American	8252	506
Native Hawaiian or Other Pacific Islander	106	9
White	13596	629
Unknown	7340	136

Ordinary logistic regression without charlson and American Indian or Alaska Native

	Estimate	Pr(>|z|)	
(Intercept)	-5.67032	< 2e-16	***
sexM	0.56641	< 2e-16	***
raceBlack or African American	0.93508	0.000455	***
raceNative Hawaiian or Other Pacific Islander	1.84405	4.43e-05	***
raceUnknown	0.20332	0.462623	
raceWhite	0.52408	0.048481	*
cancer	-0.13146	0.191646	
MI	0.68390	1.26e-13	***
CHF	0.36755	1.21e-05	***
cerebro	0.05860	0.516509	
dementia	0.26049	0.007378	**
CPD	0.21867	0.002579	**
RD	0.15874	0.328141	
PUD	-0.30948	0.111342	

Ordinary logistic regression results (cont'd.)

	Estimate	Pr(>|z|)	
mildliver	0.25547	0.030074	*
diabetes	0.11679	0.106026	
hemipara	-0.28824	0.128170	
renal	0.77121	< 2e-16	***
modsevliv	1.01456	0.000216	***
metastaticcancer	0.54953	0.002871	**
aids	0.52133	0.059725	.
PVD	-0.11793	0.205512	
agecat	0.69127	< 2e-16	***

Calculating probability estimates

- Most at risk subject is 80–90 year-old male Native Hawaiian or other Pacific Islander with all comorbidities except `cancer`, `PUD`, `hemipara`, and `PVD`

$$
\begin{aligned}
x &= -5.67032 + 0.56641 + 1.84405 + 0.68390 + 0.36755 + 0.05860 \\
&\quad + 0.26049 + 0.21867 + 0.15874 + 0.11679 + 0.77121 + 1.01456 \\
&\quad + 0.54953 + 0.52133 + 0.69127 \times 4 \\
&= 4.22659 \\
P(Y = 1) &= \frac{\exp(x)}{1 + \exp(x)} = 0.9856
\end{aligned}
$$

- There is no such subject in the data

Calculating probability estimates (cont'd.)

- `modsevliv` (moderate/severe liver disease) has largest coefficient
- Probability for 80–90 year-old male Native Hawaiian or other Pacific Islander with moderate/severe liver disease and **no other comorbidity**:

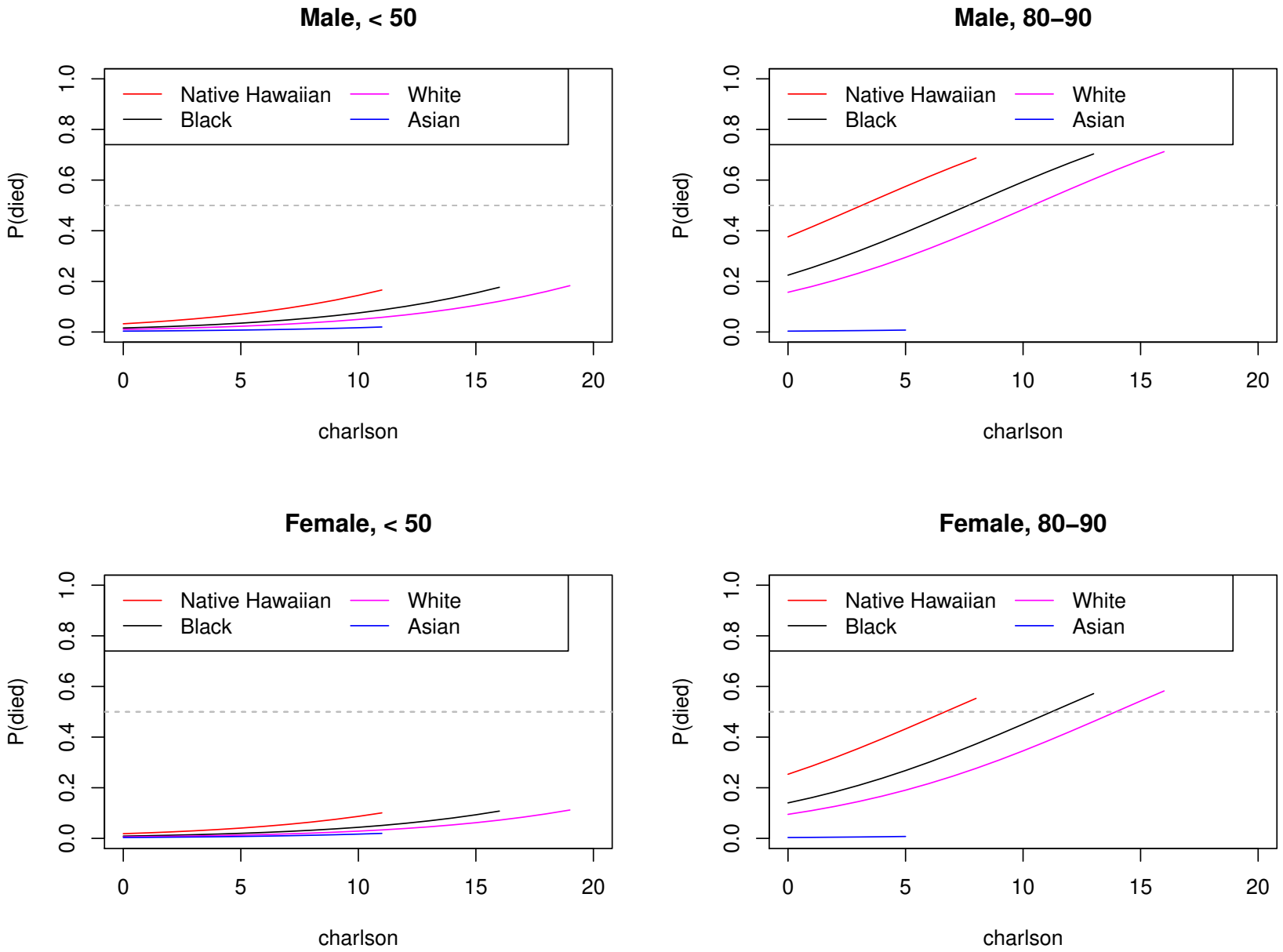
$$\begin{aligned} x &= -5.67032 + 0.56641 + 1.84405 + 1.01456 + 0.69127 \times 4 \\ &= 0.51978 \\ P(Y=1) &= \frac{\exp(x)}{1+\exp(x)} = 0.6271 \end{aligned}$$

- There is no such subject in the data

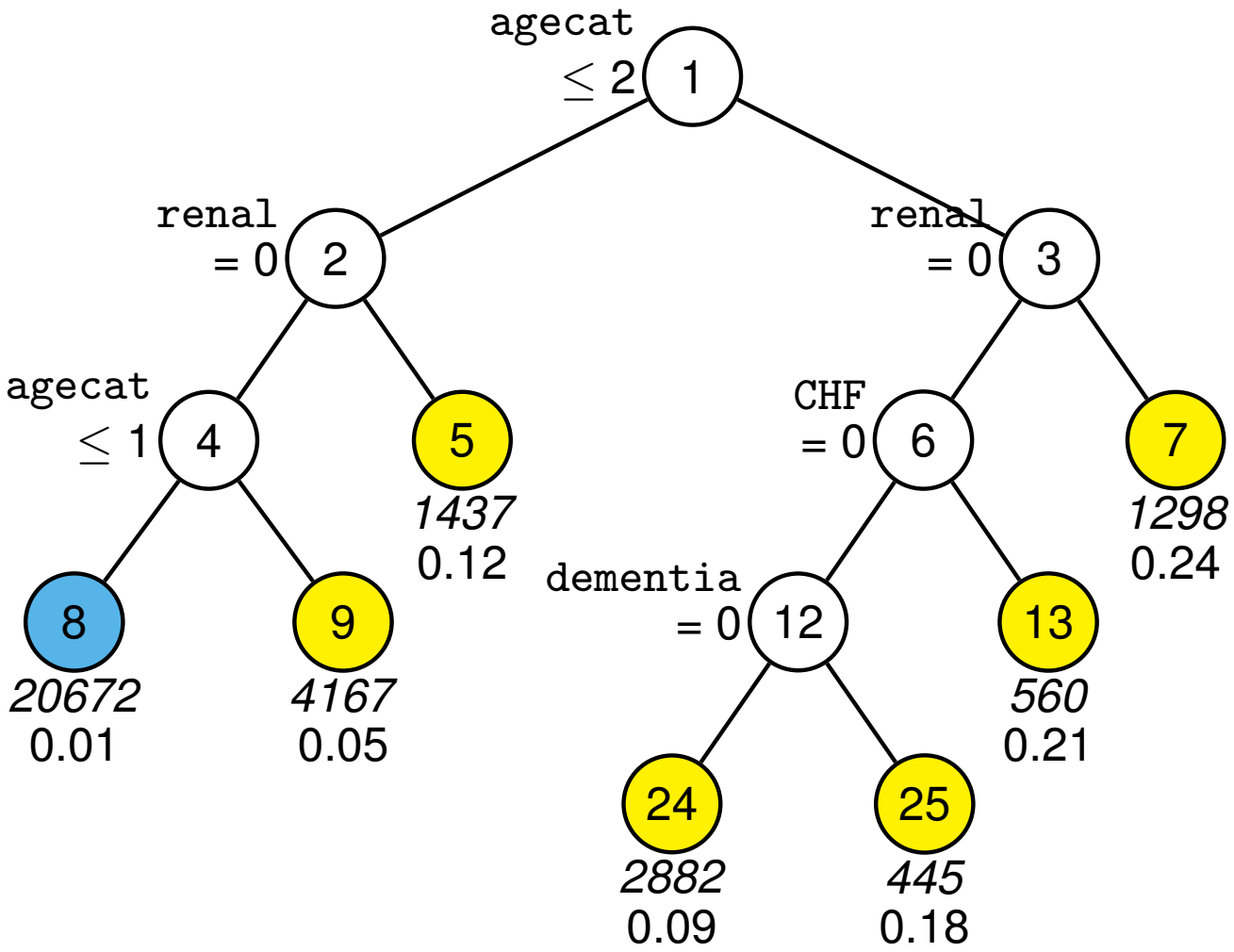
Logistic regression with charlson but without comorbidities and Amer. Indian & Alaska Native

	Estimate	Pr(>|z|)	
(Intercept)	-5.681569	< 2e-16	***
sexM	0.574101	< 2e-16	***
raceBlack or African American	0.979740	0.000214	***
raceNative Hawaiian or Other Pac. Islander	1.712625	0.000158	***
raceUnknown	0.219208	0.425239	
raceWhite	0.538004	0.041245	*
agecat	0.721921	< 2e-16	***
charlson	0.161760	< 2e-16	***

Fitted curves from logistic regression

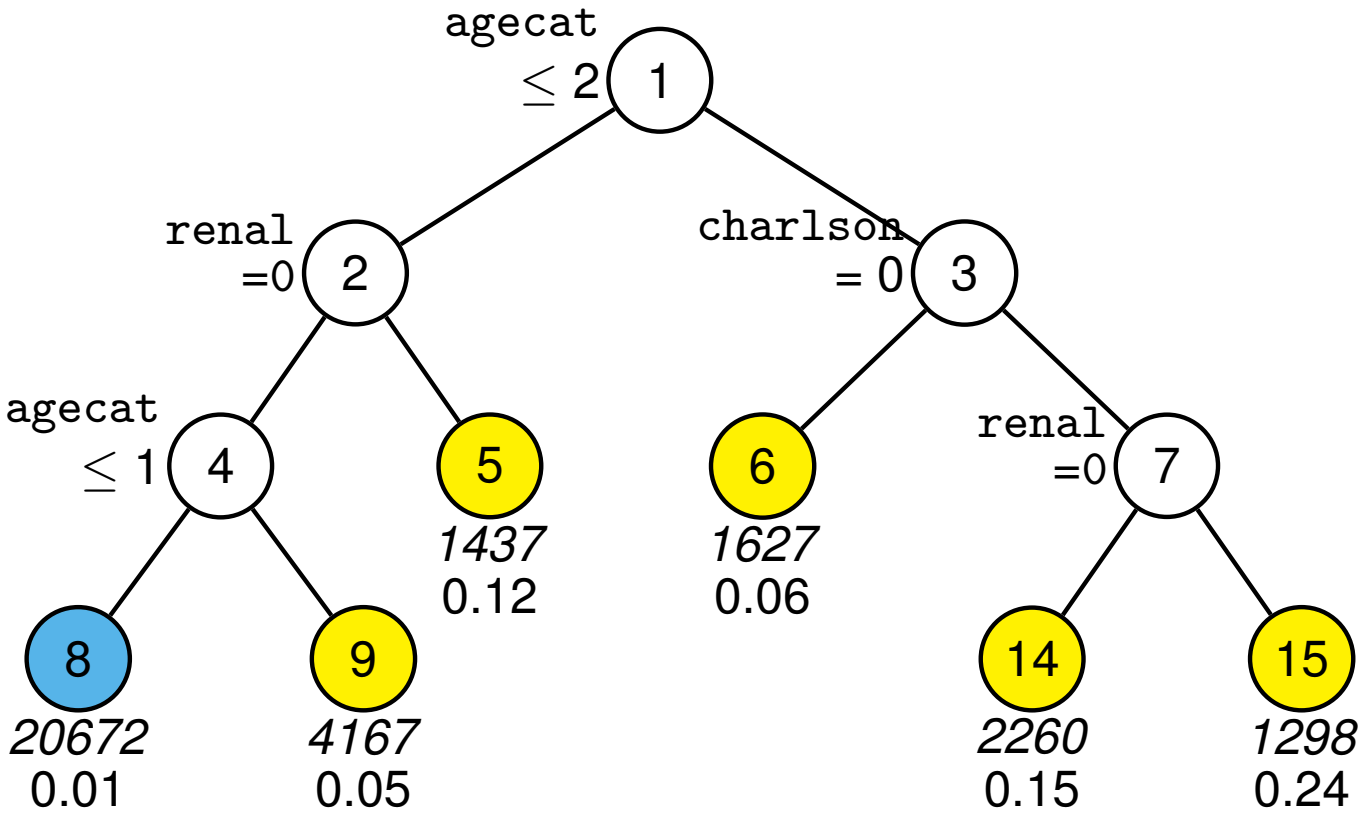


Regression tree (31461 obs, without charlson)



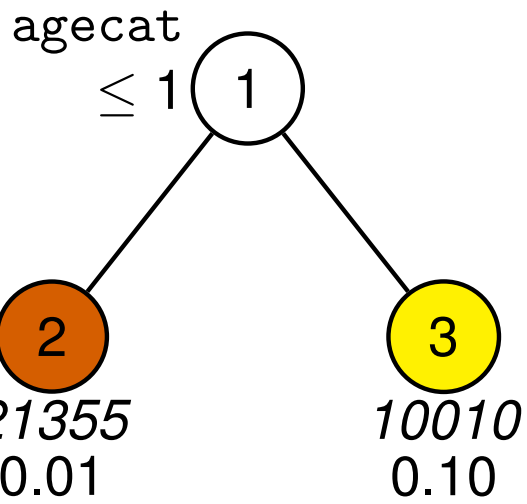
- Sample size (*in italics*) and mortality rate printed below nodes
- Terminal nodes with mortality rates above and below value of 0.04 at root node are colored yellow and skyblue, respectively

Regression tree (31461 obs, with charlson)



- Sample size (*in italics*) and mortality rate printed below nodes
- Terminal nodes with mortality rates above and below value of 0.04 at root node are colored yellow and skyblue, respectively

Logistic regression tree (without charlson and American Indian & Alaska Native)

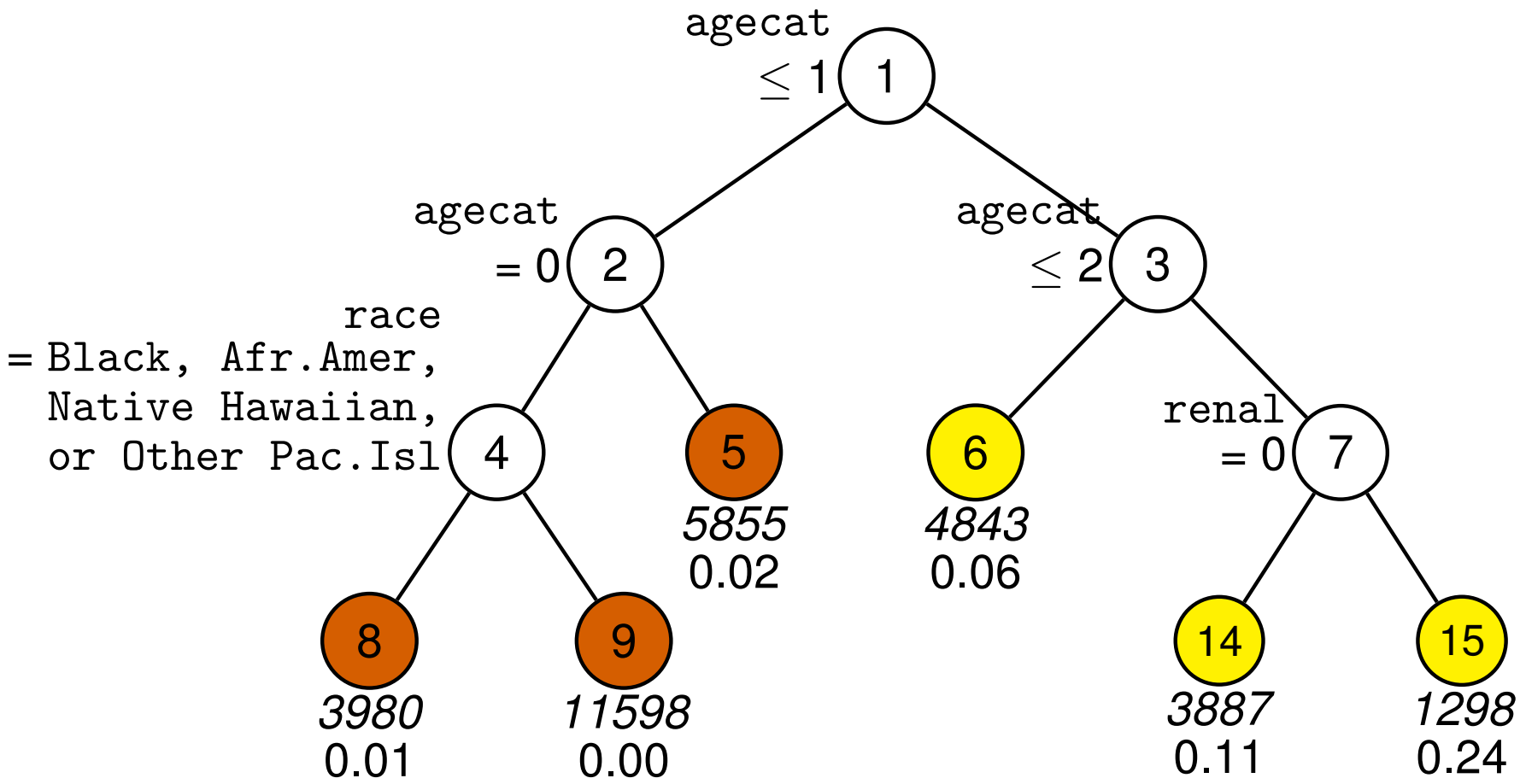


- At each split, an observation goes to the left branch if and only if the condition is satisfied
- Sample size (*in italics*) and estimated probability of death below nodes
- Logistic regression model fitted to each node

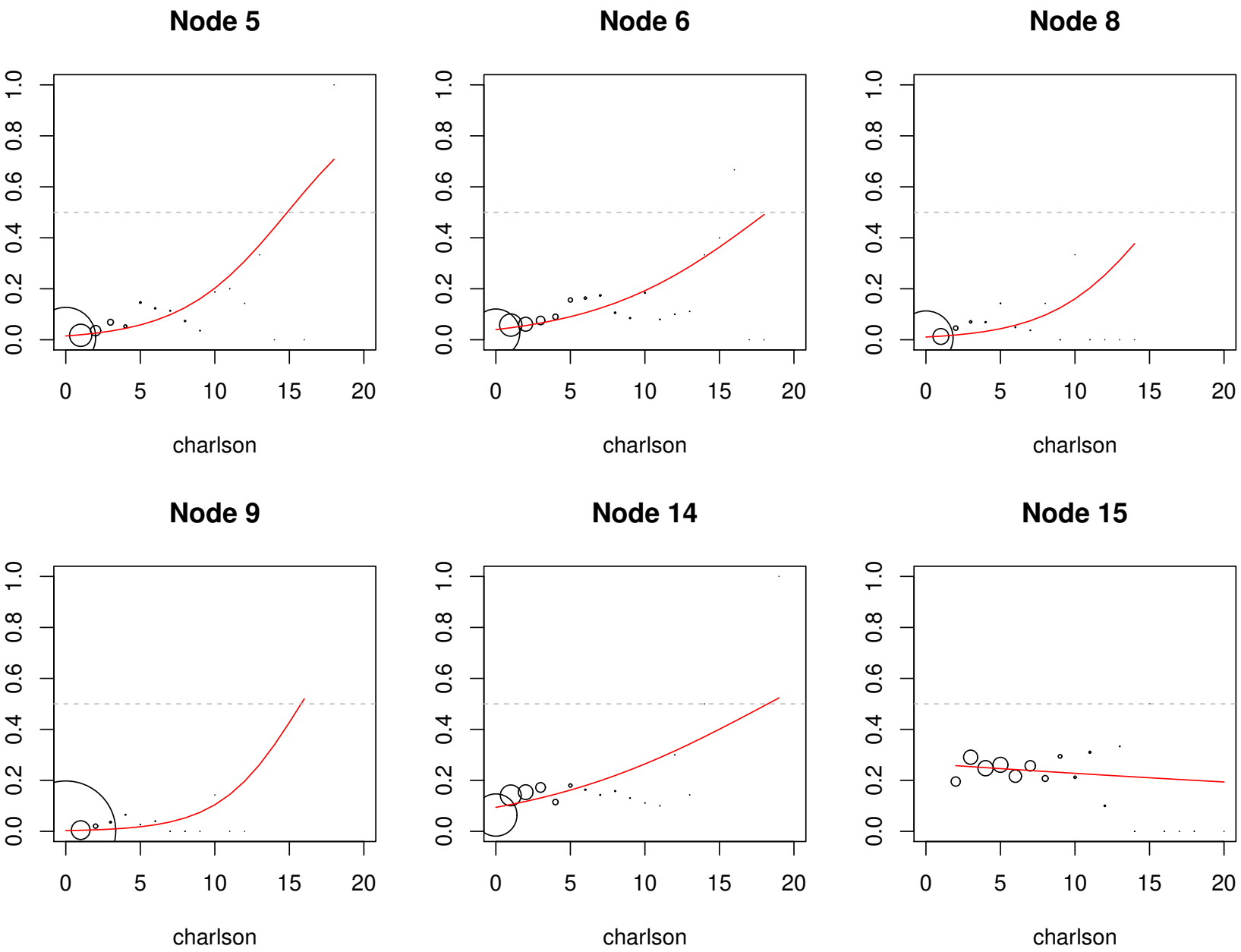
	agecat ≤ 1		agecat > 1	
	Coef	P-value	Coef	P-value
(Intercept)	-6.626	0.000	-4.760	0.000
renal	1.192	0.000	0.686	0.000
agecat	0.828	0.248	0.492	0.000
CHF	0.470	0.034	0.357	0.000
MI	0.940	0.000	0.608	0.000
PVD	0.128	0.630	-0.103	0.288
cerebro	0.101	0.709	0.071	0.447
dementia	0.550	0.425	0.339	0.001
diabetes	0.233	0.166	0.036	0.653
cancer	0.112	0.713	-0.129	0.219
CPD	0.293	0.069	0.192	0.017
mildliver	0.605	0.005	-0.009	0.949
modsevliv	1.488	0.000	0.546	0.135

	agecat ≤ 1		agecat > 1	
	Coef	P-value	Coef	P-value
sex.M	0.817	0.000	0.463	0.000
race.Black or African American	1.478	0.036	0.733	0.011
race.Native Hawaiian or Other Pacific	1.248	0.310	2.020	0.000
race.Unknown	0.398	0.580	0.204	0.496
race.White	0.727	0.305	0.457	0.110
metastatic	0.881	0.080	0.479	0.014
hemipara	0.523	0.211	-0.513	0.016
RD	0.135	0.719	0.115	0.514
PUD	-0.088	0.835	-0.396	0.067
aids	0.527	0.174	0.212	0.599

Logistic regression tree (31,461 obs, with charlson as sole linear predictor)



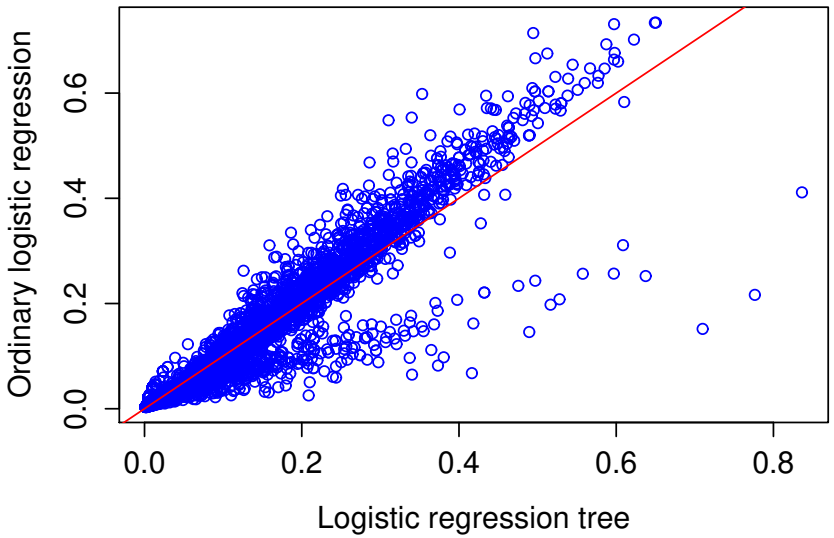
Fitted logistic curves in nodes of tree

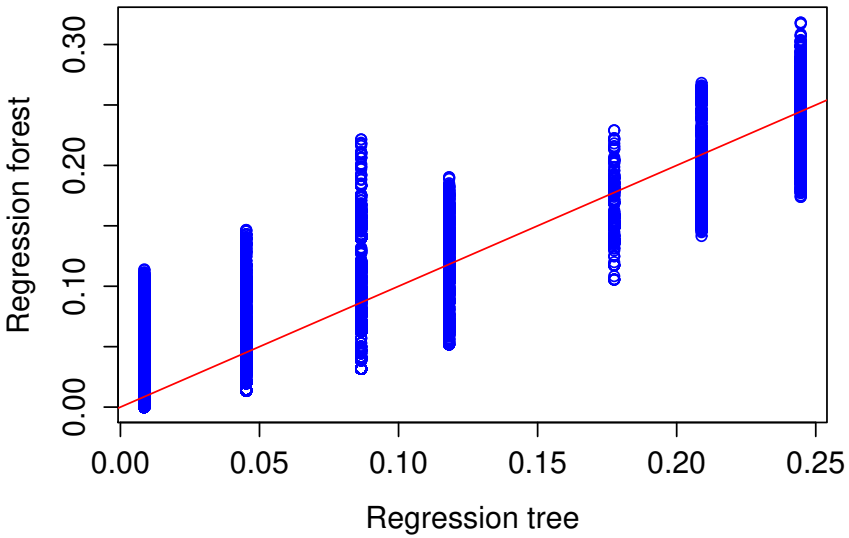


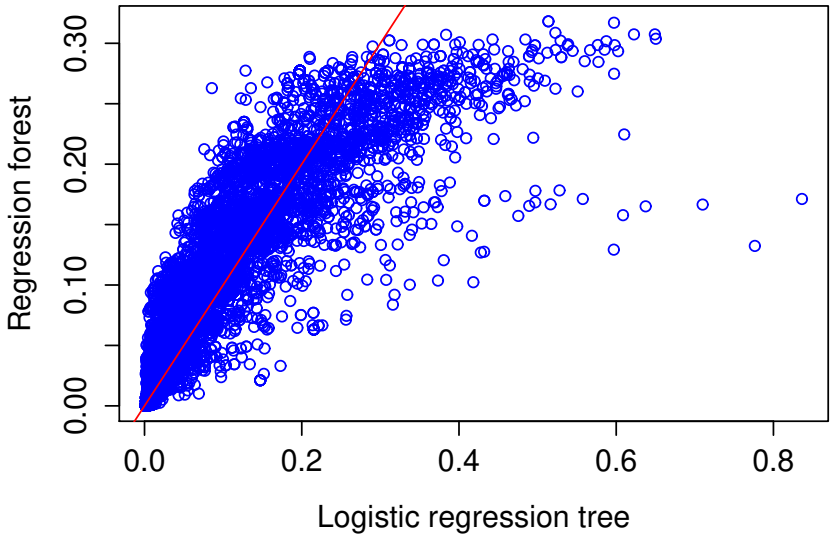
GUIDE models

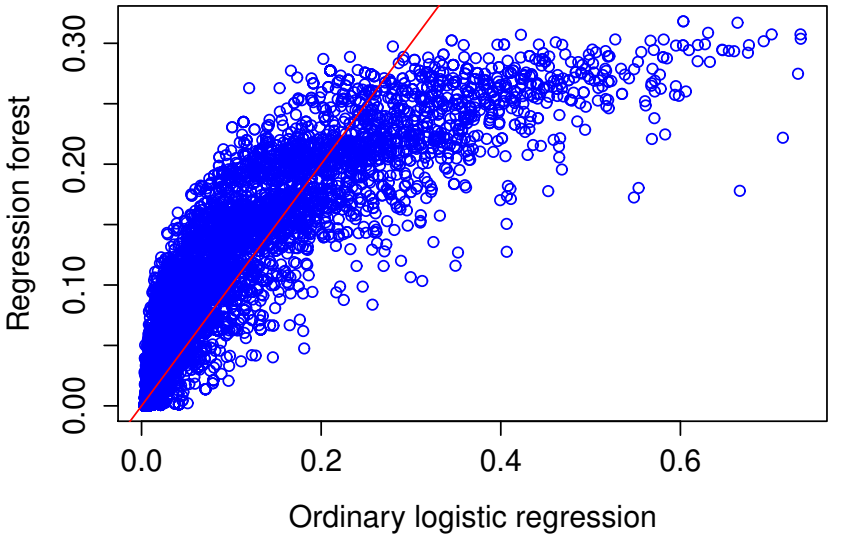
- Regression tree with/without `American Indian & Alaska Native` and with/without `charlson`
- Regression forest with/without `American Indian & Alaska Native` and and with/without `charlson`
- Logistic regression tree without `American Indian & Alaska Native` and with/without `charlson`

P(died) w/o charlson & Am. Indian & Alaska Nat.









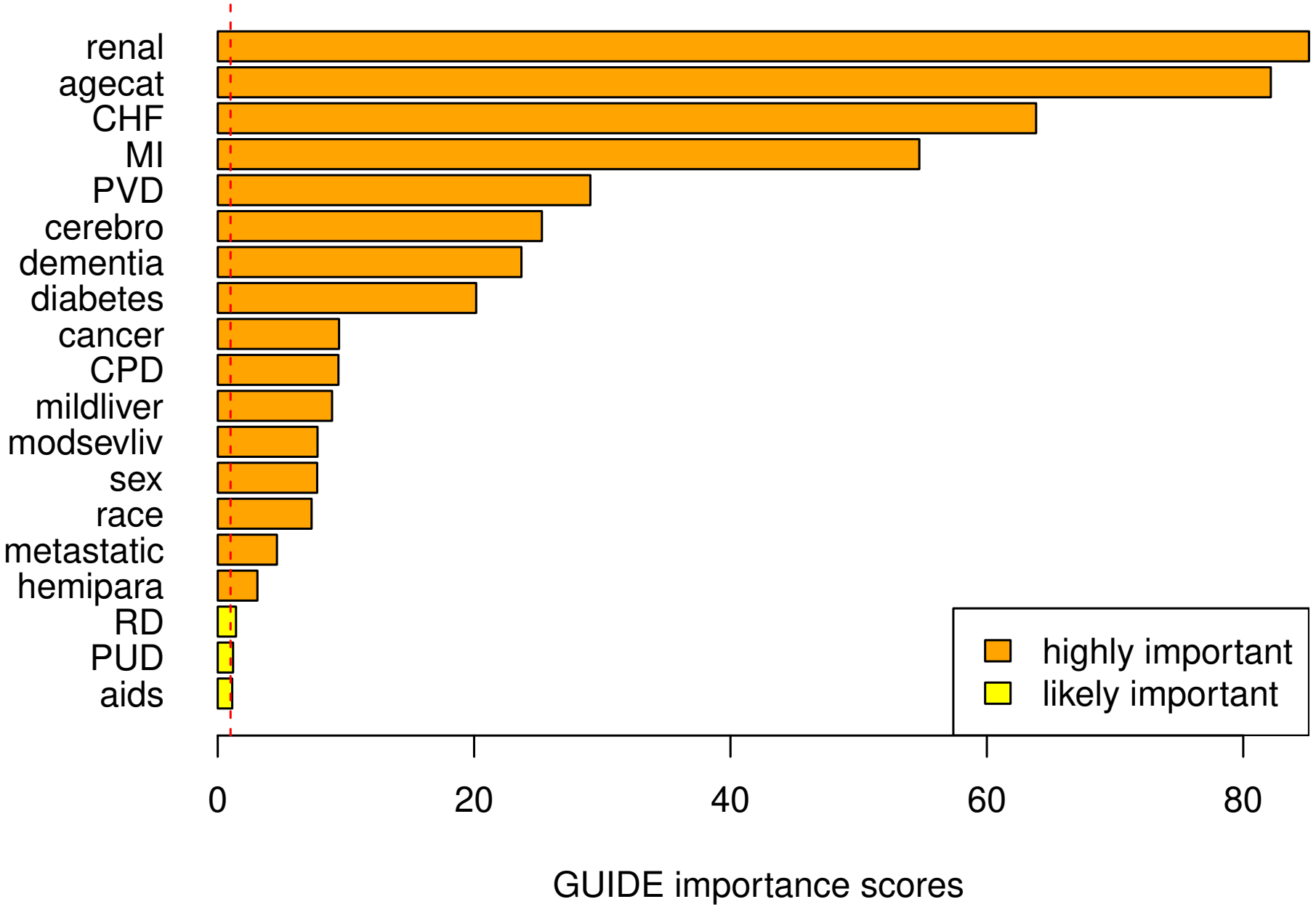
Accuracy vs interpretability

Most accurate		Most interpretable
Classification or regression forest	Logistic regression tree	Piecewise constant tree

Logistic models

- Ordinary logistic regression does not allow `charlson` and `American Indian & Alaska Native`
- Logistic regression tree does not allow `American Indian & Alaska Native`

GUIDE importance scores



About GUIDE

- GUIDE algorithm and software have been in development for 30+ years
- GUIDE manual and free compiled code for Linux, Mac OS X and Windows are available at `www.stat.wisc.edu/~loh/guide.html`
- GUIDE is not implemented in R but can be used in R (see manual)
- Key references: Loh and Vanichsetakul (1988), Chaudhuri et al. (1994, 1995), Loh and Shih (1997), Kim and Loh (2001), Loh (2002, 2009, 2014, 2019), Loh and Zheng (2013), and Loh et al. (2015, 2016, 2019a,b)

Things to do

1. Go to `http://pages.stat.wisc.edu/~loh/guide.html` to download the GUIDE datasets, manual and software:
   - `http://www.stat.wisc.edu/~loh/treeprogs/guide/datafiles.zip`
   - `http://www.stat.wisc.edu/~loh/treeprogs/guide/guideman.pdf`
2. Follow the instructions in the GUIDE manual to install the GUIDE and LaTeX software
3. For a brief introduction to classification and regression trees, see Loh (2011), Classification and regression trees, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol.1, 14–23
   `http://www.stat.wisc.edu/~loh/treeprogs/guide/wires11.pdf`

Useful commands to use in terminal windows

Mac OSX

`https://www.taniarascia.com/how-to-use-the-command-line-for-apple-macos-and-linux/`

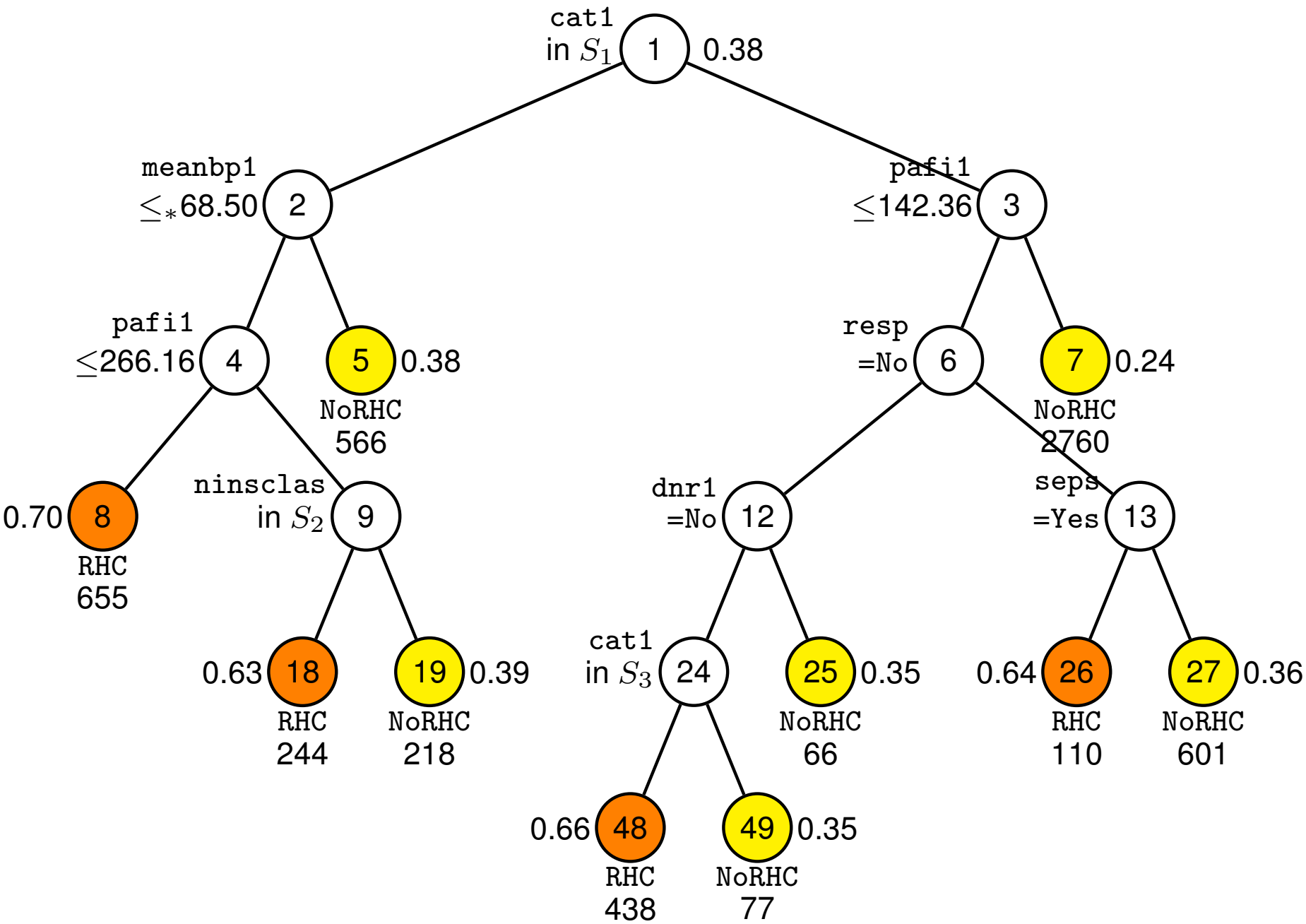
Windows

`https://www.computerhope.com/overview.htm`

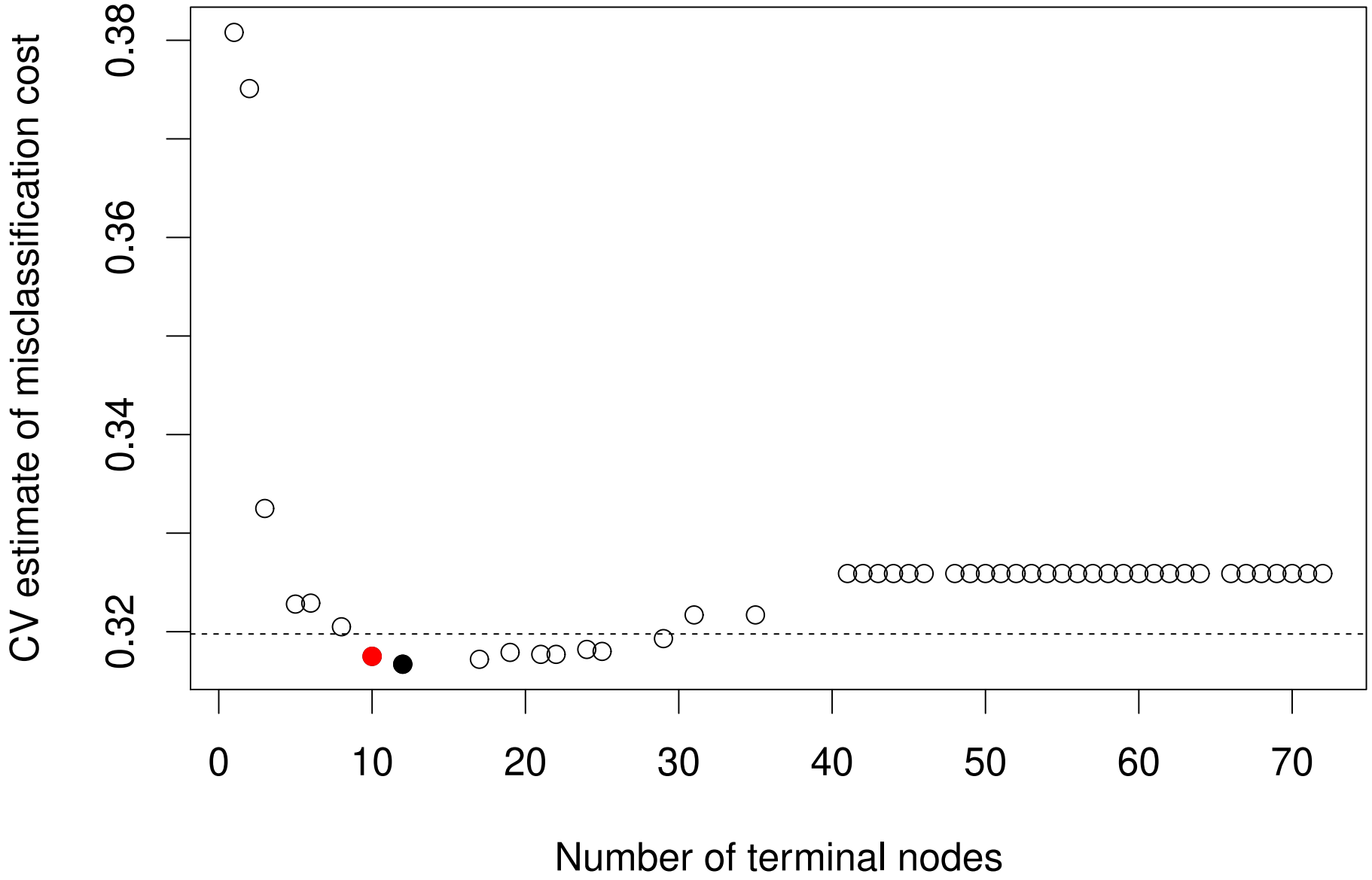
For information on opening a Command window:

`https://www.computerhope.com/issues/chusedos.htm`

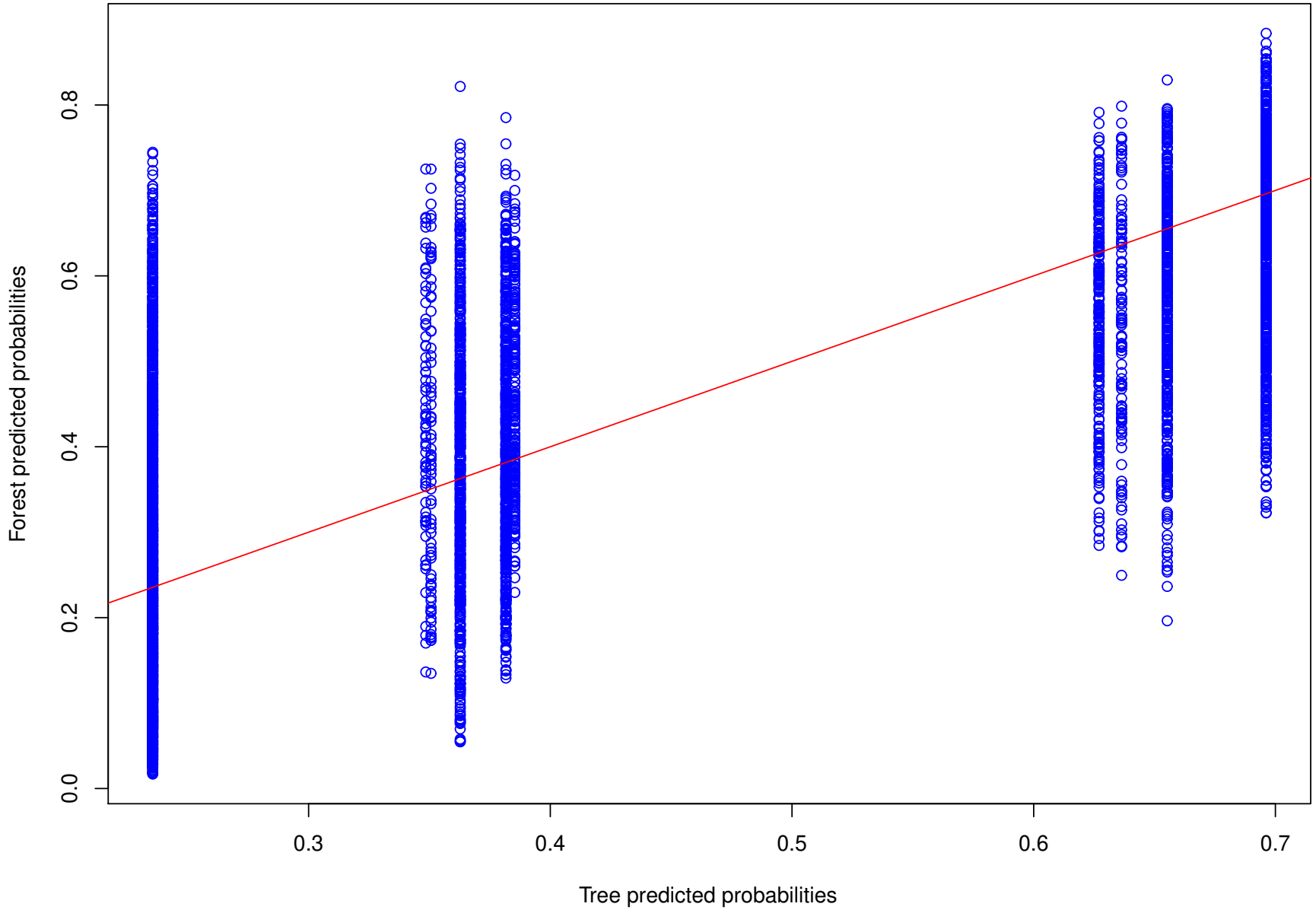
RHC classification tree



CV estimates of misclassification cost



RHC predicted probabilities of forest vs tree

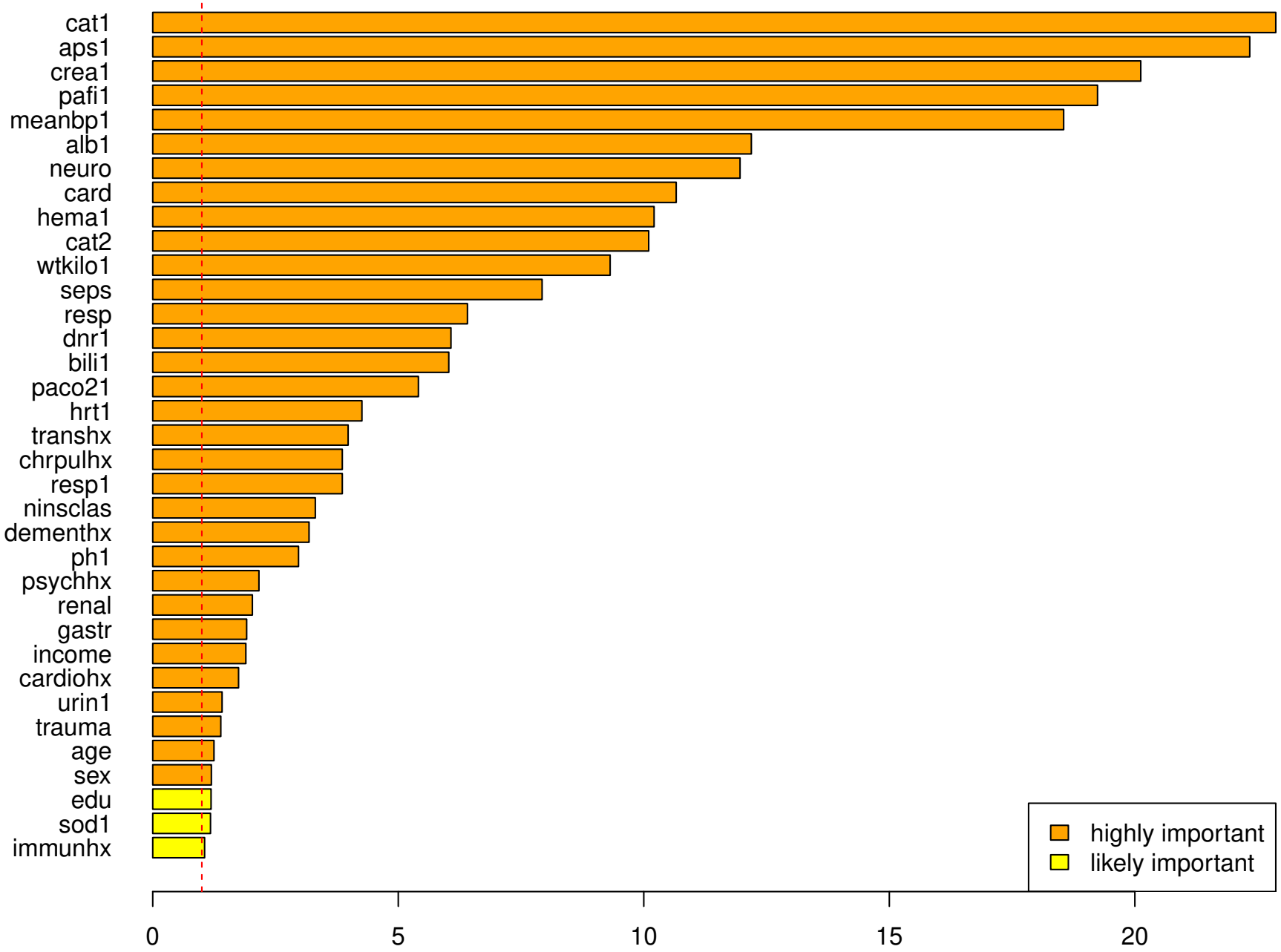


R code for plotting predicted probabilities

```
tree <- read.table("classpred.txt",header=TRUE)
forest <- read.table("forestpred.txt",header=TRUE)
tree.p <- tree[,6]
forest.p <- forest[,3]

plot(forest.p ~ tree.p, xlab="Tree predicted probabilities",
                ylab="Forest predicted probabilities",col="blue")
abline(c(0,1),col="red")
```

Scores of important RHC variables



R code for plotting importance scores

```
par(las=1,mar=c(5,12,4,2),cex=1)
leg.col <- c("orange","yellow")
leg.txt <- c("highly important","likely important")
x <- read.table("imp.scr",header=TRUE)
score <- x$Score
vars <- x$Variable
type <- x$Type
barcol <- rep("orange",length(vars))
barcol[type == "L"] <- "yellow"
barcol[type == "U"] <- "cyan"
n <- sum(x$Type != "U")
barplot(rev(score[1:n]),names.arg=rev(vars[1:n]),
    col=rev(barcol[1:n]),horiz=TRUE,xlab="GUIDE importance scores")
abline(v=1,col="red",lty=2)
legend("bottomright",legend=leg.txt,fill=leg.col)
```

Creating GUIDE description file in R

```
dat <- read.table("data.txt",header=TRUE)
nvar <- ncol(dat)
varnames <- names(dat)
roles <- rep("s",nvar)
c.vars <- c("cat1","cat2",...)
roles[varnames %in% c.vars] <- "c"
x.vars <- c("xvar1","xvar2",...)
roles[varnames %in% x.vars] <- "x"
d.var <- "dvariable"
roles[varnames %in% d.var] <- "d"
write("data.txt",file="desc.txt")
write("NA",file="desc.txt",append=TRUE)
write("2",file="desc.txt",append=TRUE)
write.table(cbind(1:nvar,varnames,roles),file="desc.txt",
    row.names=FALSE,col.names=FALSE,quote=FALSE,append=TRUE)
```

Homework 2 (RHC cont'd.)
due in Canvas by 9:30AM, Thu Feb 25, 2021

1. Build a logistic regression model to estimate P(RHC). State clearly how you deal with missing values and how you arrive at your logistic model (e.g., how and why you chose the terms, including interactions, if any). Give the R code and output for reproducing your logistic fit.
2. Reproduce the GUIDE importance scores plot on slide 55.
3. What does the logistic model say about the effects of the variables? How does it agree or disagree with the GUIDE tree and importance scores?
4. Build a GUIDE forest model to estimate P(RHC) and plot the estimates versus those from logistic regression (see slide 53). Say which method is more accurate and why you think so.
5. Include in your report the input and output files for GUIDE forest and importance scores.

Right heart catheterization (RHC) data

- Doctors believe that direct measurement of cardiac function by right heart catheterization (RHC) for some critically ill patients yields better outcomes
- Benefit of RHC has not been demonstrated in a randomized clinical trial, because physicians refuse to allow their patients to be randomized
- In observational studies, relative risk of death is higher in elderly and patients with acute myocrdial infarction who received RHC
- In such studies, decision to use RHC is at discretion of physician
- Treatment selection is confounded with patient factors that are also related to outcomes, e.g., patients with low blood pressure are more likely to get RHC, and such patients are also more likely to die
- Data consist of observations on more than 60 variables for 5735 patients from 5 medical centers over 5 years (Connors et al., 1996)
- Response variables are `dth30` (death within 30 days) and `death` (death within 6 months)

Demographic and disease category variables

Name	Description	#missing
age	Age (18.04–101.85)	
sex	Sex (Female, Male)	
race	Race (black, white, other)	
edu	Years of education (0–30)	
income	Income bracket (Under \$11k, \$11–\$25k \$25–\$50k, $>$ \$50k)	
ninsclass	Medical insurance (Medicaid, Medicare, Medicare & Medicaid, Private, Private & Medicare, No insurance)	
cat1	Primary disease category (ARF, COPD, CHF, cirrhosis, coma, colon cancer, lung cancer, MOSF w/malignancy, MOSF w/sepsis)	
cat2	Secondary disease category (same categories as cat1)	4535

Admission diagnosis variables

Name	Description
ca	Cancer (No, Yes, Metastatic)
card	Cardiovascular (No, Yes)
gastr	Gastrointestinal (No, Yes)
hema	Hematologic (No, Yes)
meta	Metabolic (No, Yes)
neuro	Neurological (No, Yes)
ortho	Orthopedic (No, Yes)
renal	Renal (No, Yes)
resp	Respiratory (No, Yes)
seps	Sepsis (No, Yes)
trauma	Trauma (No, Yes)

Comorbidity illness indicator variables

Name	Description
amihx	Definite myocardial infarction
cardiohx	Acute MI, vascular disease, severe cardiovascular symptoms
chfhx	Congestive heart failure
chrpulhx	Pulmonary disease
dementhx	Dementia, stroke, Parkinson's
gibledhx	Upper GI bleeding
immunhx	Immunosuppression, organ transplant, HIV, diabetes, connective tissue disease
liverhx	Cirrhosis, hepatic failure
malighx	Solid tumor, metastatic disease, leukemia, myeloma, lymphoma
psychhx	Psychiatric history, psychosis, severe depression
renalhx	Renal disease

Day 1 variables

Name	Description	#missing
alb1	Albumin (0.3–29.0; normal is 3.4–5.4)	
aps1	Acute physiology component of APACHE III (3–147)	
bili1	Bilirubin (0.09999–58.19531; normal is 0.3–1.2)	
crea1	Creatinine (0.09999–25.09766; normal is 0.84–1.21)	
dnr1	Do-not-resuscitate status (No, Yes)	
hema1	Hematocrit (No, Yes)	
hrt1	Heart rate (8–250; normal is 60–100)	159
meanbp1	Mean blood pressure (10–259; normal is 70–100)	80
pafi1	PaO2/(0.01*FIO2) (11.6–937.5; normal is $\geq$ 400)	
paco21	PaCo2 (1–156; normal is 35–45)	

Day 1 variables (cont'd.)

Name	Description	#missing
ph1	PH (6.579–7.77; normal is 7.35–7.45)	
pot1	Potassium (1.2–11.898; normal is 3.6–5.2)	
resp1	Respiratory rate (2–100; normal is 12–16)	136
scoma1	Glasgow Coma Score (0–100)	
sod1	Sodium (101–178; normal is 135–145)	
temp1	Temperature (27–43; normal is 36.1–37.2)	
urin1	urine output (0–9000; normal is 800–2000)	3028
wblc1	White blood cell count (0–192; normal is 4.5–11)	
wtkilo1	Weight (19.5–244)	515

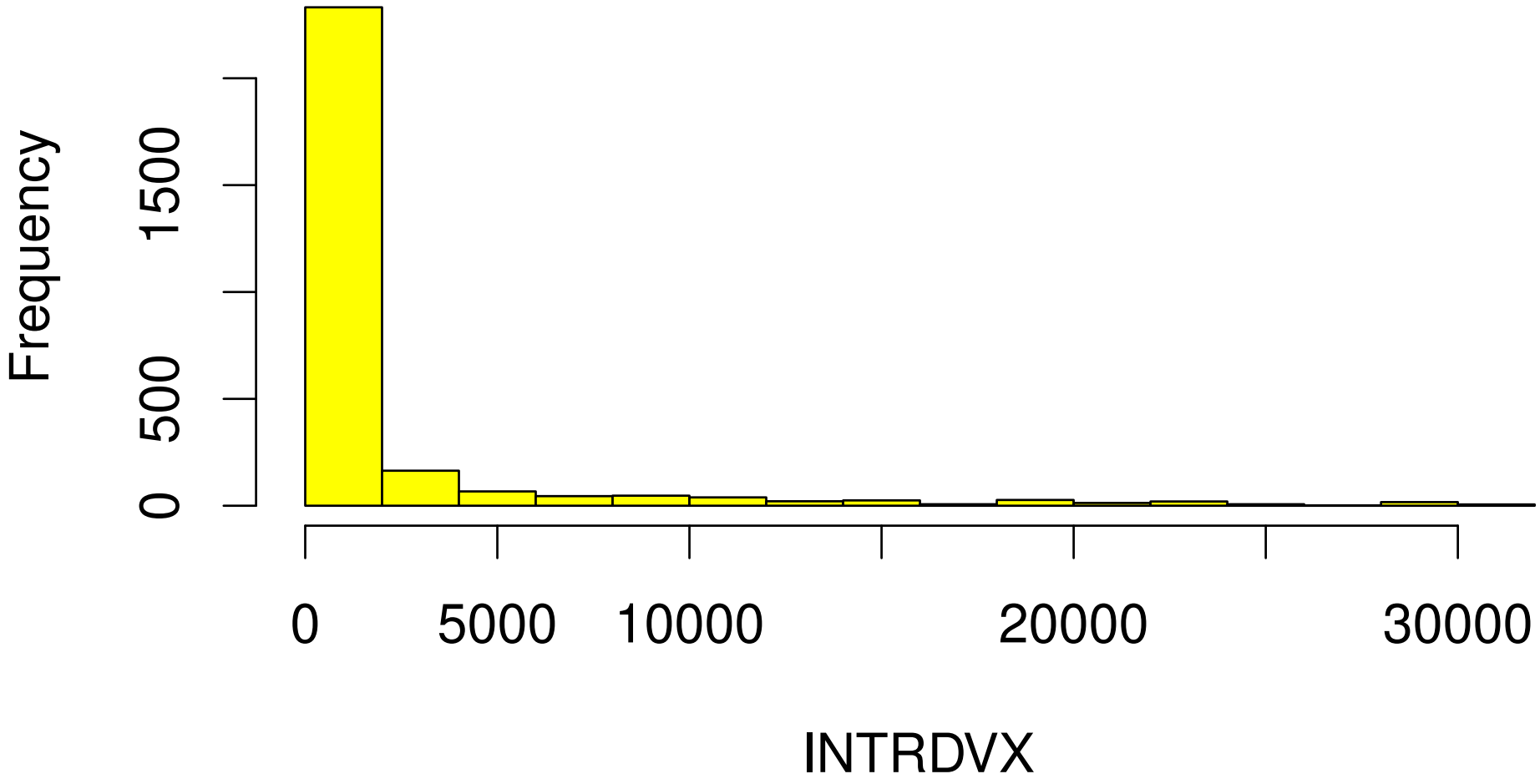
2013 Consumer Expenditure (CE) Survey Data

- 2013 Consumer Expenditure Survey, Bureau of Labor Statistics
- 25,822 consumer units (CUs) interviewed quarterly on hundreds of items
- Goal: estimate population mean interest and dividend (INTRDVX)
- About 21,000 CUs have valid nonresponse on INTRDVX
- Top 3% of INTRDVX are “topcoded” (above $32,000 changed to $98,338)
- Remaining 4609 CUs: 1771 missing and 2838 nonmissing INTRDVX
- 546 predictor variables
- 124 (20%) variables have missing values; 67 have more than 95% missing

Missing value flag codes

A	valid nonresponse: a response is not anticipated
B	invalid nonresponse
C	“don’t know”, refusal, or other type of nonresponse
D	valid data value
T	topcoding applied to value

Histogram of 2838 nonmissing INTRDVX values



Some variables and their proportions missing

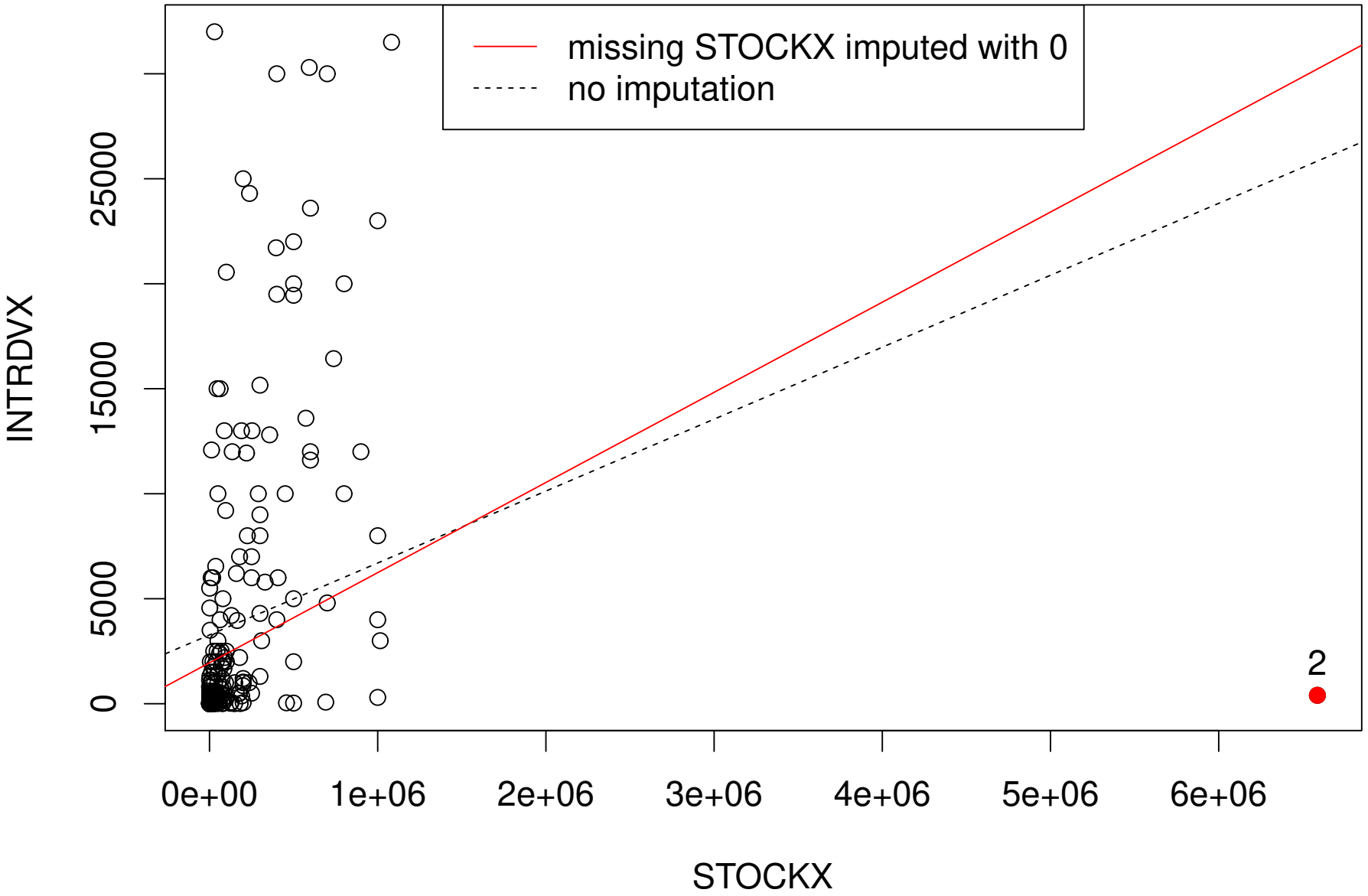
Name	Definition	Prop
AGE_REF	Age of reference person	
AGE2	Age of spouse	0.41
AS_COMP3	Number of males age 2 through 15 in CU	
BUILT	Year range property was built	0.13
CUTENURE	Housing tenure	
EDUCA2	Education of spouse	0.41
ERANKH	Percent expenditure outlay rank	0.08
FEDRFNDX	Federal income tax refund to all CU members	0.55
EARNCOMP	Composition of earners	
EOWNDWLP	Owned home outlays last quarter	
FFTAXOWE	Estimated Federal tax liabilities for entire CU	
FINCATAX	CU income after taxes in past 12 months	
FINCBTAX	CU income before taxes in past 12 months	

FINLWT21	Sampling weight	
FJSSDEDX	Estimated amount contributed to Social Security by all CU members past 12 mos.	
FRRETIRX	Social security and railroad retirement income	
FSALARYX	Wage and salary income of all members past 12 mos.	
FSTAXOWE	Estimated state tax owed	
GASMOPQ	Gasoline and motor oil last quarter	
HIGH_EDU	Highest level of education	
INC_HRS1	Number hours worked per week by reference person	0.30
INC_RANK	Income rank of CU to total population	0.08
INCLASS	Income class of CU based on income before taxes	
INCLASS2	Income class based on INC_RANK	
INCNONW1	Reason for not working during past 12 months	0.63

INCOMEY1	Employer paying most earnings in past 12 months	0.37
INCOMEY2	Employer from which spouse received most earnings during the past 12 months	0.61
IRAX	Total value of retirement accounts	0.84
LIQUIDB	Bracket range of bank accounts	0.97
LIQUIDX	Total value of checking, savings, CD, etc., accounts	0.83
LIQUIDX_	Flag variable for LIQUIDX	
LIQUDYRX	Total value of bank accounts one year ago	0.84
OCCUCOD1	Highest paid occupation last 12 months	0.37
OFSTPARK	Off street parking	0.25
PERINSPQ	Personal insurance and pensions last quarter	
PERSOT64	Number of persons over 64 in CU	
POV_CY	Is income below current year's poverty threshold?	0.08
POV_PY	Is income below previous year's poverty threshold?	0.08
PROPTXPQ	Property taxes last quarter	

PSU	Primary sampling unit	0.56
RENTEQVX	Monthly rent if home rented today	0.14
RESPSTAT	Completeness of income response	
RETSRVBX	Median value of bracket range for RETSURVB	0.99
RETSURVB	Range for amount received in retirement, survivor, or disability pensions during past 12 months	0.99
RETSURVX	Retirement, survivor, disability pensions past 12 mos.	0.76
SLOCTAXX	Total amount paid for state and local income taxes	0.87
STATE	State identifier	0.11
STOCKX	Value of directly-held stocks, bonds, mutual funds	0.94
STOCKYRX	Median value of bracket range of STOCKX	0.93
TOTXEST	Estimated total taxes paid	
UTILCQ	Utilities, fuels and public services this quarter	

Plot of 6% of data nonmissing STOCKX



Standard approaches to mean estimation

- Let n be the sample size and μ the population mean
- Let S be the sample subset of non-missing y_i and $\bar{S}$ be its complement
- Let $\hat{\pi}_i$ be the estimated probability that y_i is nonmissing
- Let w_i be the sampling weight

Weighting. The *inverse probability weighted* (IPW) estimate of μ is

$$\Big(\sum_{i\in S} w_i/\hat{\pi}_i\Big)^{-1}\sum_{i\in S} w_i y_i/\hat{\pi}_i$$

Imputation. Let $\hat{y}_j$ be the imputed value of y_j in $\bar{S}$. The estimate of μ is

$$(\sum_i w_i)^{-1}\Big(\sum_{i\in S} w_i y_i + \sum_{j\in\bar{S}} w_j\hat{y}_j\Big)$$

Imputation methods

Hot deck. Impute missing values by random sampling of non-missing values within 'adjustment cells'

Maximum likelihood. Sample from fitted posterior distribution. AMELIA (Honaker et al., 2011) uses multivariate normal model and EM algorithm.

Sequential regression. Initialize missing values with means and modes. Then iteratively fit a parametric regression model to one variable at a time, updating missing values with predicted values.

MICE (van Buuren and Groothuis-Oudshoorn, 2011) uses multiple linear regression for ordinal variables and multinomial logistic regression for categorical variables.

Practical issues with imputation

Hot deck. Adjustment cells often obtained by logistic regression, but their number can be excessive — one Current Population Survey used 11,520 cells from 7 predictor variables (Bollinger and Hirsch, 2006). Logistic regression requires all X variables to be nonmissing.
Circular logic: impute missing X values so that logistic regression can be used by hot deck to impute the same missing values!

Maximum likelihood and EM. Requires specification of likelihood

1. consequences of incorrect specification unknown
2. multivariate normality often used but is slow for high dimensional data
3. categorical variables converted to indicator variables

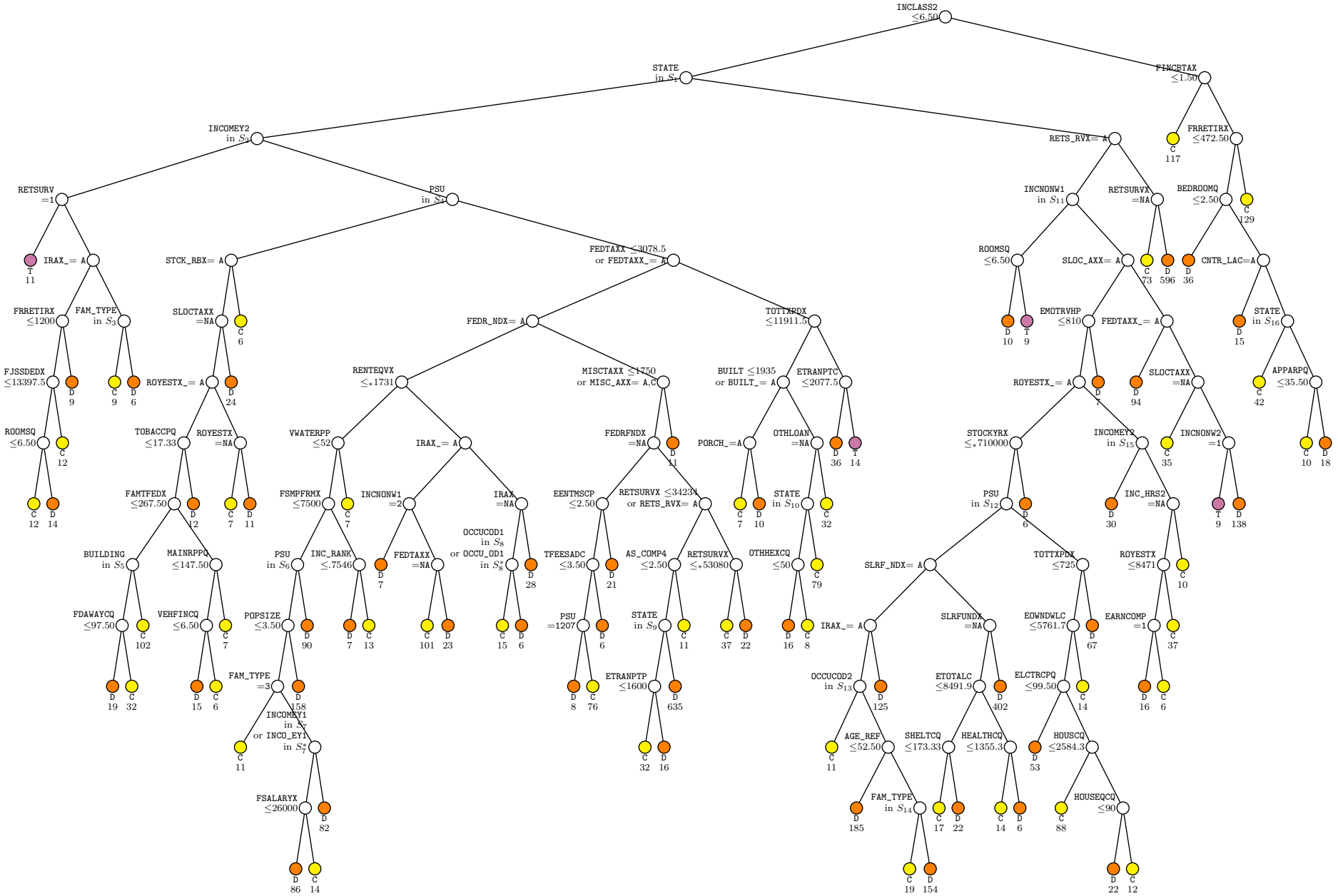
Sequential regression. Faster than MLE but:

1. linear regression limited by multicollinearity
2. logistic regression stopped by quasi-complete separation
3. variable relationships potentially violated (e.g., spouse age and education level imputed for unmarried)

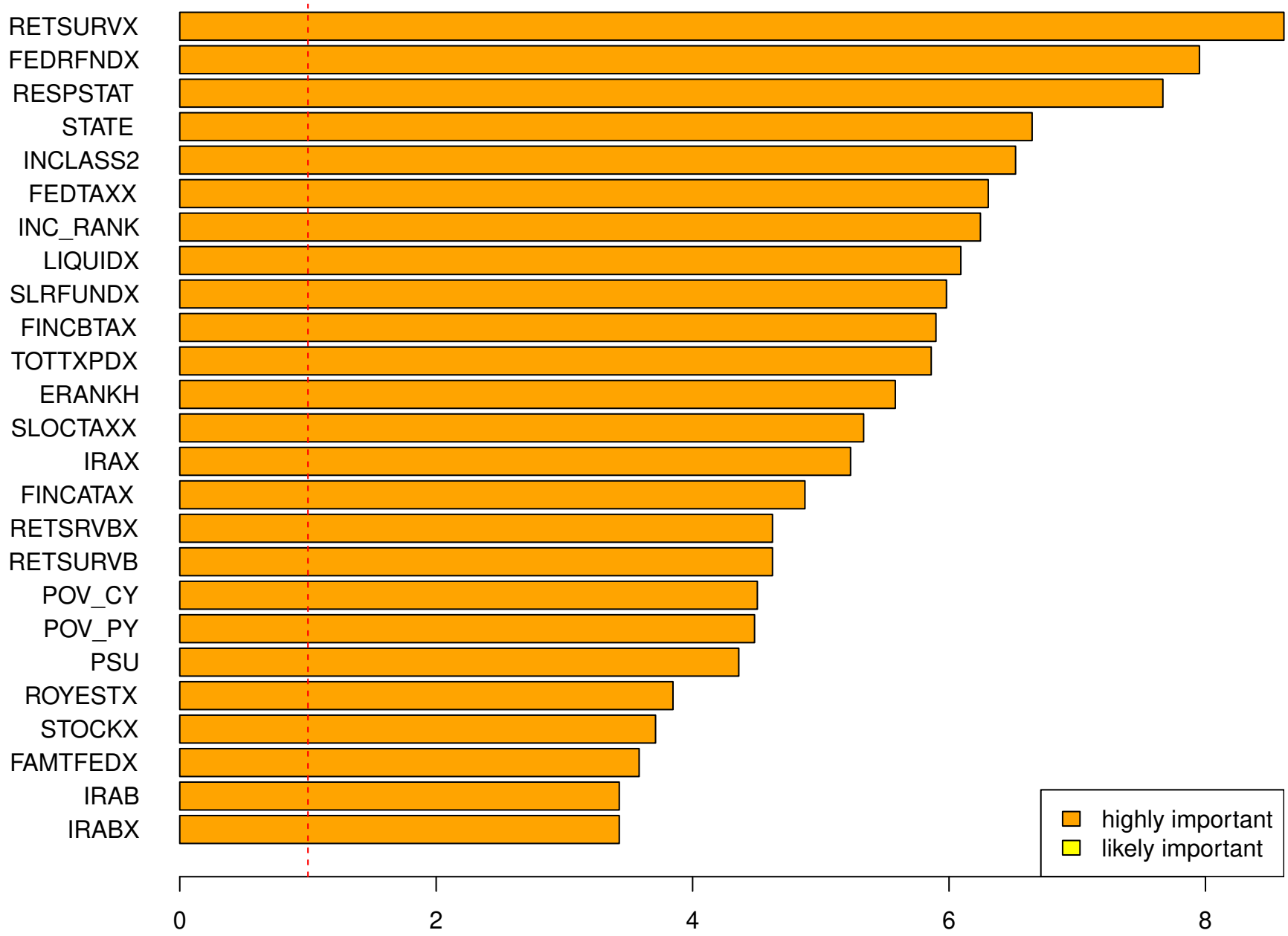
Contents of `ceclass.dsc` (M and W descriptors)

```
cedata.txt
NA
2
1 DIRACC C
2 DIRACC_ M
3 AGE_REF N
4 AGE_REF_ M
5 AGE2 N
6 AGE2_ M
:
50 FINLWT21 W
:
513 INTRDVX X
514 INTRDVX_ D
:
```

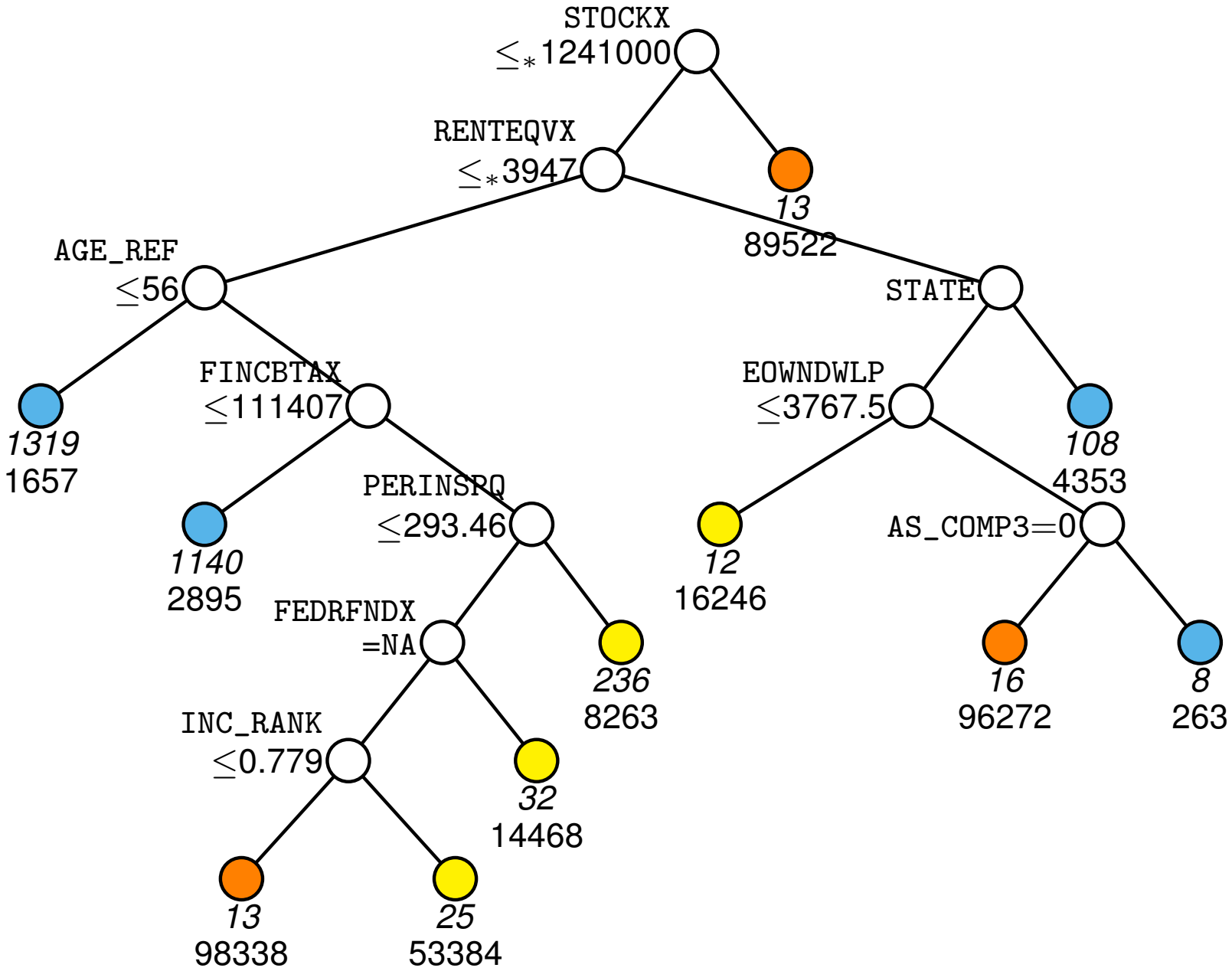
Classification tree for predicting INTRDVX_



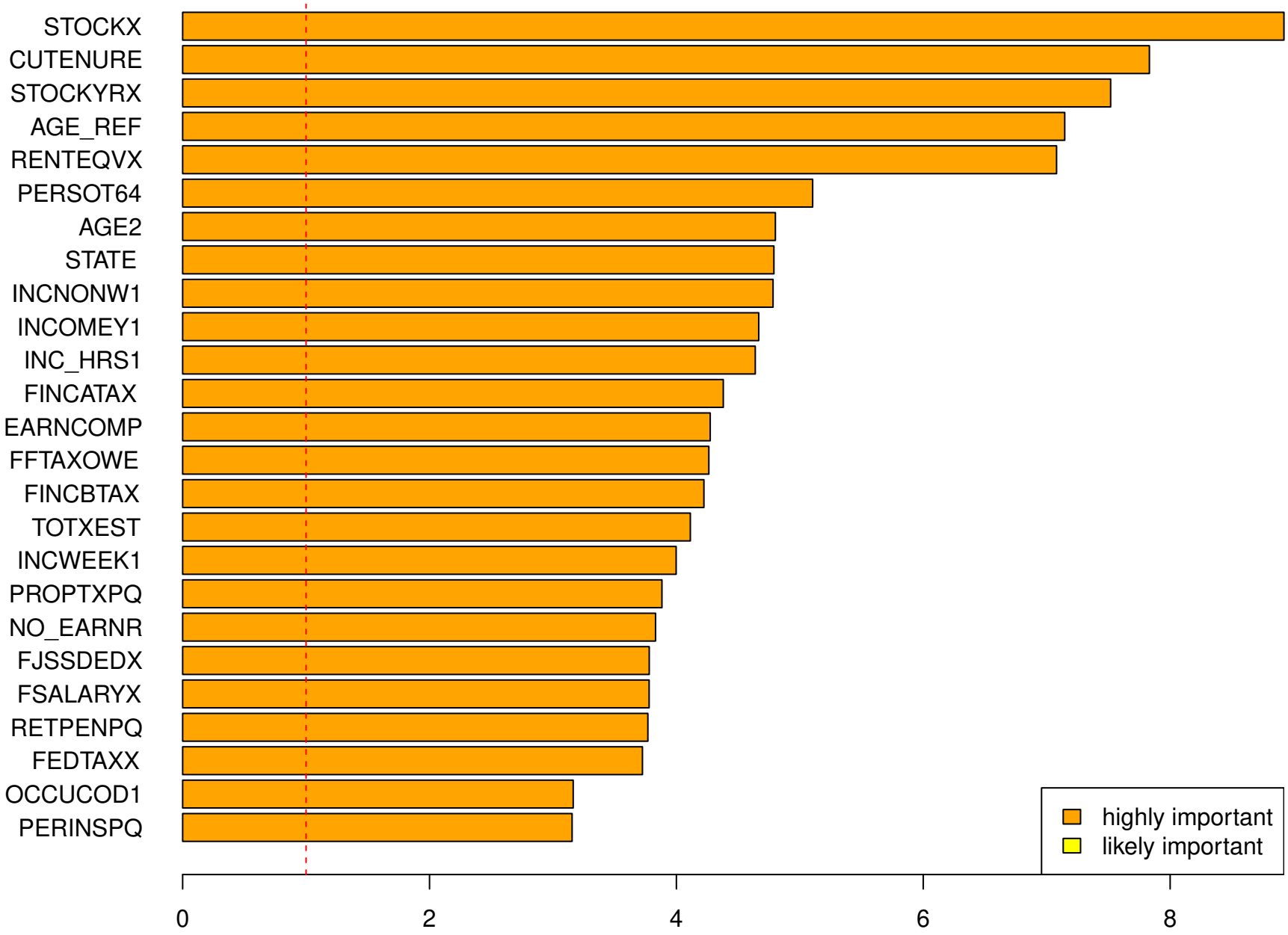
Top 25 predictors of INTRDVX_



Regression tree for predicting INTRDVX



Top 25 predictors of INTRDVX



Estimates of population mean INTRDVX

Simple method ignoring missing INTRDVX values	4697
IPW by GUIDE classification tree	4303
Weighted mean by GUIDE regression tree	4390

Contents of `ceclass.fit`

train	node	observed	predicted	"P(C)"	"P(D)"	"P(T)"
y	2689	"D"	"D"	0.21603E+00	0.78396E+00	0.38132E-05
y	19460	"D"	"D"	0.46510E+00	0.53490E+00	0.38132E-05
y	2691	"D"	"D"	0.13682E+00	0.86070E+00	0.24876E-02
y	2689	"D"	"D"	0.21603E+00	0.78396E+00	0.38132E-05
y	23	"D"	"D"	0.19631E+00	0.78691E+00	0.16779E-01
y	23	"D"	"D"	0.19631E+00	0.78691E+00	0.16779E-01
y	617	"C"	"D"	0.33334E+00	0.66665E+00	0.38132E-05
y	2433	"D"	"D"	0.23421E+00	0.76579E+00	0.38132E-05
:						

Contents of `cereg.fit`

train	node	observed	predicted
y	18	1.300000E+01	2.895361E+03
y	18	2.000000E+00	2.895361E+03
y	8	2.270000E+02	1.657023E+03
y	8	2.000000E+02	1.657023E+03
y	8	9.000000E+01	1.657023E+03
y	11	3.150000E+04	4.352667E+03
n	18	NA	2.895361E+03
y	8	1.000000E+00	1.657023E+03
y	8	4.000000E+02	1.657023E+03
y	18	2.300000E+02	2.895361E+03
n	18	NA	2.895361E+03
:			

R code for computing estimates

```
z <- read.table("cedata.txt",header=TRUE)
w <- z$FINLWT21  ### sampling weights

zclass <- read.table("ceclass.fit",header=TRUE)
probmissing <- zclass[,5] ### estimated P(INTRDVX_ = C)
p <- 1-probmissing ### estimated P(INTRDVX is nonmissing)
group <- !is.na(z$INTRDVX) ### group of nonmissing INTRDVX obs
ipw <- sum(w[group]*z$INTRDVX[group]/p[group])
         /sum(w[group]/p[group])
zreg <- read.table("cereg.fit",header=TRUE)
yhat <- zreg$predicted
imputed <- (sum(w[group]*z$INTRDVX[group])
            +sum(w[!group]*yhat[!group]))/sum(w)
simple <- sum(w[group]*z$INTRDVX[group])/sum(w[group])
```

Grading report on Homework 2

Q1. None of the students excluded the variables with 'x' in description file in the logistic regression. Some fit full model first, and then fit the significant variables again. Others used GUIDE to "cheat". For example, some students used the important variables in importance score plot to fit logistic regression. Some even used the logistic function of GUIDE directly for this homework. I didn't penalize them for that. But I think the point of this question is to get a sense of how painful it is without GUIDE.

Q2. A lot of students missed some important variables in importance score plot. I can't see how they get that wrong from their report. I directed them to Siyu.

Q4. A lot of them plotted the estimations of forest against tree instead of logistic regression. I guess it is because they copy the code directly from the slides.

Homework 3 (CE data)
due in Canvas by 9:30AM, Thu Mar 11, 2021

1. Fit a GUIDE classification forest to the CE data to find $\hat{\pi}_i$, the estimated probability that INTRDVX $\neq$ C (INTRDVX non-missing); data and description files are `cedata.txt` and `ceclass.dsc` in `"CE Data Folder"`
2. Use the **IPW** weighting method on slide 73 to estimate the population mean of INTRDVX, with sampling weight (w_i) variable `FINLWT21`
3. Fit a GUIDE regression forest to the CE data to obtain $\hat{y}_j$, the estimates of the missing values of INTRDVX (note: use `cereg.dsc` or change `INTRDVX` to `D` and `INTRDVX_` to `X` in `ceclass.dsc`)
4. Use the **imputation** method on slide 73 to estimate the population mean of INTRDVX with sampling weight variable `FINLWT21`
5. Include input and output files (but not predicted value files) in your report

R packages (and limitations) for trees & forests

rpart: tree models

party: trees (ctree) and forests (cforest)

- no class priors
- case weights treated as replicate weights, not as weights as in weighted least squares

partykit: trees (ctree) and forests (cforest)

- no variables with all NAs
- no categorical variables with only 1 non-NA level
- no class priors
- case weights treated as replicate weights, not as weights as in weighted least squares
- categorical variables must have ≤ 31 levels
- cannot predict observations with new categorical levels

randomForest: no NAs allowed; no sampling weights

Data preparation steps

1. Change the descriptor for INTRDVX from ‘x’ to ‘n’ in ceclass.dsc
2. Create a data set from cedata.txt that contains only the non-excluded (‘x’) variables (change the ‘x’ descriptor for INTRDVX to ‘n’)
3. Run the R methods on resulting data set

Removing 'x' variables from cedata.txt

```
 0. Read the warranty disclaimer
 1. Create a GUIDE input file
 Input your choice: 1
 Name of batch input file: subset.in
 Input 1 for model fitting, 2 for importance or DIF scoring,
       3 for data conversion ([1:3], <cr>=1): 3
 Name of batch output file: subset.out
 Input 1 if D variable is categorical, 2 if real ([1:2], <cr>=1):
 Input name of data description file (max 100 characters);
 enclose with matching quotes if it has spaces: ceclass.dsc
 :
 Warning: "x" variables will be excluded
 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
 :
 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
-------------------------------------------------------------------
 0                                   abort this job
 Input your choice ([0:11], <cr>=1):
 Input name of new data file: subset.txt
 Input file is created!
 Run GUIDE with the command: guide < subset.in
```

Contents of subset.txt

```
# vartype <- rep("numeric",638)
# vartype[c(1,2,4,6,8,10,12,14,16,18,20,22,23,24,25,26,27,29,31,33,34,35,37,39,
#           41,43,45,47,49,52,54,56,58,60,62,64,66,68,70,72,73,74,75,76,77,78,79,
#           80,82,84,86,88,89,90,92,94,96,97,98,99,100,102,104,106,108,109,110,
#           111,112,113,114,115,116,118,119,120,122,123,124,125,126,128,130,131,
#           132,133,135,137,139,303,304,305,306,307,308,309,310,311,312,313,314,
#           315,316,317,318,319,321,323,325,331,333,407,409,410,411,453,454,456,
#           458,460,462,464,465,466,467,468,470,472,474,476,477,478,479,482,484,
#           486,488,490,492,494,496,497,498,499,500,502,504,506,508,510,512,514,
#           516,518,520,522,524,526,528,530,532,534,536,538,540,542,544,546,548,
#           550,552,554,556,558,560,562,564,566,568,570,572,574,576,578,580,582,
#           584,585,586,588,590,592,594,596,598,600,602,604,606,608,610,612,614,
#           616,618,620,622,624,626,628,630,632,635,637)] <- "factor"
# z <- read.table("subset.txt",header=TRUE,colClasses=vartype)
#
DIRACC DIRACC_ AGE_REF AGE_REF_ AGE2 AGE2_ AS_COMP1 AS_C_MP1 AS_COMP2 AS_C_MP2 AS_C
"1" "D" 82 "D" 87 "T" 1 "D" 1 "D" 0 "D" 0 "D" 0 "D" 2 "D" 2 "D" 1 "1" "D" "1" "D" "
```

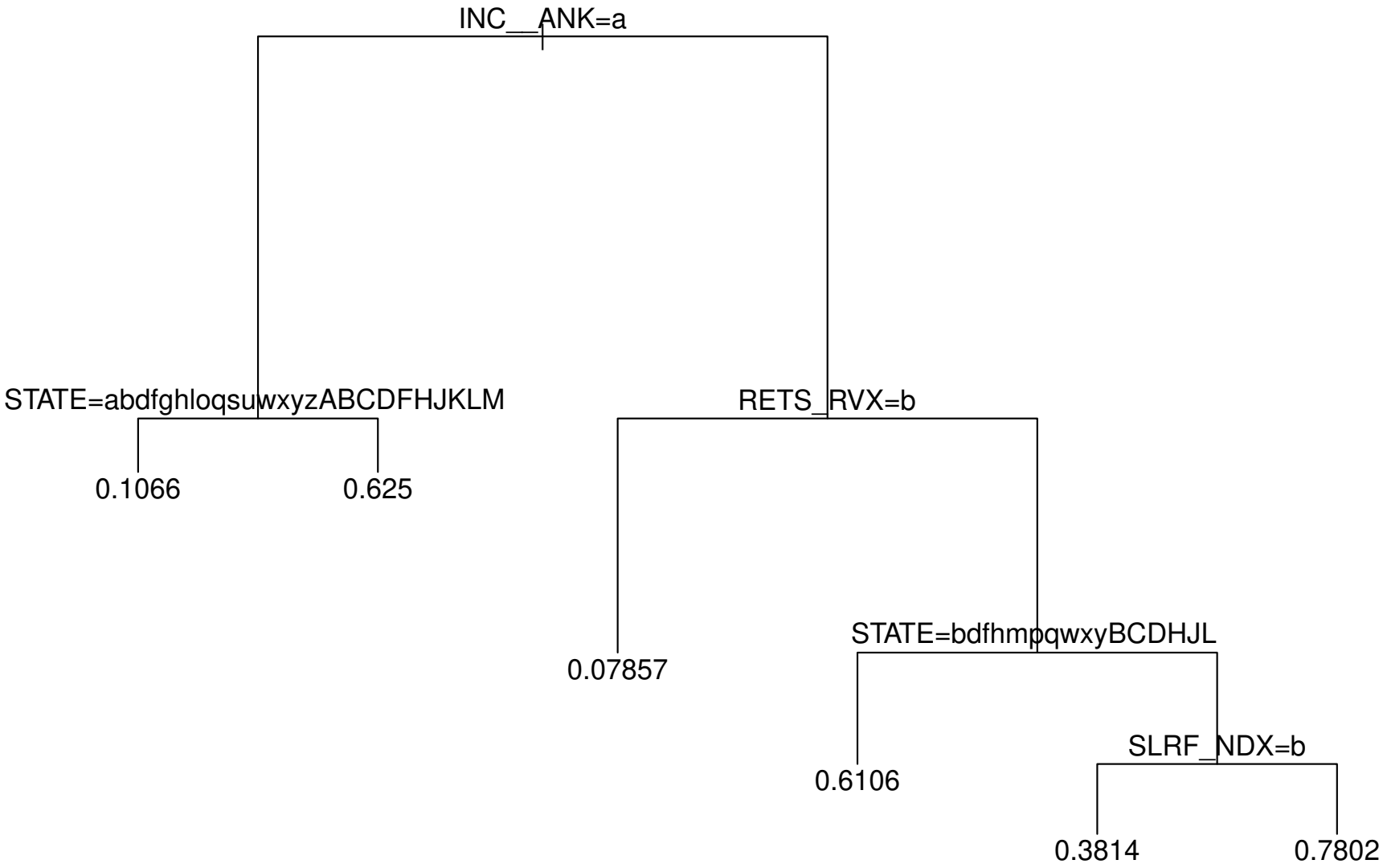
Code for reading subset.txt into R

```
library(rpart)
vartype <- rep("numeric",638)
vartype[c(1,2,4,6,8,10,12,14,16,18,20,22,23,24,25,26,27,29,31,33,34,35,37,39,
          41,43,45,47,49,52,54,56,58,60,62,64,66,68,70,72,73,74,75,76,77,78,79,
          80,82,84,86,88,89,90,92,94,96,97,98,99,100,102,104,106,108,109,110,
          111,112,113,114,115,116,118,119,120,122,123,124,125,126,128,130,131,
          132,133,135,137,139,303,304,305,306,307,308,309,310,311,312,313,314,
          315,316,317,318,319,321,323,325,331,333,407,409,410,411,453,454,456,
          458,460,462,464,465,466,467,468,470,472,474,476,477,478,479,482,484,
          486,488,490,492,494,496,497,498,499,500,502,504,506,508,510,512,514,
          516,518,520,522,524,526,528,530,532,534,536,538,540,542,544,546,548,
          550,552,554,556,558,560,562,564,566,568,570,572,574,576,578,580,582,
          584,585,586,588,590,592,594,596,598,600,602,604,606,608,610,612,614,
          616,618,620,622,624,626,628,630,632,635,637)] <- "factor"
z <- read.table("subset.txt",header=TRUE,colClasses=vartype)
```

R code for IPW estimate using RPART

```
tmp <- rep(NA,nrow(z))
tmp[z$INTRDVX_ == "C"] <- 0
tmp[z$INTRDVX_ == "D" | z$INTRDVX_ == "T"] <- 1
z$INTRDVX_ <- tmp  ### convert INTRDVX to binary variable
### regression tree without INTRDVX and FINLWT21
rp <- rpart(INTRDVX_ ~ . - INTRDVX - FINLWT21, data=z, method="anova")
plot(rp,compress=TRUE,margin=0.1)
text(rp)  ### plot is on next page
p <- predict(rp) ### predicted prob(INTRDVX_ = 1)
w <- z$FINLWT21
y <- z$INTRDVX
gp <- !is.na(y)
ipw <- sum(w[gp]*y[gp]/p[gp])/sum(w[gp]/p[gp])
print(ipw)
```

RPART tree for π = P(INTRDVX is nonmissing)



RPART tree (text form) for estimating π

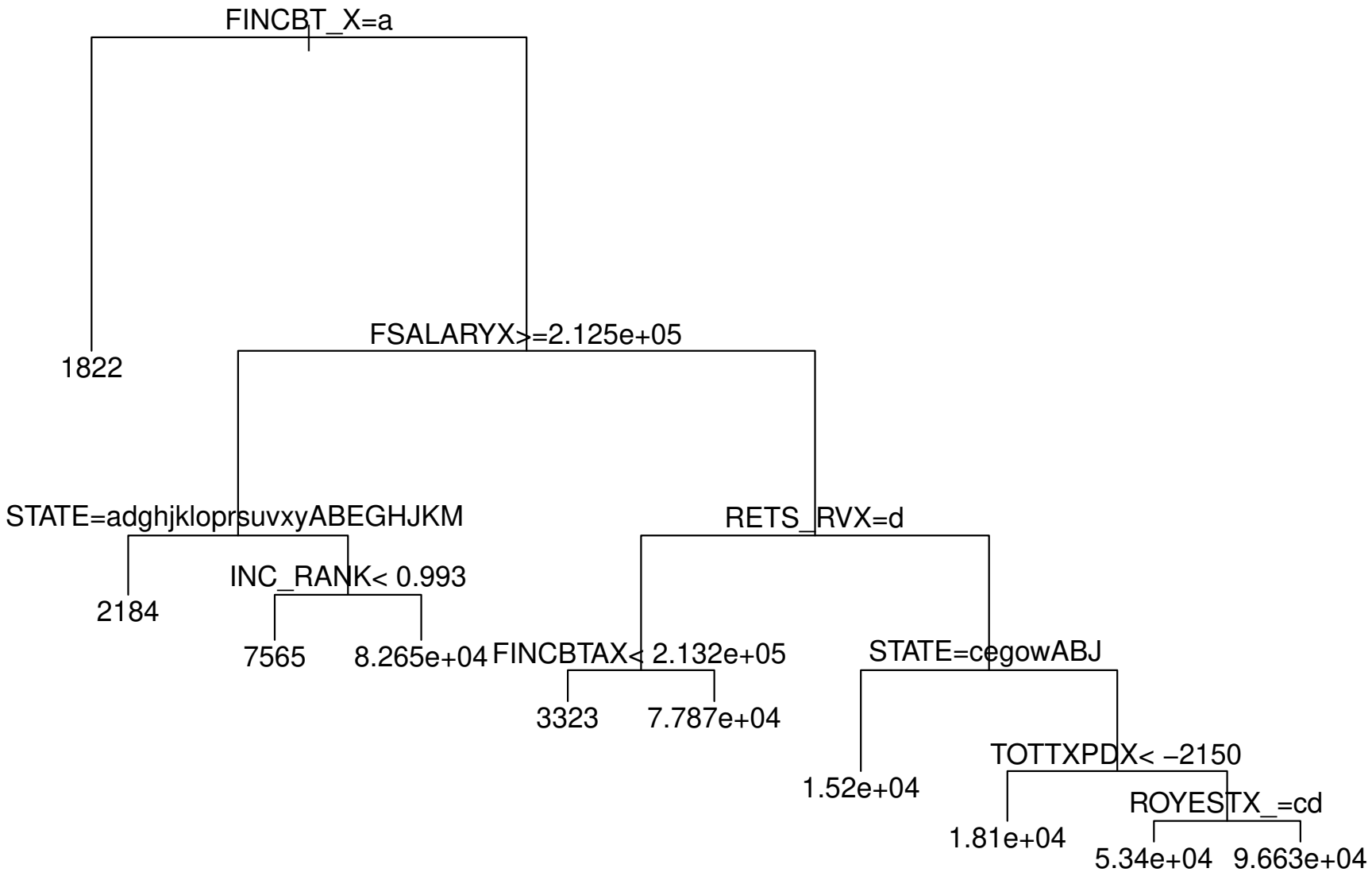
```
node), split, n, deviance, yval
      * denotes terminal node
 1) root 4693 1102.67700 0.62262940
   2) INC__ANK=A 367   52.83924 0.17438690
     4) STATE=1,10,12,15,16,17,21,24,26,29,32,34,36,39,4,41,42,45,47,49,53,55,6,
              8,9 319   30.37618 0.10658310 *
     5) STATE=13,18,2,22,25,27,48,51 48   11.25000 0.62500000 *
   3) INC__ANK=D 4326  969.84370 0.66065650
     6) RETS_RVX=C 140   10.13571 0.07857143 *
     7) RETS_RVX=A,D,T 4186  910.68630 0.68012420
      14) STATE=10,12,15,17,22,25,26,34,36,39,42,45,47,53,55,8
                2242  533.06740 0.61061550 *
      15) STATE=1,11,13,16,18,2,20,21,23,24,27,29,31,32,33,4,41,48,49,51,54,6,9
                1944  354.29420 0.76028810
        30) SLRF_NDX=C 97   22.88660 0.38144330 *
        31) SLRF_NDX=A,D,T 1847  316.75470 0.78018410 *
```

R code for computing $\hat{y}$ with RPART

```
rp2 <- rpart(INTRDVX ~ . - INTRDVX_, weight=FINLWT21, data=z,
                    method="anova")
plot(rp2,compress=TRUE,margin=0.1)
text(rp2)

y <- z$INTRDVX
w <- z$FINLWT21
miss <- is.na(y) ## obs with missing INTRDVX
yhat <- predict(rp2,newdata=z)
popmean <- (sum(w[!miss]*y[!miss])+sum(w[miss]*yhat[miss]))/sum(w)
print(popmean)
```

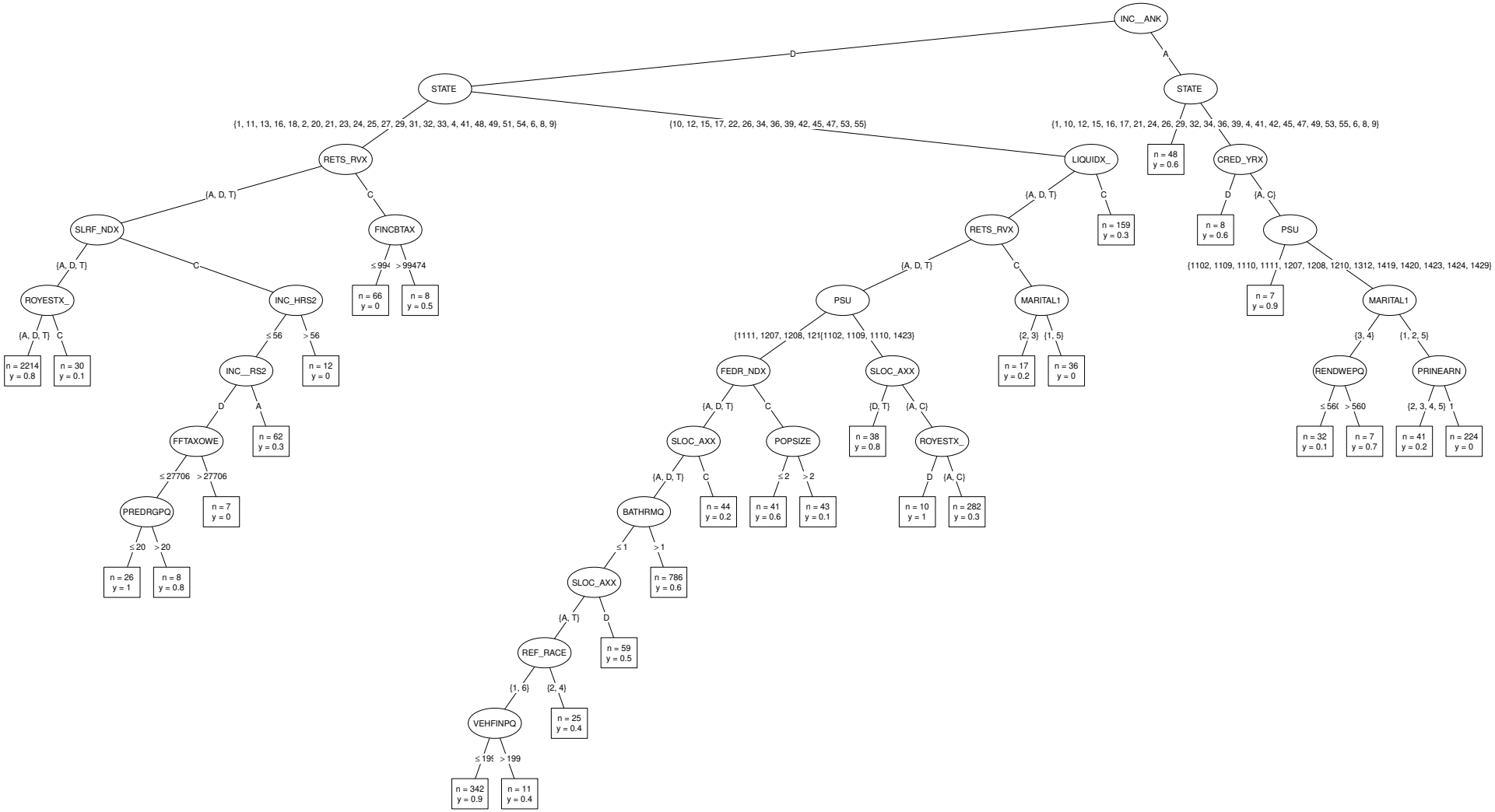
RPART regression tree for INTRDVX



RPART tree (text form) for computing $\hat{y}$

```
1) root 2922 1.627725e+16  4696.624
  2) FINCBT_X=D 2624 1.187559e+15  1821.816 *
  3) FINCBT_X=T 298 1.062296e+16 31918.830
    6) FSALARYX>=212524.5 148 1.977165e+15 10078.690
     12) STATE=1,12,16,17,2,20,21,24,25,27,29,32,33,36,39,41,42,48,51,53,55,6,9 1
     13) STATE=11,26,34,47 22 1.078676e+15 49096.130
       26) INC_RANK< 0.99305 10 1.314239e+14  7565.234 *
       27) INC_RANK>=0.99305 12 3.042116e+14 82649.860 *
    7) FSALARYX< 212524.5 150 6.014321e+15 54019.280
     14) RETS_RVX=T 50 6.222578e+14 11980.150
       28) FINCBTAX< 213208.5 43 3.714334e+13  3322.542 *
       29) FINCBTAX>=213208.5 7 1.429290e+14 77873.200 *
     15) RETS_RVX=A,D 100 3.473378e+15 70862.200
       30) STATE=11,13,16,24,34,41,42,55 20 2.897553e+14 15197.080 *
       31) STATE=1,12,15,17,18,20,25,26,27,29,32,33,36,39,4,45,48,51,53,6,8,9 80
         62) TOTTXPDX< -2150 8 1.263134e+14 18098.420 *
         63) TOTTXPDX>=-2150 72 9.093254e+14 91097.050
          126) ROYESTX_=D,T 10 4.113645e+14 53399.160 *
          127) ROYESTX_=A 62 2.030268e+14 96632.950 *
```

Ctree (party) tree for π = P(INTRDVX nonmissing)



Ctree tree (text form) for estimating π

```
1) INC__ANK == {D}; criterion = 1, statistic = 340.382
  2) STATE == {1, 11, 13, 16, 18, 2, 20, 21, 23, 24, 25, 27, 29, 31, 32, 33, 4, 41, 48, 49, 51,
              54, 6, 8, 9}; criterion = 1, statistic = 262.161
    3) RETS_RVX == {A, D, T}; criterion = 1, statistic = 189.379
      4) SLRF_NDX == {A, D, T}; criterion = 1, statistic = 111.512
        5) ROYESTX_ == {A, D, T}; criterion = 1, statistic = 84.918
          6)*  weights = 2214
        5) ROYESTX_ == {C}
          7)*  weights = 30
      4) SLRF_NDX == {C}
        8) INC_HRS2 <= 56; criterion = 0.997, statistic = 35.683
          9) INC__RS2 == {D}; criterion = 1, statistic = 31.184
            10) FFTAXOWE <= 27706; criterion = 0.999, statistic = 24.43
              11) PREDRGPQ <= 20; criterion = 0.997, statistic = 30
                12)*  weights = 26
              11) PREDRGPQ > 20
                13)*  weights = 8
            10) FFTAXOWE > 27706
              14)*  weights = 7
          9) INC__RS2 == {A}
            15)*  weights = 62
        8) INC_HRS2 > 56
          16)*  weights = 12
    3) RETS_RVX == {C}
      17) FINCBTAX <= 99474; criterion = 0.984, statistic = 17.41
```

```
      18)*  weights = 66
    17) FINCBTAX > 99474
      19)*  weights = 8
2) STATE == {10, 12, 15, 17, 22, 26, 34, 36, 39, 42, 45, 47, 53, 55}
  20) LIQUIDX_ == {A, D, T}; criterion = 1, statistic = 84.91
    21) RETS_RVX == {A, D, T}; criterion = 1, statistic = 79.685
      22) PSU == {1111, 1207, 1208, 1210, 1320}; criterion = 1, statistic = 75.365
        23) FEDR_NDX == {A, D, T}; criterion = 1, statistic = 39.217
          24) SLOC_AXX == {A, D, T}; criterion = 1, statistic = 40.86
            25) BATHRMQ <= 1; criterion = 0.999, statistic = 33.192
              26) SLOC_AXX == {A, T}; criterion = 1, statistic = 32.318
                27) REF_RACE == {1, 6}; criterion = 0.999, statistic = 45.581
                  28) VEHFINPQ <= 199; criterion = 0.966, statistic = 32.536
                    29)*  weights = 342
                  28) VEHFINPQ > 199
                    30)*  weights = 11
                27) REF_RACE == {2, 4}
                  31)*  weights = 25
              26) SLOC_AXX == {D}
                32)*  weights = 59
            25) BATHRMQ > 1
              33)*  weights = 786
          24) SLOC_AXX == {C}
            34)*  weights = 44
        23) FEDR_NDX == {C}
          35) POPSIZE <= 2; criterion = 0.993, statistic = 21.409
            36)*  weights = 41
          35) POPSIZE > 2
```

```
            37)*  weights = 43
          22) PSU == {1102, 1109, 1110, 1423}
            38) SLOC_AXX == {D, T}; criterion = 1, statistic = 33.196
              39)*  weights = 38
            38) SLOC_AXX == {A, C}
              40) ROYESTX_ == {D}; criterion = 0.982, statistic = 26.582
                41)*  weights = 10
              40) ROYESTX_ == {A, C}
                42)*  weights = 282
        21) RETS_RVX == {C}
          43) MARITAL1 == {2, 3}; criterion = 0.969, statistic = 22.237
            44)*  weights = 17
          43) MARITAL1 == {1, 5}
            45)*  weights = 36
      20) LIQUIDX_ == {C}
        46)*  weights = 159
1) INC__ANK == {A}
  47) STATE == {13, 18, 2, 22, 25, 27, 48, 51}; criterion = 1, statistic = 93.544
    48)*  weights = 48
  47) STATE == {1, 10, 12, 15, 16, 17, 21, 24, 26, 29, 32, 34, 36, 39, 4, 41, 42, 45, 47, 49, 53,
                 55, 6, 8, 9}
    49) CRED_YRX == {D}; criterion = 0.996, statistic = 42.276
      50)*  weights = 8
    49) CRED_YRX == {A, C}
      51) PSU == {1313, 1422}; criterion = 1, statistic = 56.706
        52)*  weights = 7
      51) PSU == {1102, 1109, 1110, 1111, 1207, 1208, 1210, 1312, 1419, 1420, 1423, 1424, 1429}
        53) MARITAL1 == {3, 4}; criterion = 0.996, statistic = 33.4
```

```
  54) RENDWEPQ <= 560; criterion = 0.996, statistic = 24
    55)*  weights = 32
  54) RENDWEPQ > 560
    56)*  weights = 7
53) MARITAL1 == {1, 2, 5}
  57) PRINEARN == {2, 3, 4, 5}; criterion = 1, statistic = 38.199
    58)*  weights = 41
  57) PRINEARN == {1}
    59)*  weights = 224
```

R code for IPW estimate using Ctree

```
tmp <- rep(NA,nrow(z))
tmp[z$INTRDVX_ == "C"] <- 0
tmp[z$INTRDVX_ == "D" | z$INTRDVX_ == "T"] <- 1
z$INTRDVX_ <- tmp
### regression tree without INTRDVX and FINLWT21
fmla <- formula(INTRDVX_ ~ . - INTRDVX - FINLWT21)
ct <- ctree(fmla, data=z)
plot(ct, type="simple",            # no terminal plots
  inner_panel=node_inner(ct,
       abbreviate = FALSE,           # do not shorten variable names
       pval = FALSE,                 # no p-values
       id = FALSE),                  # no id of node
  terminal_panel=node_terminal(ct,
       abbreviate = TRUE,
       digits = 1,                   # few digits on numbers
       fill = c("white"),            # make box white not grey
       id = FALSE)
   )
```

R code for IPW estimate using Ctree (cont'd.)

```
y <- z$INTRDVX
p <- predict(ct)
gp <- !is.na(y)
ipw <- sum(w[gp]*y[gp]/p[gp])/sum(w[gp]/p[gp])
print(ipw)
```

Ctree, Cforest and randomForest are inapplicable for computing $\hat{y}$ (regression) because they do not allow sampling weights

R code for IPW estimate using CFOREST

```
fmla <- formula(INTRDVX_ ~ . - INTRDVX - FINLWT21)
cf <- cforest(fmla, data=z)
p <- predict(cf,newdata=z)
w <- z$FINLWT21
y <- z$INTRDVX
gp <- !is.na(y)
ipw <- sum(w[gp]*y[gp]/p[gp])/sum(w[gp]/p[gp])
print(ipw)
```

R code for IPW estimate using randomForest after mean imputation

```
library(randomForest)
vartype <- rep("numeric",638)
vartype[c(1,2,4,6,8,10,12,14,16,18,20,22,23,24,25,26,27,29,31,33,34,35,37,39,
          41,43,45,47,49,52,54,56,58,60,62,64,66,68,70,72,73,74,75,76,77,78,79,
          80,82,84,86,88,89,90,92,94,96,97,98,99,100,102,104,106,108,109,110,
          111,112,113,114,115,116,118,119,120,122,123,124,125,126,128,130,131,
          132,133,135,137,139,303,304,305,306,307,308,309,310,311,312,313,314,
          315,316,317,318,319,321,323,325,331,333,407,409,410,411,453,454,456,
          458,460,462,464,465,466,467,468,470,472,474,476,477,478,479,482,484,
          486,488,490,492,494,496,497,498,499,500,502,504,506,508,510,512,514,
          516,518,520,522,524,526,528,530,532,534,536,538,540,542,544,546,548,
          550,552,554,556,558,560,562,564,566,568,570,572,574,576,578,580,582,
          584,585,586,588,590,592,594,596,598,600,602,604,606,608,610,612,614,
          616,618,620,622,624,626,628,630,632,635,637)] <- "factor"
z <- read.table("subset.txt",header=TRUE,colClasses=vartype)
```

```
### change INTRDVX_ to 0, 1
tmp <- rep(NA,nrow(z))
tmp[z$INTRDVX_ == "C"] <- 0
tmp[z$INTRDVX_ == "D" | z$INTRDVX_ == "T"] <- 1
z$INTRDVX_ <- tmp
gp <- tmp == 1  ### obs with nonmissing INTRDVX
### replace missing values in non-categorical vars with means
for(k in 1:ncol(z)){
    x <- z[,k]
    if(!is.factor(x)){
        m0 <- sum(is.na(x))
        if(m0 > 0){
            ##            print(c(k,names(z)[k],m0))
            m <- mean(x,na.rm=TRUE)
            x[is.na(x)] <- m
            z[,k] <- x
        }
    }
}
```

```
### create a 'missing' level for factor variables
for(k in 1:ncol(z)){
    x <- z[,k]
    if(is.factor(x) & sum(is.na(x) > 0)){
        levels <- levels(x)
        levels[length(levels) + 1] <- "missing"
        x <- factor(x, levels = levels)
        x[is.na(x)] <- "missing"
        z[,k] <- x
    }
}
null.cols <- NULL  ### find columns with all NAs
for(k in 1:ncol(z)){
    x <- z[,k]
    if(sum(!is.na(x)) == 0){
        print(c(k,names(z)[k]))
        null.cols <- c(null.cols,k)
    }
}
z <- z[,-null.cols]  ### remove variables with all missing values
```

```
### randomForest is slow for large data sets if formula interface is used
### extract matrix of predictor variables
x <- z[,-which(names(z) %in% c("INTRDVX_","INTRDVX","FINLWT21"))]
rf <- randomForest(x=x,y=z$INTRDVX_) ## using formula is slow
### ipw estimate
w <- z$FINLWT21[gp]
y <- z$INTRDVX[gp]
p <- predict(rf,newdata=z[gp,])
ipw <- sum(w*y/p)/sum(w/p)
print(ipw)
```

Classification with categorical predictors: peptide-binding data

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

Peptide-binding data

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

1 = binder, 0 = non-binder

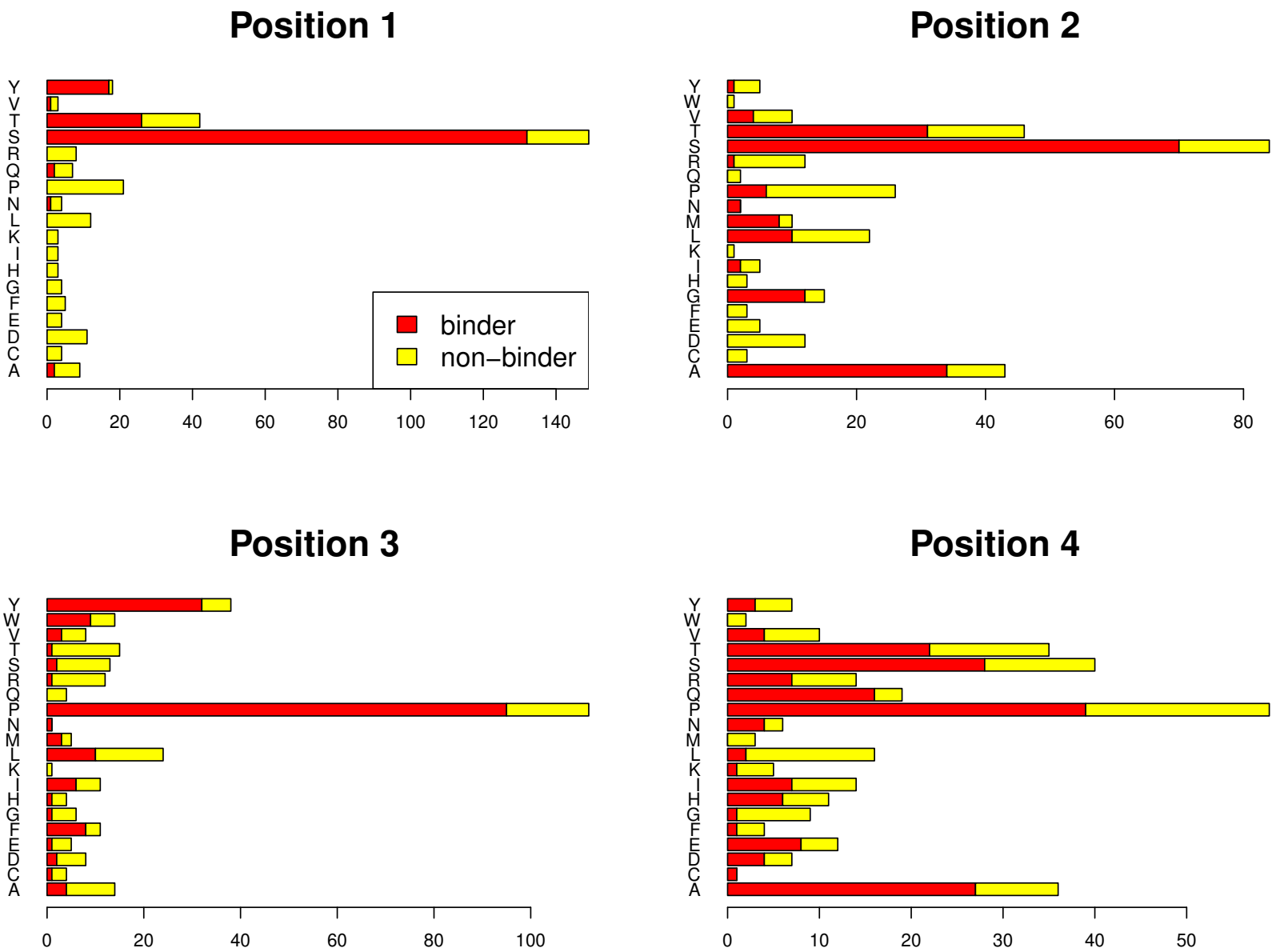
Traditional logistic regression

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

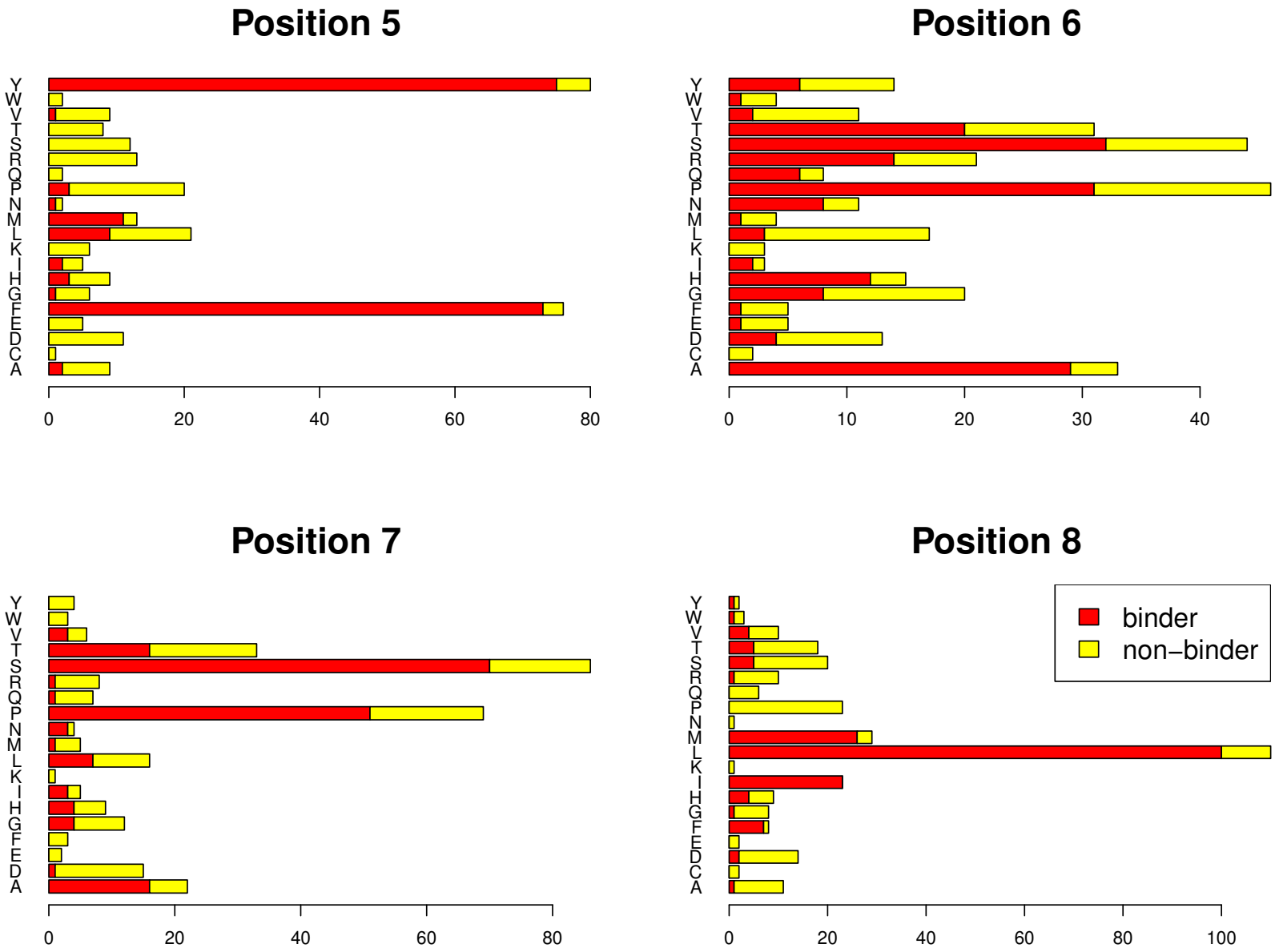
LASSO logistic regression coefficients

pos1C	-0.016	pos3D	-0.404	pos5D	-0.627	pos7D	-0.638
pos1D	-0.571	pos3P	0.738	pos5E	-0.157	pos7F	-2.834
pos1L	-0.182	pos3R	-0.510	pos5F	2.141	pos7S	0.182
pos1P	-0.708	pos3T	-1.586	pos5M	1.187	pos8I	1.010
pos1S	1.608	pos4L	-0.520	pos5N	-0.333	pos8L	1.438
pos1T	0.474	pos4P	0.330	pos5P	-0.243	pos8M	0.834
pos2D	-0.808			pos5R	-0.569	pos8P	-0.843
pos2M	0.658			pos5S	-0.696	pos8Q	-0.189
pos2N	1.476			pos5T	-0.824	pos8Y	0.870
pos2P	-0.060			pos5Y	1.807		
pos2S	0.175			pos6D	-0.088		
				pos6S	0.404		

Distributions of peptide-binding data



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



Chi-squared tests

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

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

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

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

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

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

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

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

Chi-squared tests (cont'd.)

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

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

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

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

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

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

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

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

Wilson-Hilferty χ^2_1

Pos1 (X^2 = 185.25, df = 17, p-value $<$ 2.2E-16, χ^2_1 = 105.2)

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

Pos5 (X^2 = 211.5, df = 19, p-value $<$ 2.2E-16, χ^2_1 = 119.9)

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

Pos8 (X^2 = 185.69, df = 19, p-value $<$ 2.2E-16, χ^2_1 = 100.7)

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

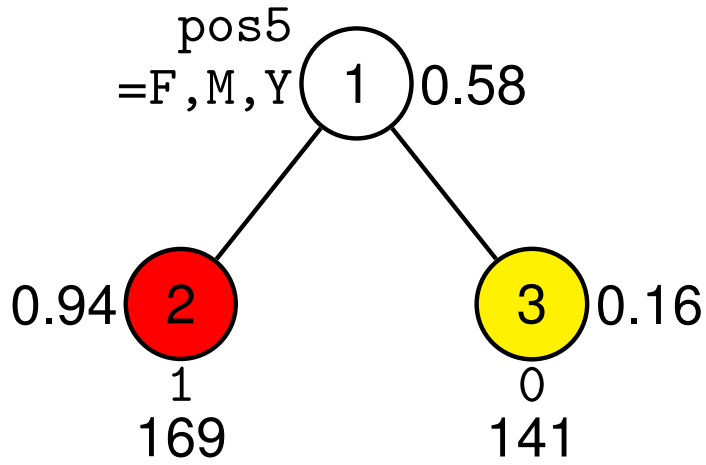
Modified Wilson-Hilferty (1931) approximation

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

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

- Then $P(\chi^2_\nu > X^2) \approx P(\chi^2_1 > W)$

GUIDE (default) classification tree

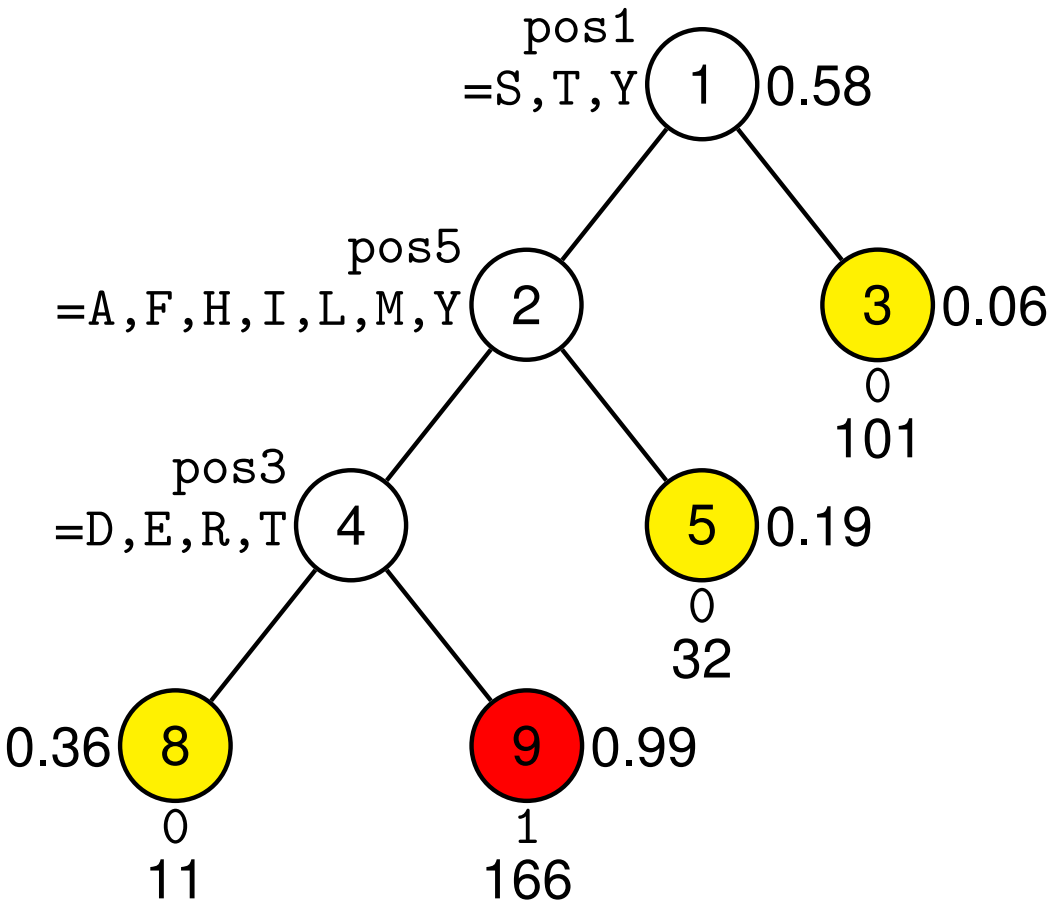


At each split, an obs. goes to left branch if and only if the condition is satisfied

Predicted class and sample size printed below terminal nodes

Class sample proportions for `bind` = `1` (red) beside nodes

GUIDE tree with 2nd best variable at root node

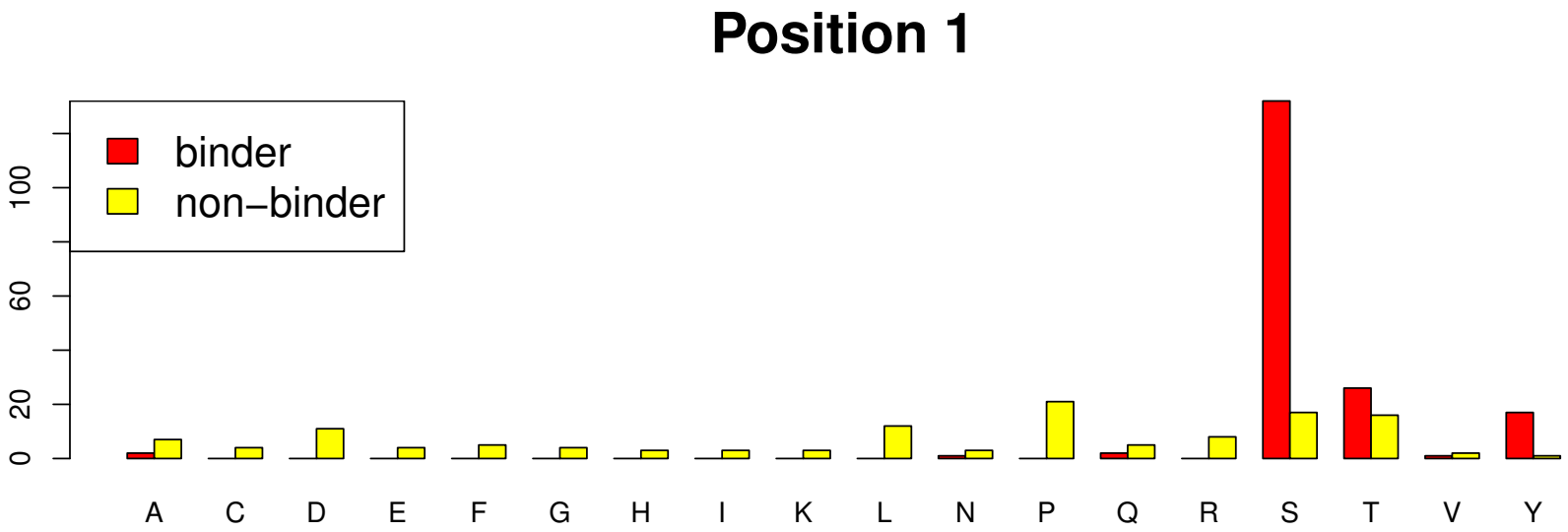


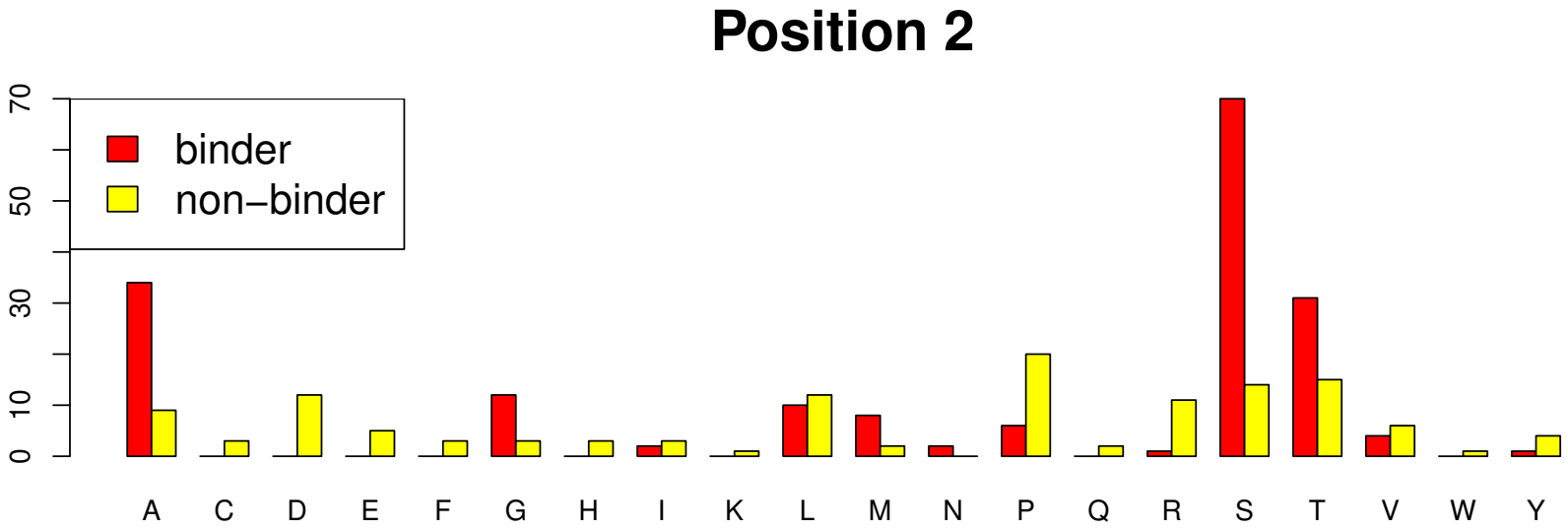
At each split, an obs. goes to left branch if and only if the condition is satisfied

Predicted class and sample size printed below terminal nodes

Class sample proportions for `bind = 1` (red) beside nodes

Classification by density estimation





GUIDE kernel density estimation nearest-neighbor options

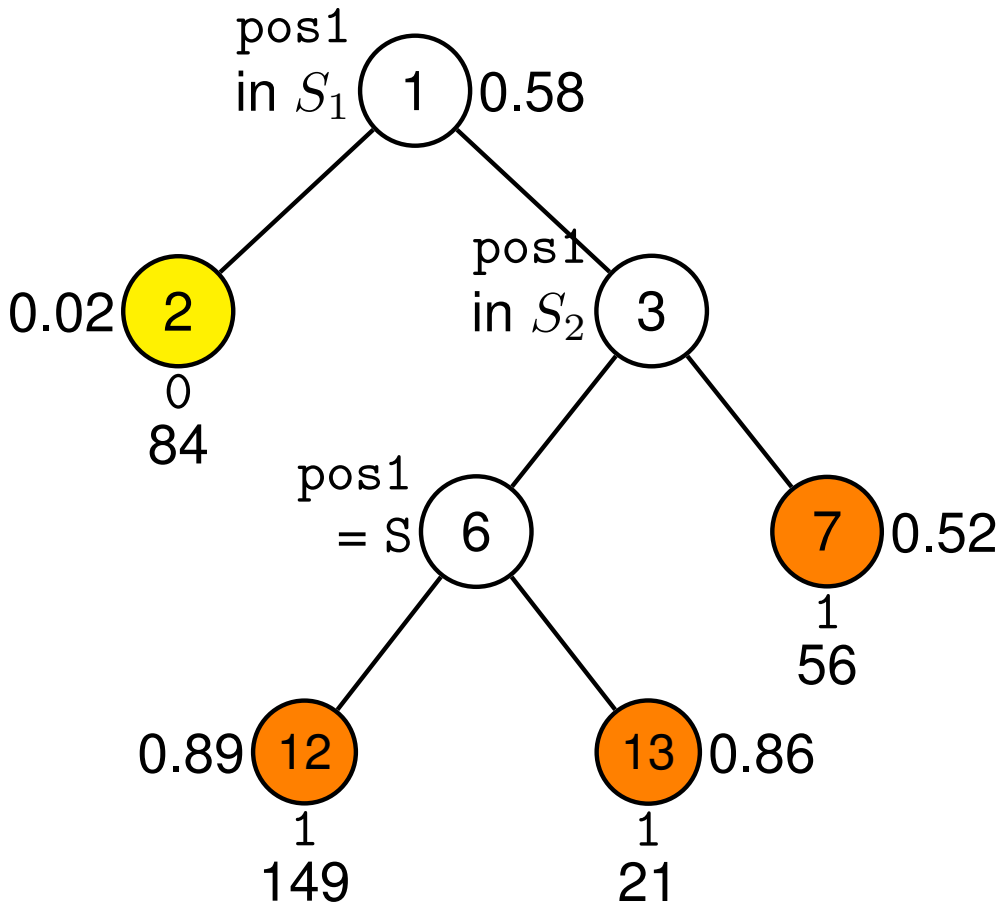
- Fit a kernel density or nearest-neighbor model in each node
- Choose between univariate and bivariate (default) node models

Bivariate kernel or NN predicted class at root node

	Pos5																		
Pos1	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
K	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
R	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S	0	0	0	0	1	1	1	1	0	1	1	0	0	0	0	0	0	1	0	1
T	1	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	1
V	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Y	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1

1 = binder, 0 = non-binder

GUIDE nearest-neighbor model

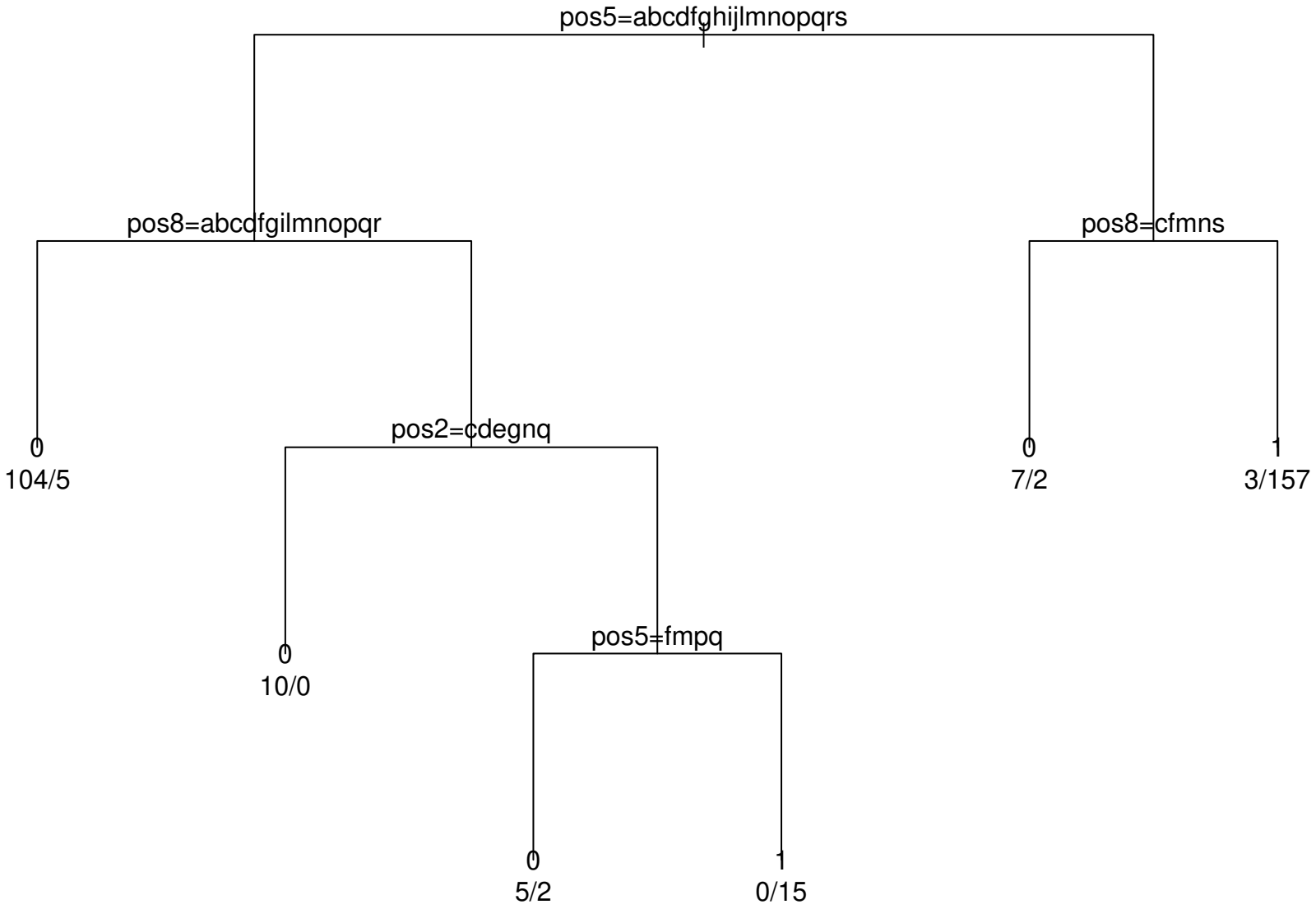


Set S_1 = {A, C, D, E, F, G, H, I, L, P, R}; set S_2 = {S, V, Y}

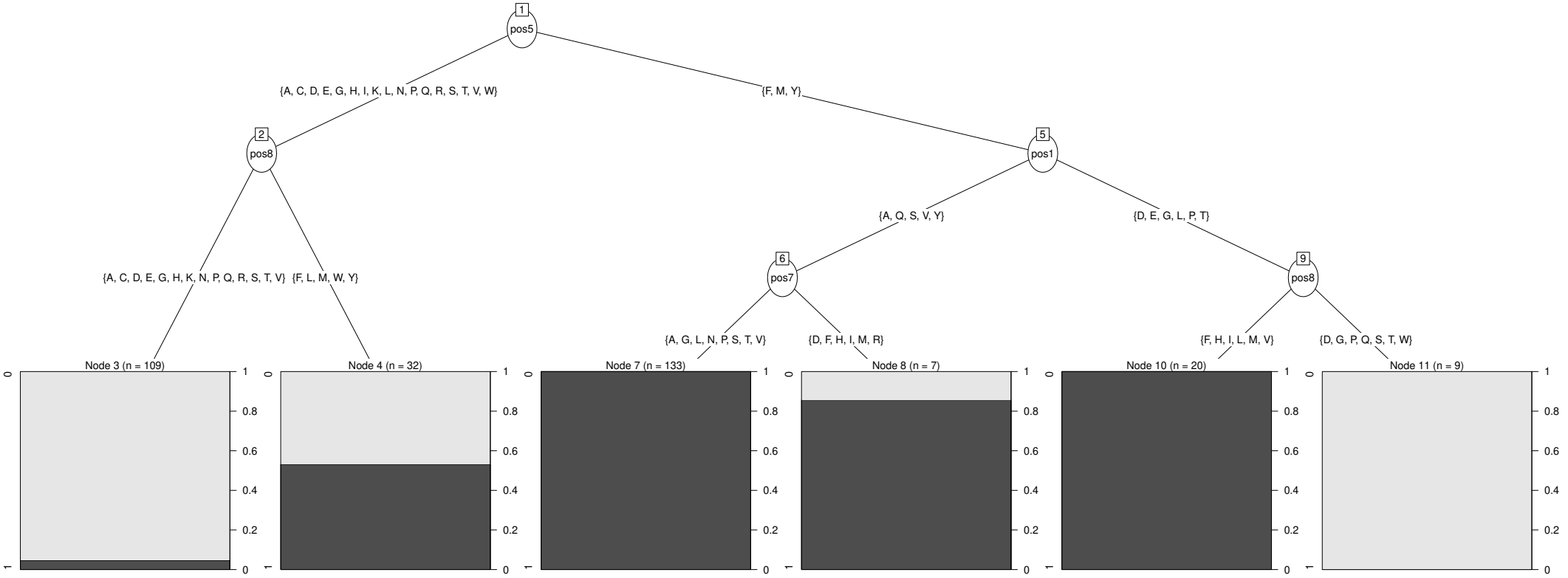
Predicted classes and sample sizes printed below terminal nodes

class sample proportion for bind = 1 beside nodes

Rpart tree for peptide data



Ctree tree for peptide data



Homework 4 (CE data)
due in Canvas by 9:30AM, Thu Mar 18, 2021

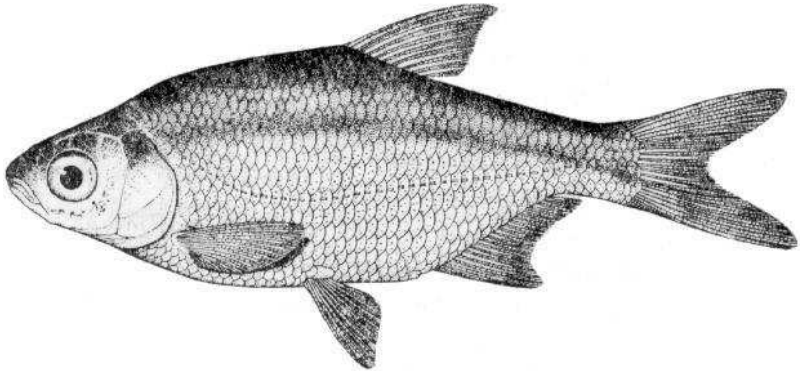
1. Fit an RPART classification tree to the CE data to find $\hat{\pi}_i$, the estimated probability that INTRDVX $\neq$ C (INTRDVX non-missing), and use the **IPW** weighting method to estimate the population mean of INTRDVX, with sampling weight (w_i) variable `FINLWT21`
2. Fit an RPART regression tree to the CE data to estimate the missing values of INTRDVX, $\hat{y}_i$, and use the **imputation** method to estimate the population mean of INTRDVX with sampling weight variable `FINLWT21`
3. Repeat Question 1 using Ctree and Cforest

Fish classification

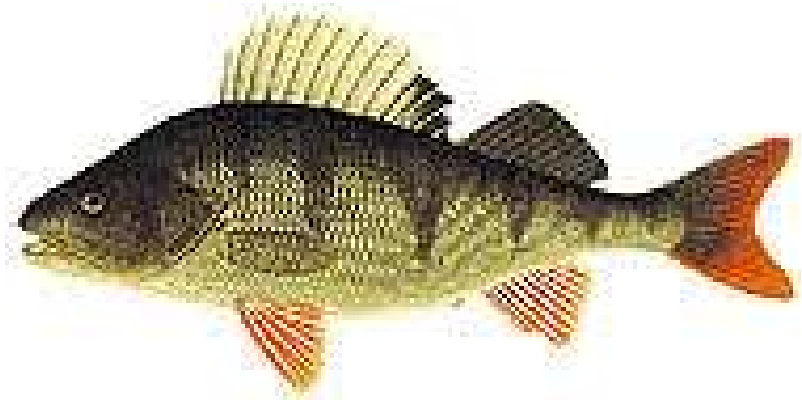
- 159 fish caught from the same lake near Tampere, Finland
- The fish are from 7 species: (1) 35 Bream, (2) 11 Parkki, (3) 56 Perch, (4) 17 Pike, (5) 20 Roach, (6) 14 Smelt, (7) 6 Whitefish

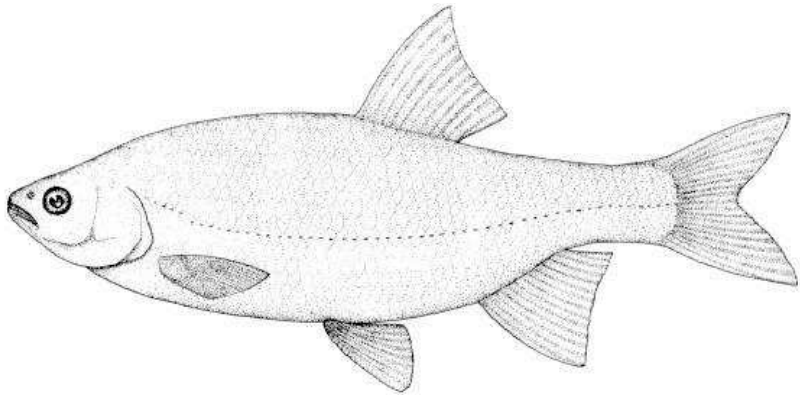
Predictor	Definition
Weight	Weight of the fish (in grams); 2 missing values
Length1	Length from the nose to the beginning of the tail (in cm)
Length2	Length from the nose to the notch of the tail (in cm)
Length3	Length from the nose to the end of the tail (in cm)
Heightpc	Maximal height as percent of Length3
Widthpc	Maximal width as percent of Length3
Sex	female, male, unknown

Bream (left) and Parkki (right)

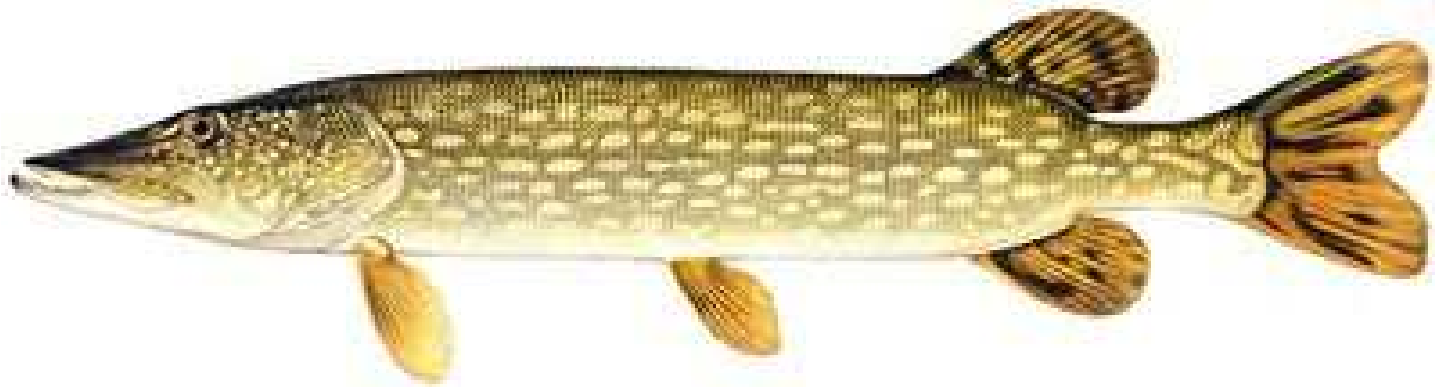


Perch (left) and Whitefish (right)

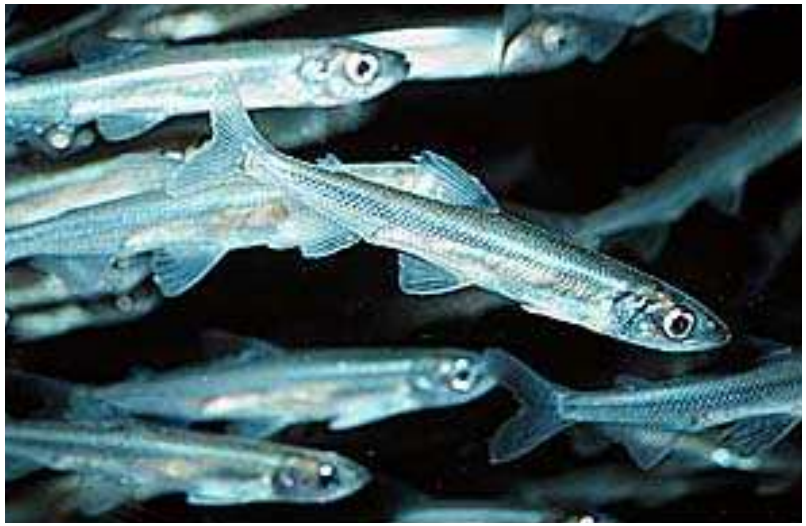




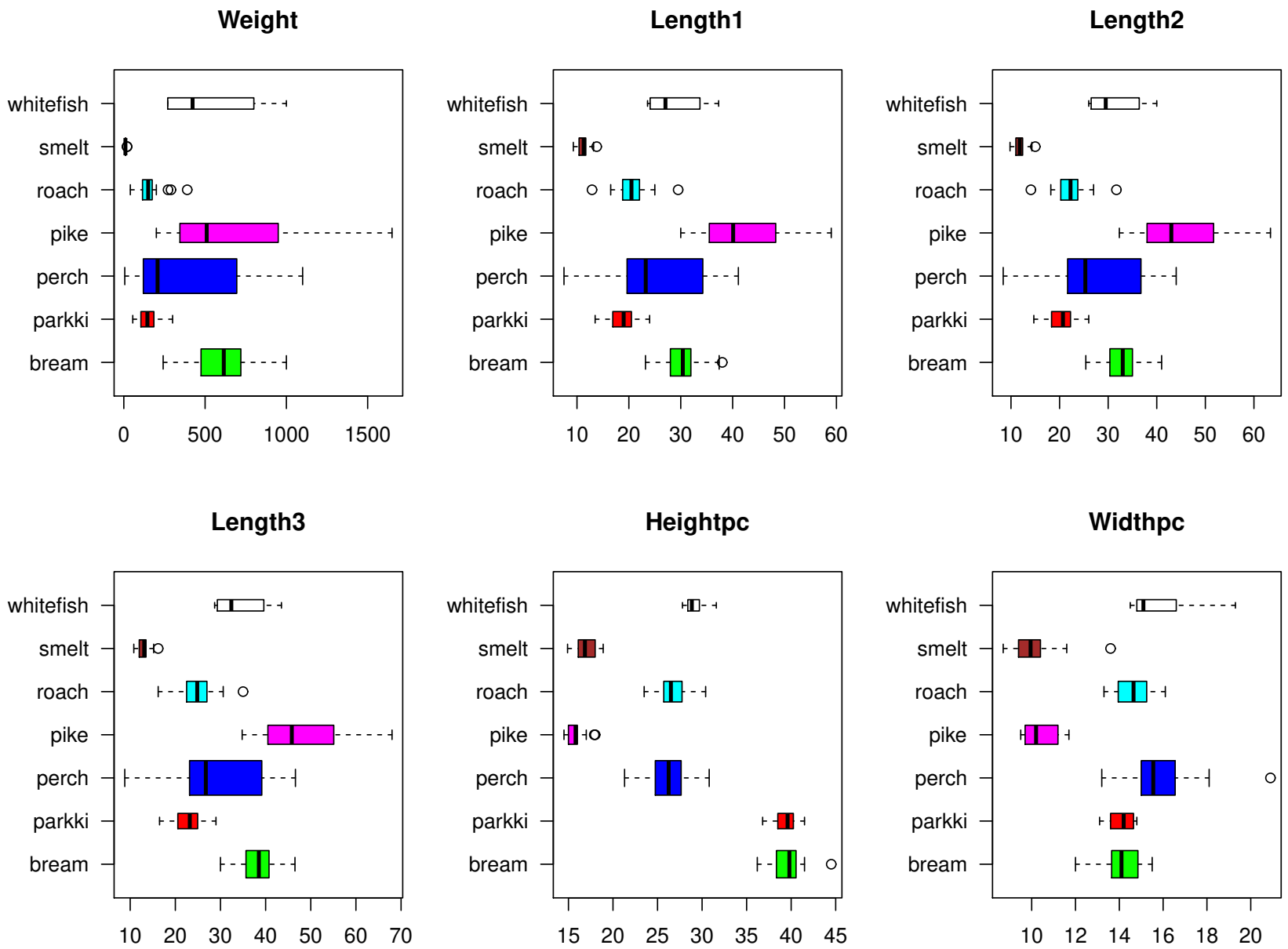
Pike



Roach (left) and Smelt (right)



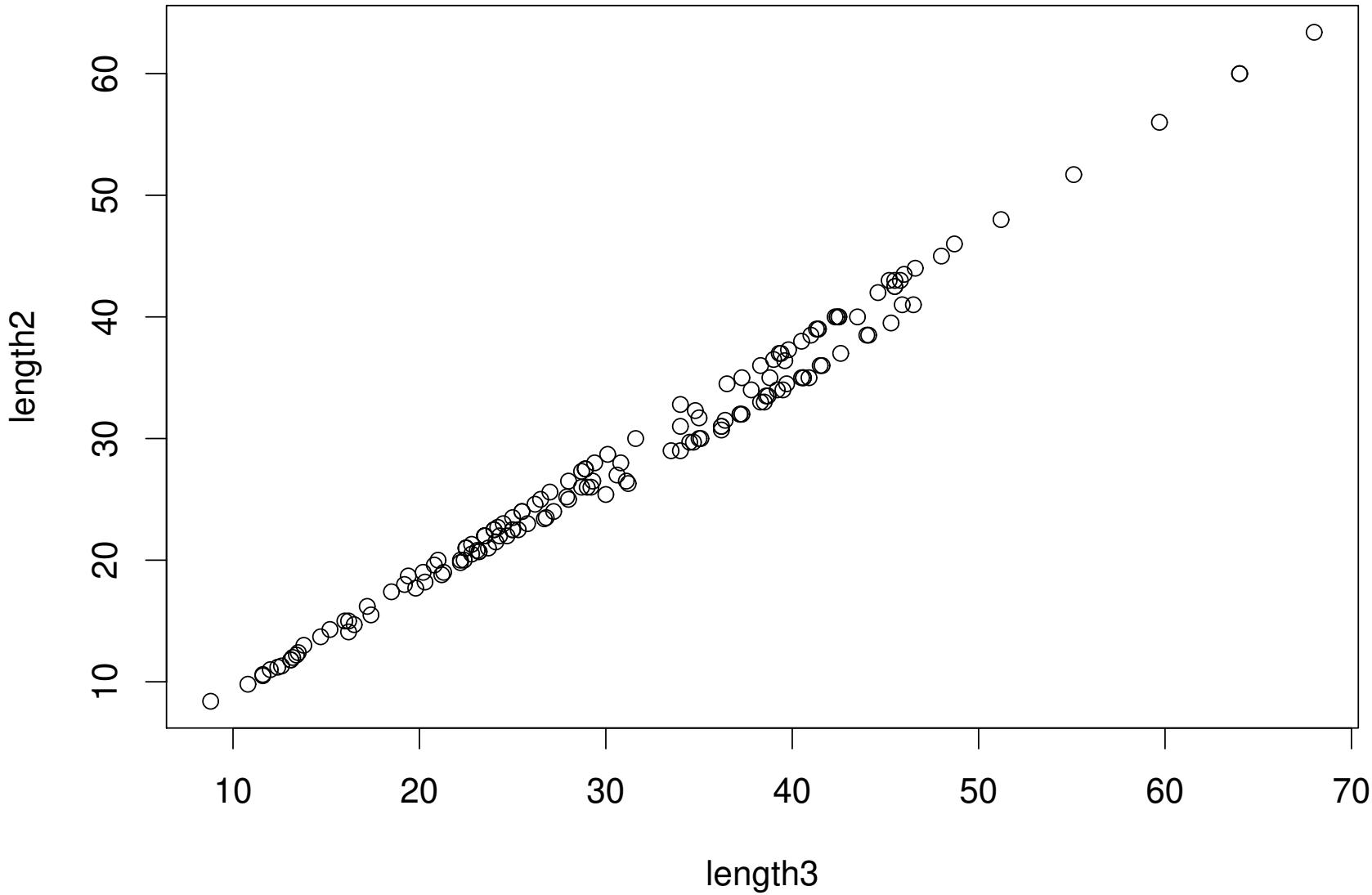
Boxplots of continuous variables



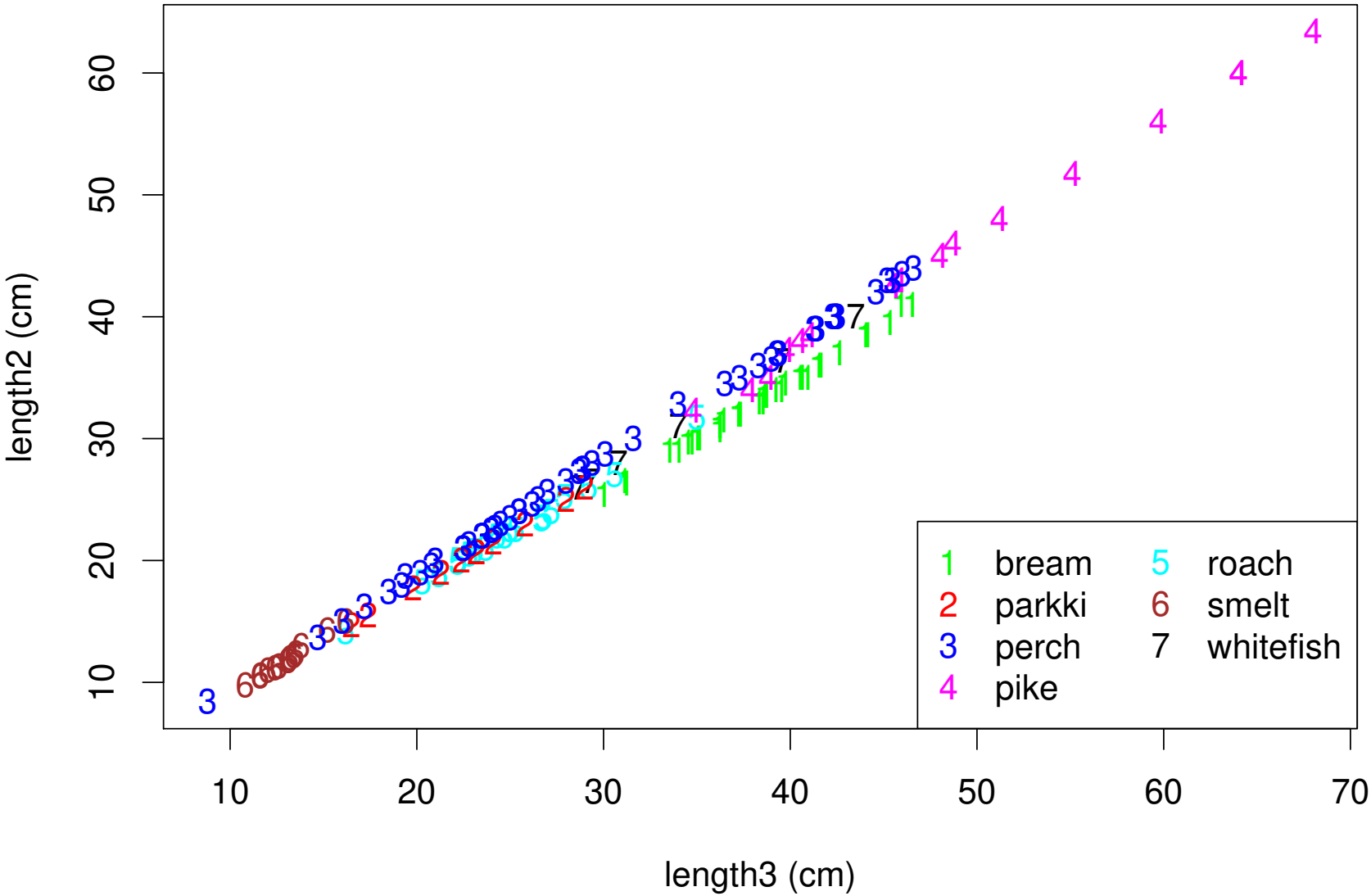
Sex by species

	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

Plot of Length2 vs. Length3



Plot of Length2 vs. Length3



Linear discriminant analysis (LDA)

- Suppose there are m classes, $G_1, G_2, \ldots, G_m$ of observations with sample sizes $n_1, n_2, \ldots, n_m$
- Let $f_j(\mathbf{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ be the normal density function, $N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ with common $\boldsymbol{\Sigma}$, for class G_j
- Likelihood for class G_j is $L_j(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}) = \prod_{i=1}^{n_j} f_j(\mathbf{x}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma})$
- Let $(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2, \ldots, \hat{\boldsymbol{\mu}_m}, \hat{\boldsymbol{\Sigma}})$ be the MLEs that maximize $\prod_{j=1}^{m} L_j(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$
- The *linear discriminant rule* classifies $\mathbf{x}$ to class k if

$$f_k(\mathbf{x}; \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}) = \max_j f_j(\mathbf{x}; \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}})$$

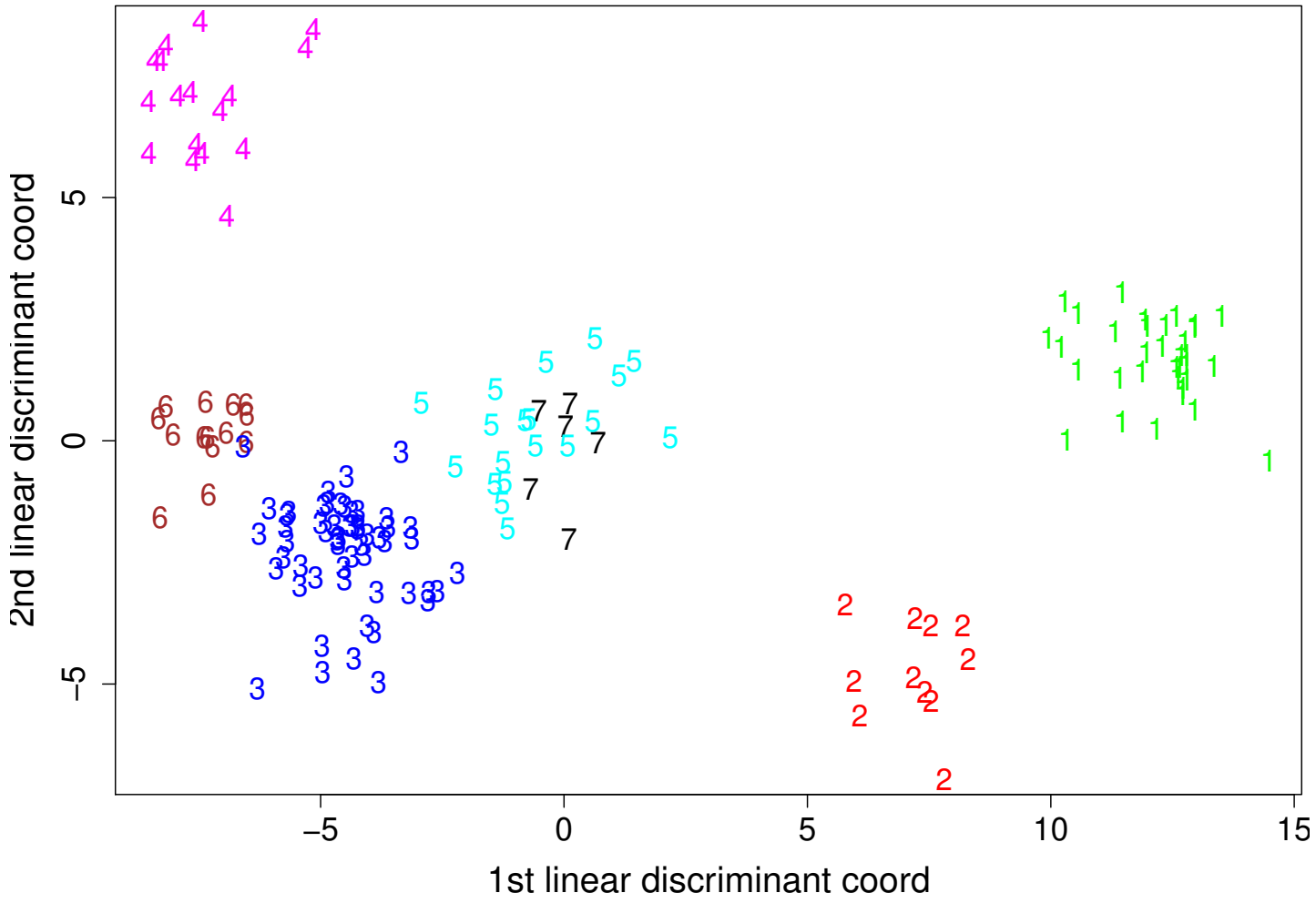
- For the fish data, LDA misclassifies 1 out of 157 complete cases

R lda function

```
> library(MASS)
> z <- read.table("fish.dat",header=TRUE,stringsAsFactors=TRUE)
> z <- na.omit(z,na.action="omit") ## omit cases with NA
> lda.model <- lda(species ~ .,data=z)
> predicted <- predict(lda.model,newdata=z)$class
> print(table(predicted,z$species))

predicted    bream parkki perch pike roach smelt whitefish
  bream         34      0     0    0     0     0          0
  parkki         0     11     0    0     0     0          0
  perch          0      0    56    0     0     0          0
  pike           0      0     0   17     0     0          0
  roach          0      0     0    0    18     0          0
  smelt          0      0     0    0     0    14          0
  whitefish      0      0     0    0     1     0          6
```

Plot of 1st two discriminant coords



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

R code for LDA plot

```
class <- as.character(z$species)
class.sym <- class
class.sym[class == "bream"] <- 1
class.sym[class == "parkki"] <- 2
class.sym[class == "perch"] <- 3
class.sym[class == "pike"] <- 4
class.sym[class == "roach"] <- 5
class.sym[class == "smelt"] <- 6
class.sym[class == "whitefish"] <- 7
class.sym <- as.numeric(class.sym)
pred2 <- predict(lda.model,dimen=2)$x
eqscplot(pred2,type="n",xlab="1st linear discriminant coord",
         ylab="2nd linear discriminant coord")
text(pred2,labels=class.sym, col=class.sym)
```

Multinomial logistic regression

- Let $Y = 1, 2, \ldots, J$ and $\pi = (\pi_1, \ldots, \pi_J)$ with $\pi_j = P(Y = j)$
- A multinomial logit model with predictors $x_1, x_2, \ldots, x_K$ has the form

$$\log(\pi_j/\pi_1) = \alpha_j + \sum_{k=1}^{K} \beta_{jk} x_k, \quad j = 2, \ldots, J$$

  i.e., $\pi_j = \pi_1 \exp\{\alpha_j + \sum_{k=1}^{K} \beta_{jk} x_k\}$

- Using $\sum_{j=1}^{J} \pi_j = 1$ and defining $\alpha_1 = \beta_{1k} = 0$, $k = 1, 2, \ldots, K$, yields

$$\pi_j = \frac{\exp(\alpha_j + \sum_k \beta_{jk} x_k)}{\sum_{i=1}^{J} \exp(\alpha_i + \sum_k \beta_{ik} x_k)}$$

Multinomial logistic regression (cont'd.)

- Let the ith observation be $y_i = (y_{i1}, y_{i2}, \ldots, y_{iJ})$, where exactly one $y_{ij} = 1$ and the others are 0
- Let $\{\hat{\alpha}_j, \hat{\beta}_{jk}\}$ be the MLEs of the multinomial loglikelihood

$$\log \prod_{i=1}^{n} \prod_{j=1}^{J} \pi_j^{y_{ij}} = \sum_{i=1}^{n} \sum_{j=1}^{J} y_{ij} \log \pi_j$$

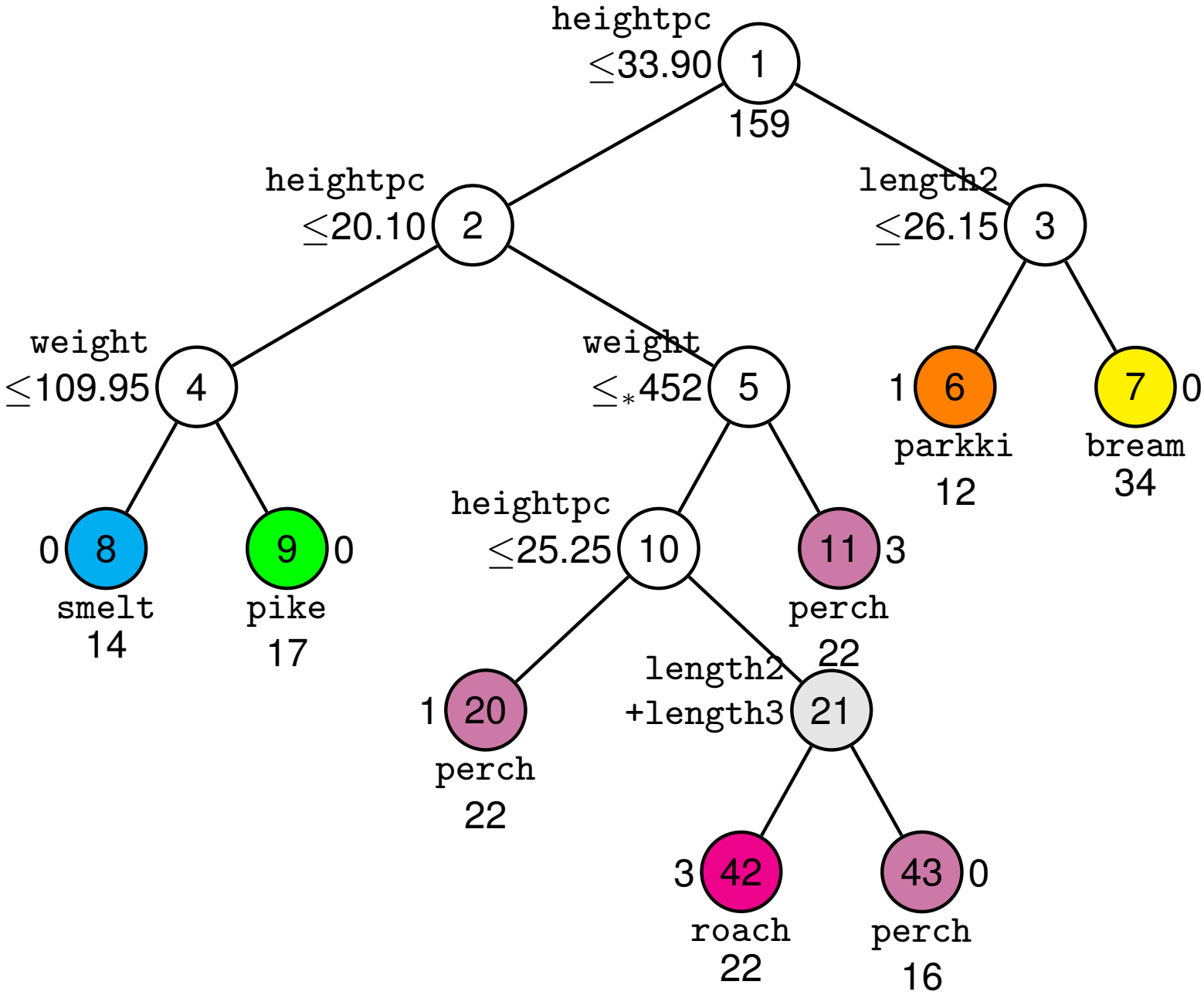
$$= \sum_{i=1}^{n} \left[\sum_{j=1}^{J} y_{ij} (\alpha_j + \sum_k \beta_{jk} x_k) - \log \left\{ \sum_{k=1}^{J} \exp(\alpha_k + \sum_k \beta_{jk} x_k) \right\} \right]$$

- An observation $\mathbf{x} = (x_1, x_2, \ldots, x_K)$ is classified to class j if

$$\hat{\pi}_j = \frac{\exp(\hat{\alpha}_j + \sum_k \hat{\beta}_{jk} x_k)}{\sum_{i=1}^{J} \exp(\hat{\alpha}_i + \sum_k \hat{\beta}_{ik} x_k)} = \max\{\hat{\pi}_i, i = 1, 2, \ldots, J\}$$

- R package `nnet` fits multinomial logistic regression

GUIDE default tree



Wilson-Hilferty χ^2_1

Heightpc (X^2 = 210.5, df = 12, p-value $<$ 2.2E-16, χ^2_1 = 139.8)

species	$\leq$ 25.3	(25.3, 29.2]	$>$ 29.2
bream	0	0	34
parkki	0	0	11
perch	21	32	3
pike	17	0	0
roach	1	18	1
smelt	14	0	0
whitefish	0	4	2

Widthpc (X^2 = 111.0, df = 12, p-value $<$ 2.2E-16, χ^2_1 = 81.1)

species	$\leq$ 13.8	(13.8, 15]	$>$ 15
bream	16	14	5
parkki	4	7	0
perch	2	17	37
pike	17	0	0
roach	3	9	8
smelt	14	0	0
whitefish	0	3	3

Wilson-Hilferty χ^2_1 (cont'd.)

Length1 (X^2 = 97.0, df = 12, p-value = 2.2E-15, χ^2_1 = 66.2)

species	$\leq$ 20.5	(20.5, 30.6]	$>$ 30.6
bream	0	18	17
parkki	8	3	0
perch	21	17	18
pike	0	1	16
roach	11	9	0
smelt	14	0	0
whitefish	0	4	2

Length2 (X^2 = 97.8, df = 12, p-value = 1.5E-15, χ^2_1 = 67.1)

species	$\leq$ 22.5	(22.5, 33.2]	$>$ 33.2
bream	0	18	17
parkki	8	3	0
perch	21	17	18
pike	0	1	16
roach	13	7	0
smelt	14	0	0
whitefish	0	4	2

Wilson-Hilferty χ^2_1 (cont'd.)

Length3 (X^2 = 98.4, df = 12, p-value = 1.1E-15, χ^2_1 = 67.8)

species	$\leq$ 24.4	(24.4, 37.5]	$>$ 37.5
bream	0	16	19
parkki	8	3	0
perch	22	18	16
pike	0	1	16
roach	9	11	0
smelt	14	0	0
whitefish	0	4	2

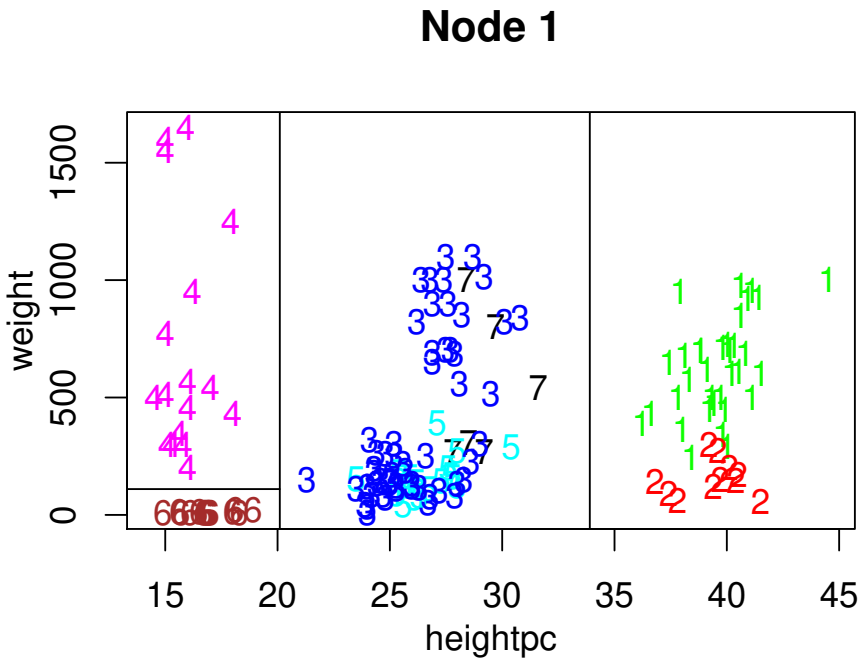
Weight (X^2 = 83.6, df = 12, p-value = 8.3E-13, χ^2_1 = 52.8)

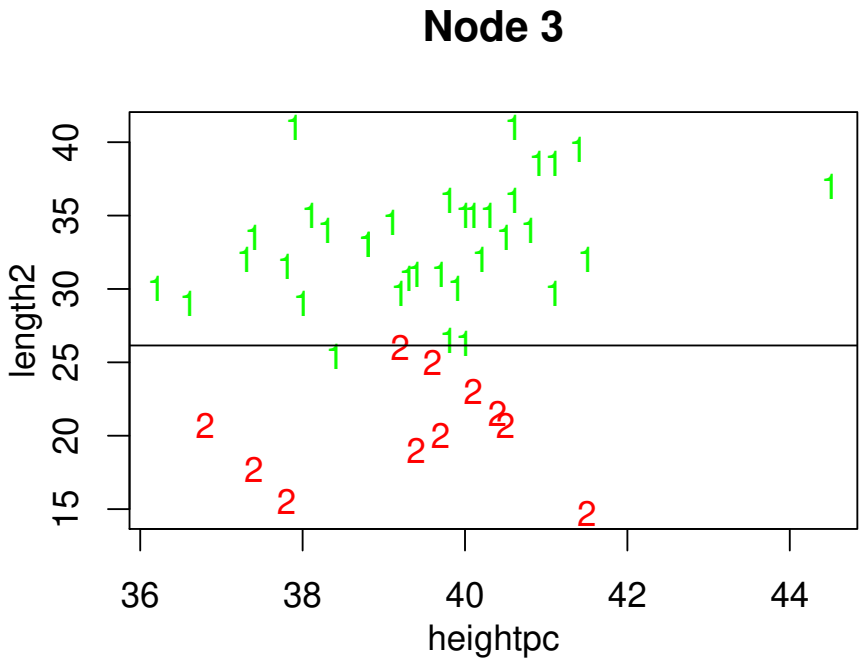
species	$\leq$ 273	$>$ 273	NA
bream	1	33	1
parkki	10	1	0
perch	34	22	0
pike	1	16	0
roach	17	2	1
smelt	14	0	0
whitefish	2	4	0

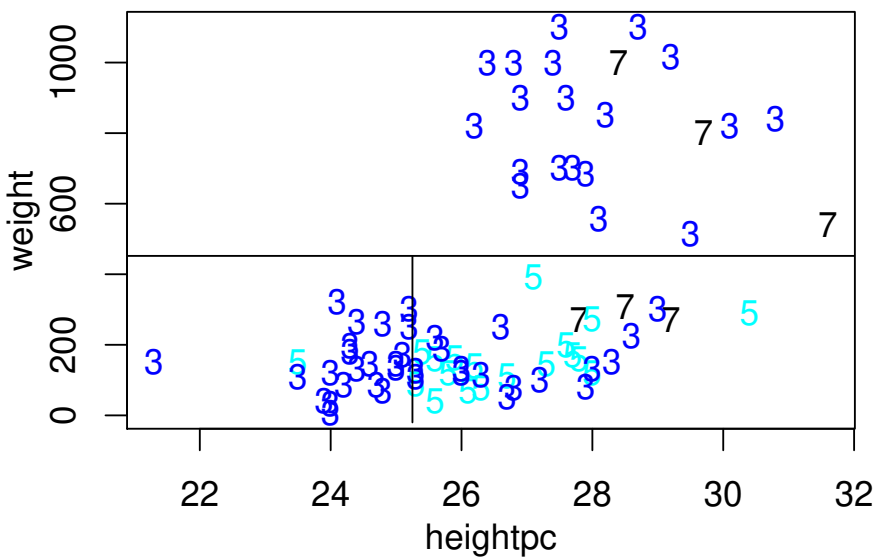
Wilson-Hilferty χ^2_1 (cont'd.)

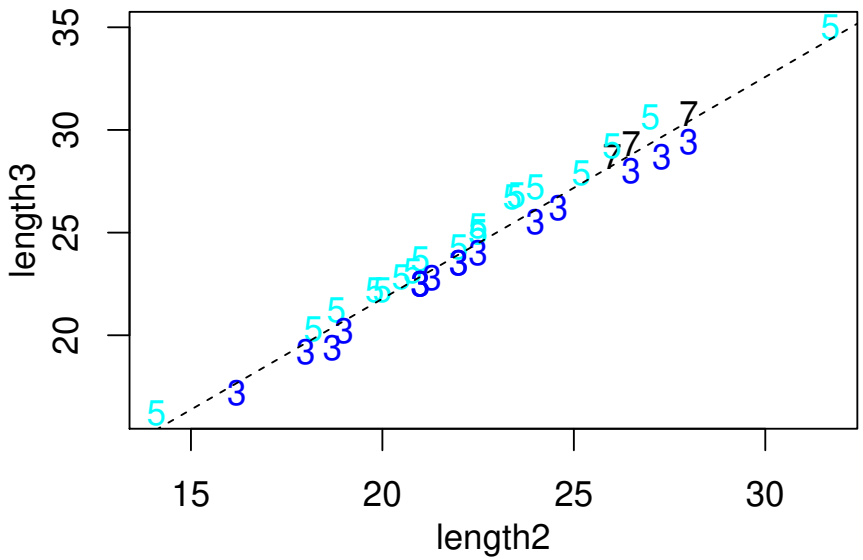
Sex (X^2 = 43.3, df = 12, p-value = 2.0E-05, χ^2_1 = 19.5)

species	female	male	unknown
bream	3	6	26
parkki	4	3	4
perch	25	2	29
pike	5	1	11
roach	8	0	12
smelt	9	5	0
whitefish	1	0	5

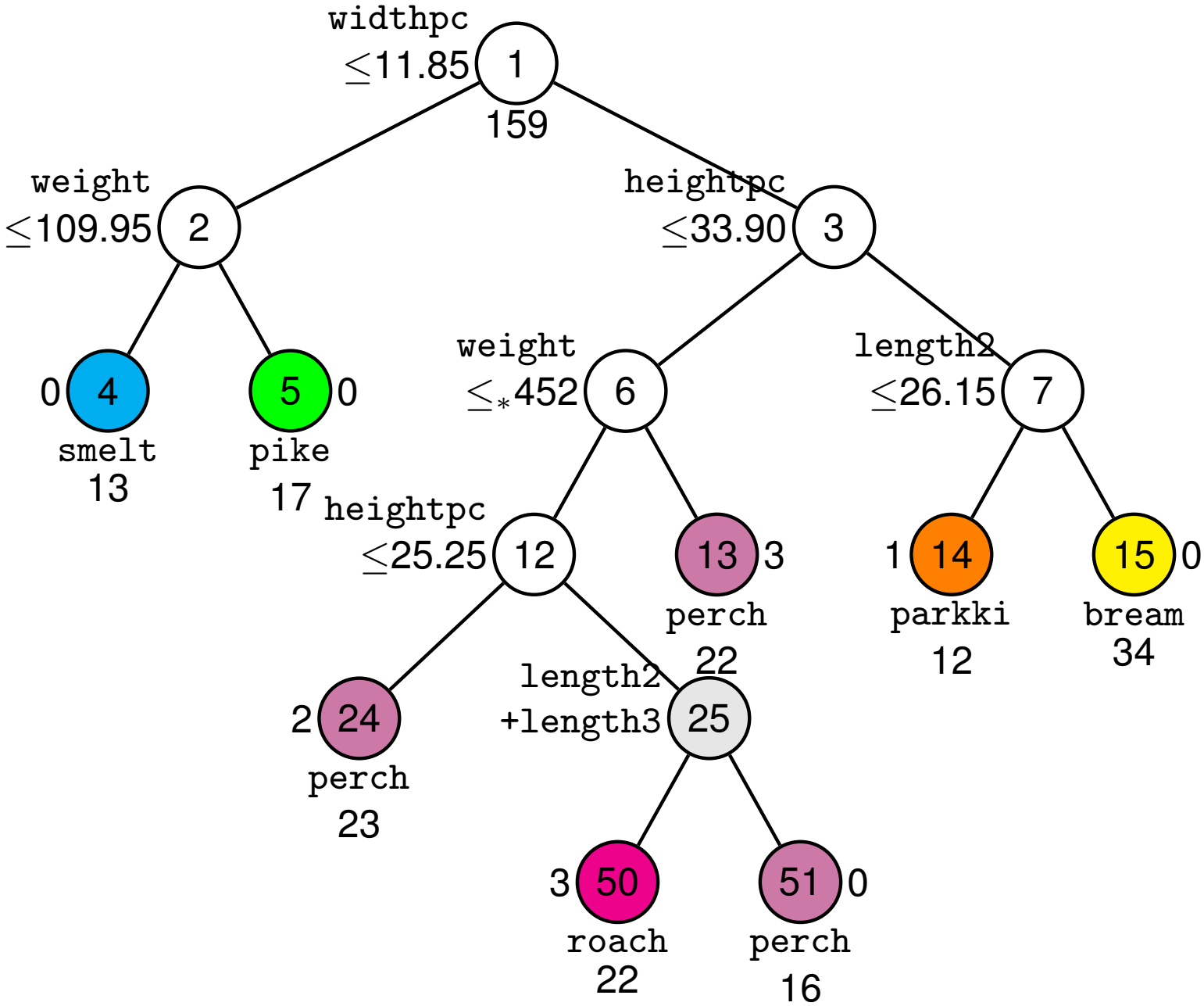

Node 1
weight
heightpc
0
500
1000
1500
15
20
25
30
35
40
45


Node 3
length2
heightpc
15
20
25
30
35
40
36
38
40
42
44

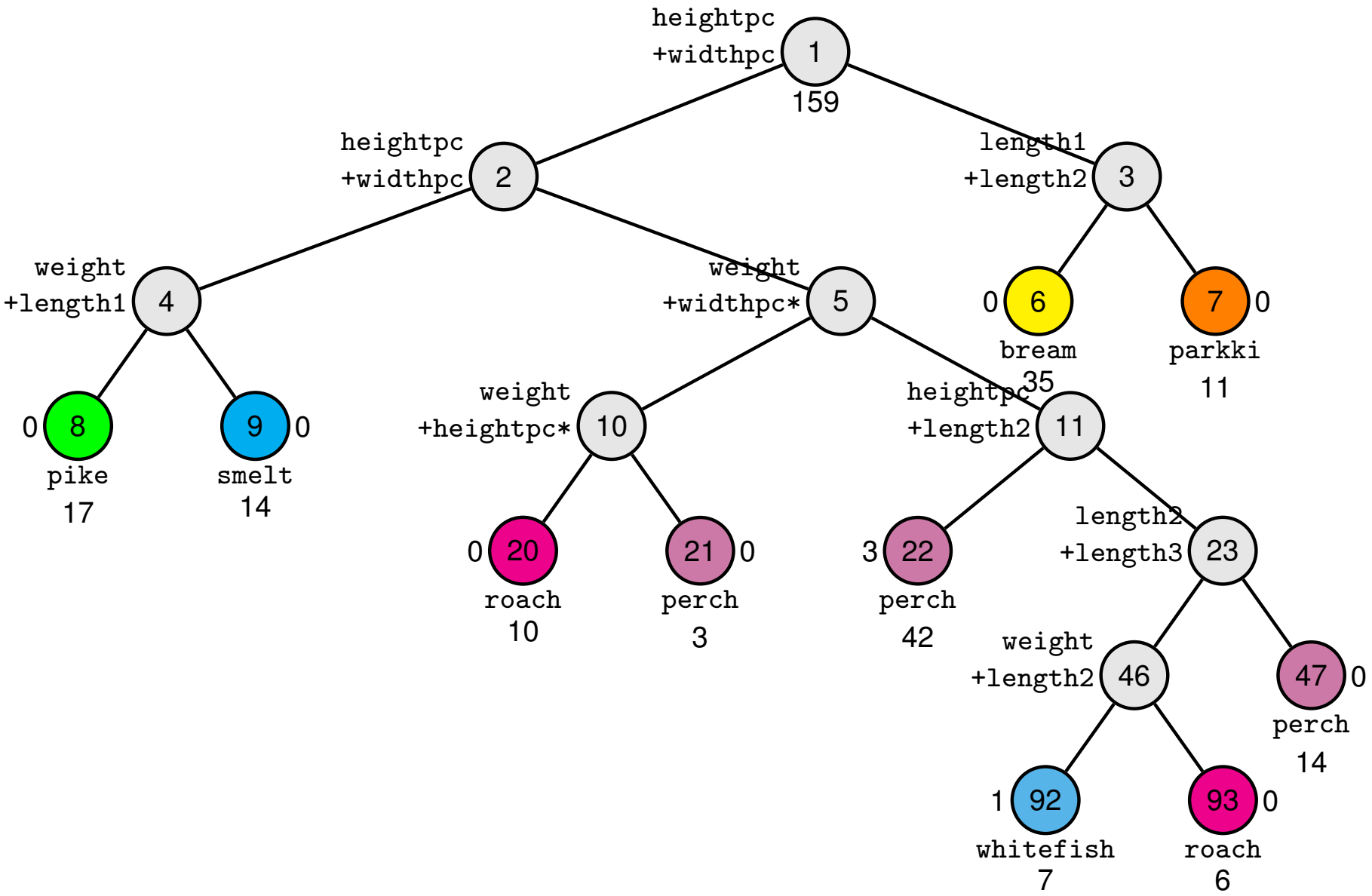

Node 5
weight
heightpc
0
200
600
1000
22
24
26
28
30
32


Node 21
length3
length2
20
25
30
35
15
20
25
30

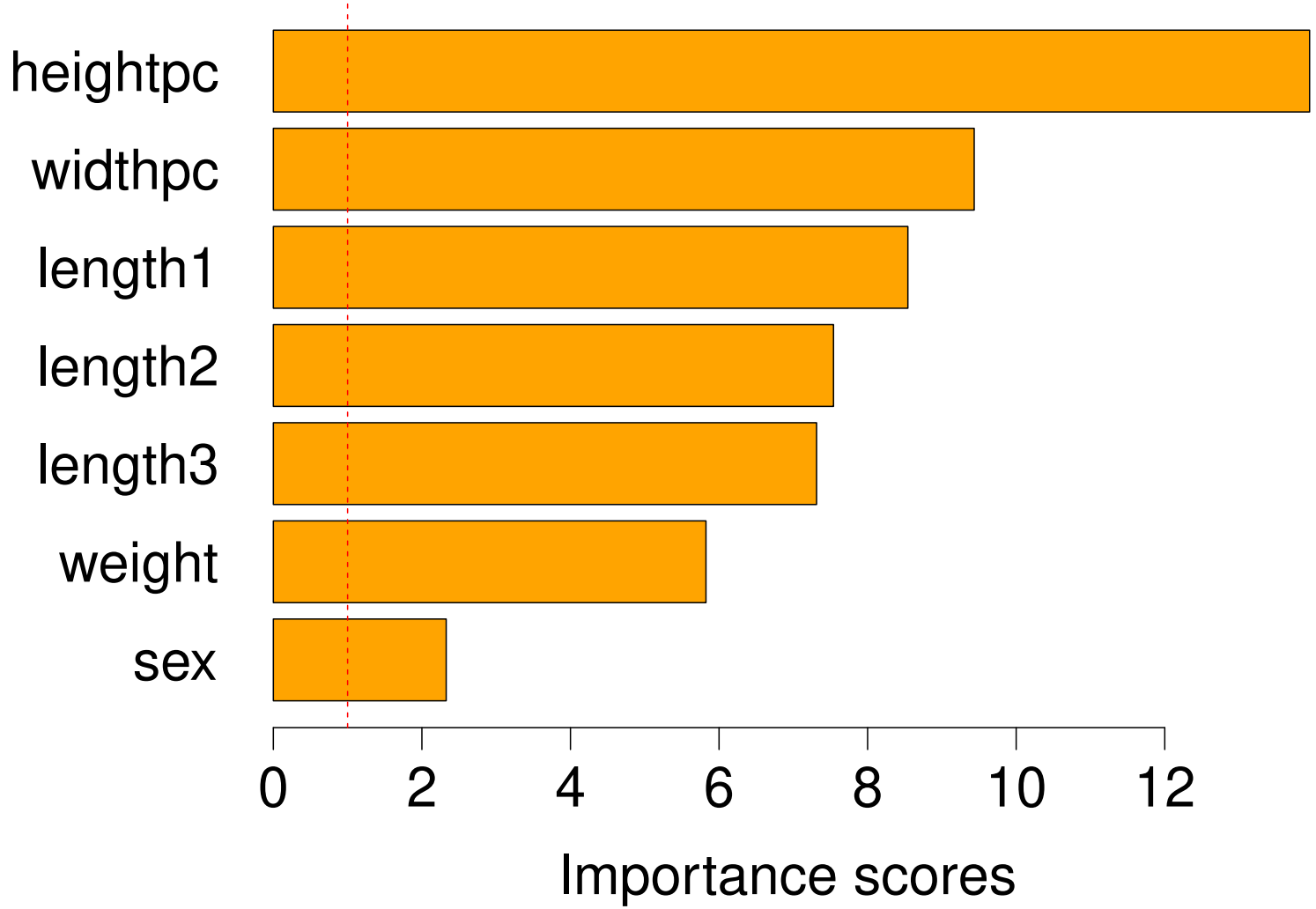
GUIDE tree with 2nd best split variable at root node



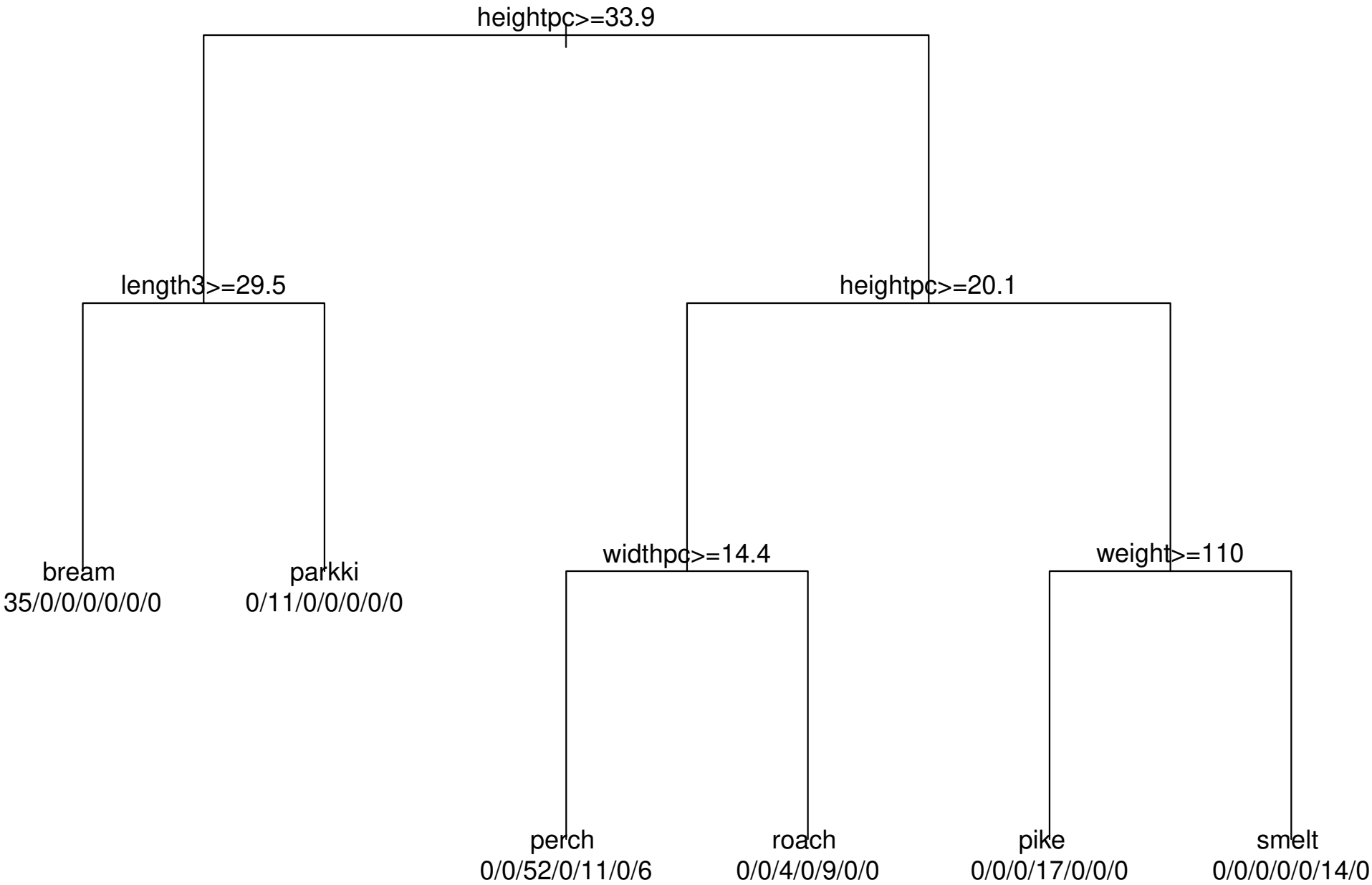
GUIDE linear splits



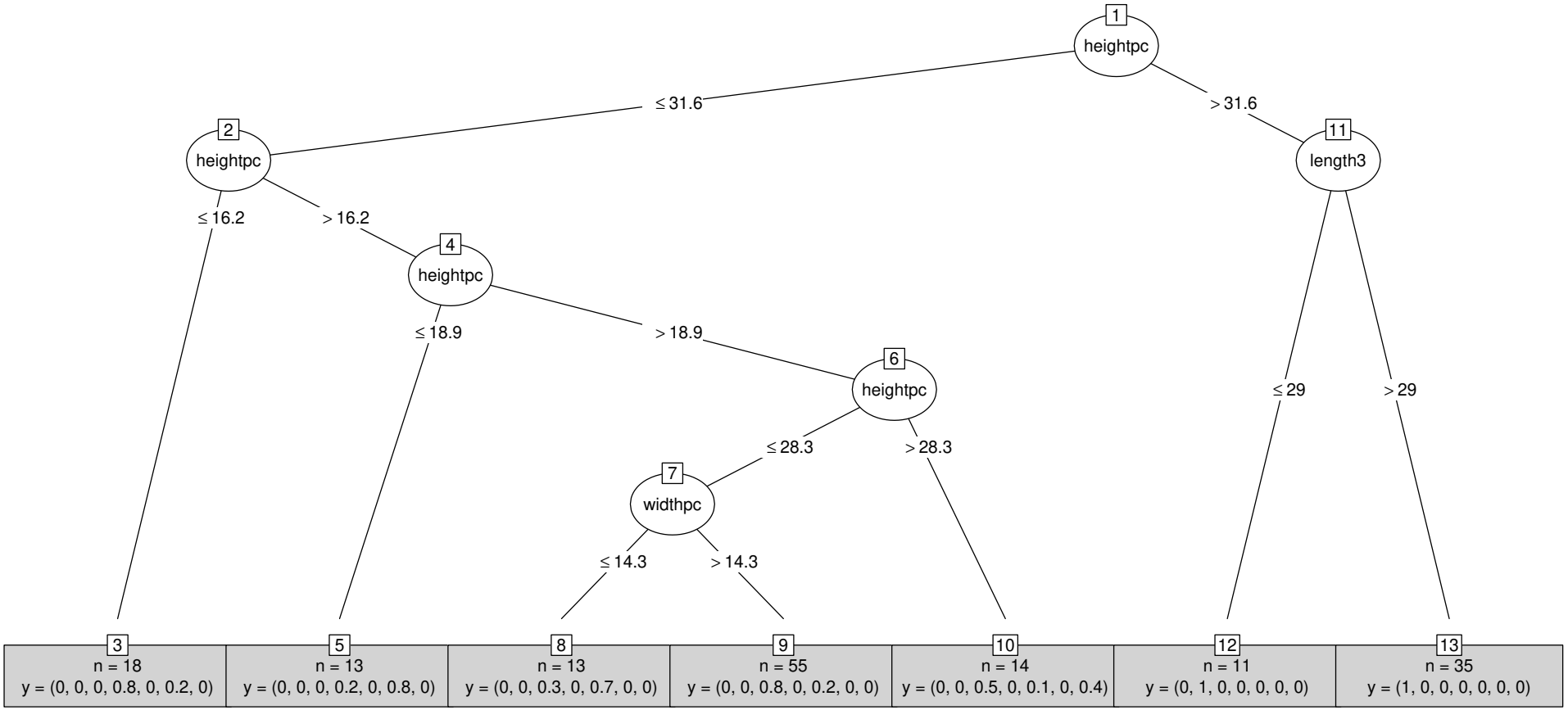
Importance scores for fish data



RPART tree for fish data



Ctree for fish data



American Community Survey (ACS) data

- Nationwide survey that collects information on social, economic, housing, and demographic characteristics about the US population every year
- Information provides an important tool for communities to use to see how they are changing
- When people fill out the ACS form, they help to ensure that decisions about the future of their community can be made using the best data available
- Decision-makers require a clear picture of their population so that scarce resources can be allocated efficiently and effectively
- Every year, the Census Bureau contacts over 3.5 million households across the country to participate in the ACS

Homework 5 (ACS data)
due in Canvas by 9:30AM, Thu Mar 25, 2021

1. Download the ACS data set assigned to you
2. Read the files in the ACS Data folder on Canvas and give an alphabetical list of the names of the variables and the reasons that they should not be used to estimate the state population mean of INTP

Course project

AIM: Use PERSON RECORD variables (from p. 30 of ACS PUMS Data Dictionary) to estimate mean μ of INTP (interest, dividends, and rental income) with sampling weight w_i in variable PWGTP

TASKS:

1. Use GUIDE tree and forest to estimate π (probability of response), then use the inverse probability weighted (IPW) method to estimate μ
2. Use GUIDE tree and forest to predict missing INTP ($\hat{y}$), then estimate μ with

$$\Big(\sum_{i\in S} w_i y_i + \sum_{j\in \bar{S}} w_j \hat{y}_j\Big) / \sum_i w_{i\in S\cup\bar{S}} \tag{1}$$

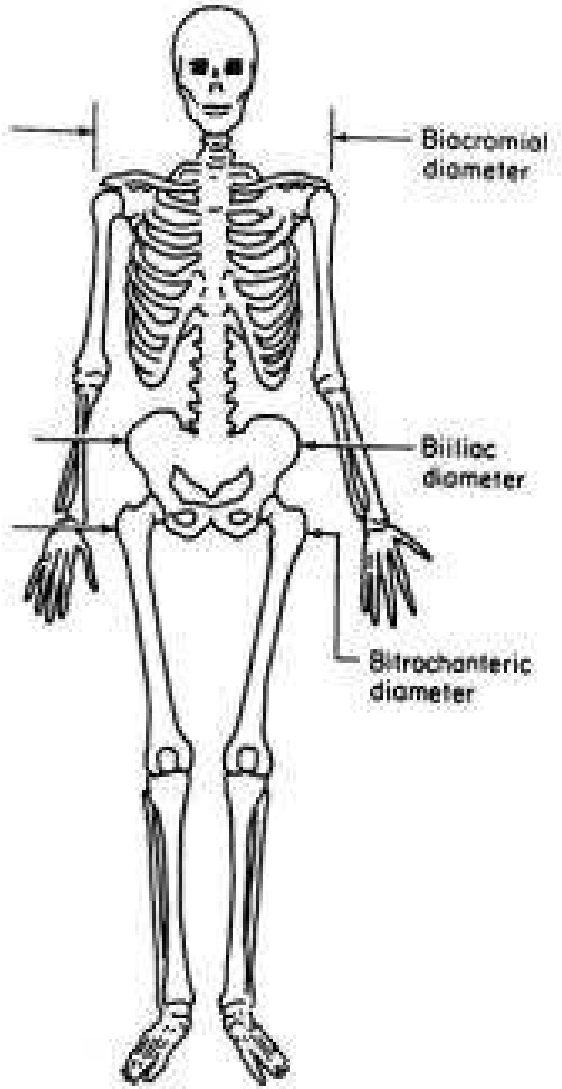
   where S observations with non-missing INTP and $\bar{S}$ is its complement
3. Repeat Tasks 1 & 2 with RPART and Ctree (`party`)
4. Use one or more methods to impute missing values in the X variables
5. Use logistic regression to find $\hat{\pi}$ and IPW to estimate μ
6. Use randomForest and Cforest (`party`) to obtain $\hat{y}$ and estimate μ with (1)
7. Comment on your methods and results

Classification with continuous predictors: predicting gender from body measurements

- Nine skeletal diameter measurements (cm) and twelve girth measurements (cm) were obtained from 507 physically active individuals
- 260 females (coded 0) and 247 males (coded 1)
- Age (years), weight (kg), and height (cm) included
- `www.amstat.org/publications/jse/v11n2/datasets.heinz.html`

Predictor variables

Skeletal measurements (cm)	Girth measurements (cm)	Others
Biacromial diameter	Shoulder girth	Age (years)
Biiliac diameter (pelvic breadth)	Chest girth	Weight (kg)
Bitrochanteric diameter	Waist girth	Height (cm)
Chest depth	Navel (abdominal) girth	
Chest diameter	Hip girth	
Sum of elbow diameters	Thigh girth	
Sum of wrist diameters	Bicep girth	
Sum of knee diameters	Forearm girth	
Sum of ankle diameters	Knee girth	
	Calf maximum girth	
	Ankle minimum girth	
	Wrist minimum girth	


Biacromial diameter
Biiliac diameter
Bitrochanteric diameter

Linear discriminant analysis

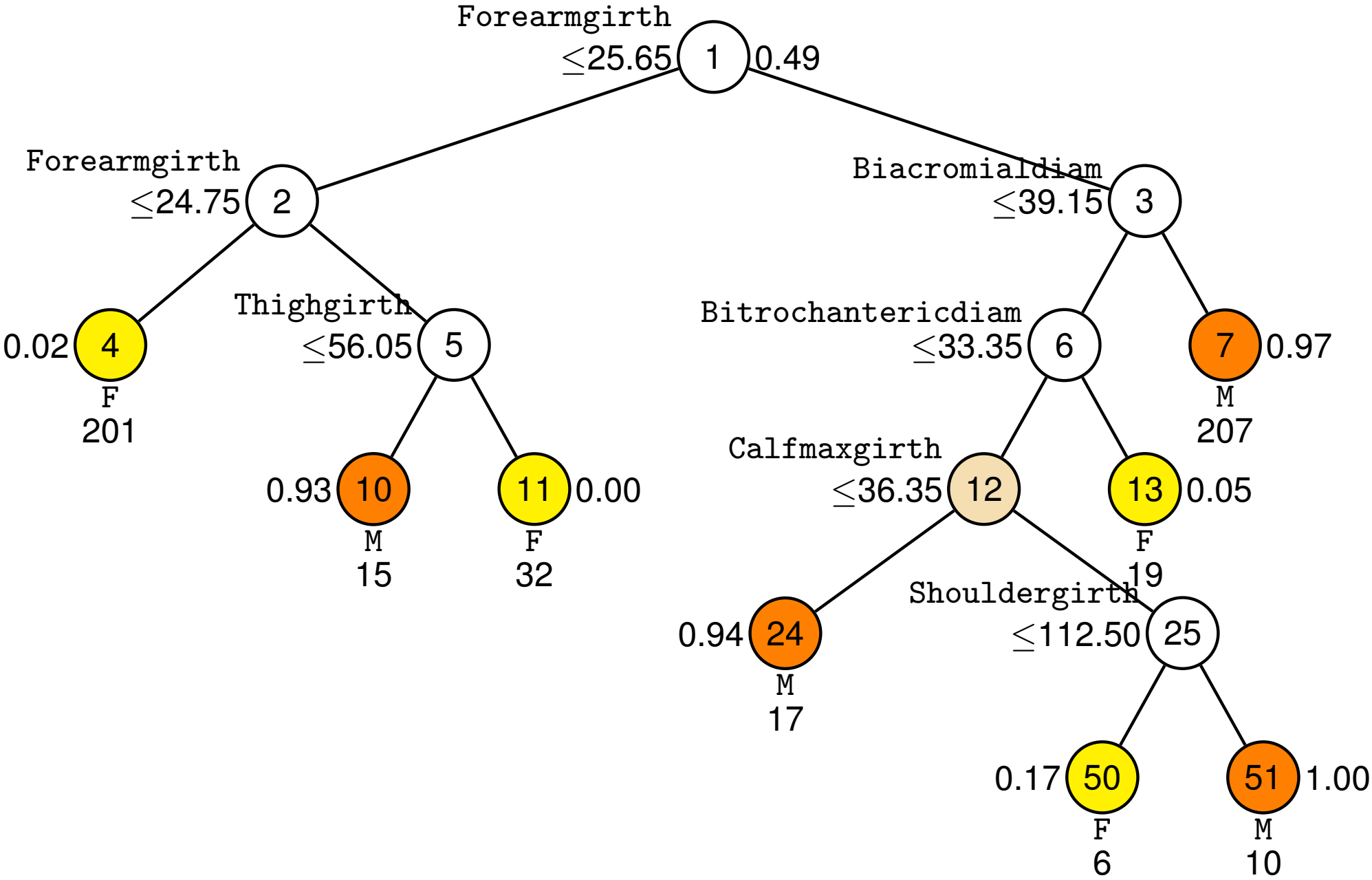
	Observed	
Predicted	Female	Male
Female	257	4
Male	3	243

SAS stepwise LDA

Step	Entered	Removed	ASCC	P(> ASCC)
1	Forearm girth		0.62710004	<.0001
2	Thigh girth		0.76754489	<.0001
3	Biacromial diameter		0.80478106	<.0001
4	Waist girth		0.82694544	<.0001
5	Navel girth		0.84936781	<.0001
6	Ankle diameter		0.85689188	<.0001
7	Bicep girth		0.86206794	<.0001
8	Height		0.86671114	<.0001
9	Weight		0.87025435	<.0001
10	Bitrochanteric diameter		0.87201627	<.0001
11	Elbow diameter		0.87309831	<.0001
12	Chest depth		0.87405146	<.0001

ASCC = average square canonical correlation

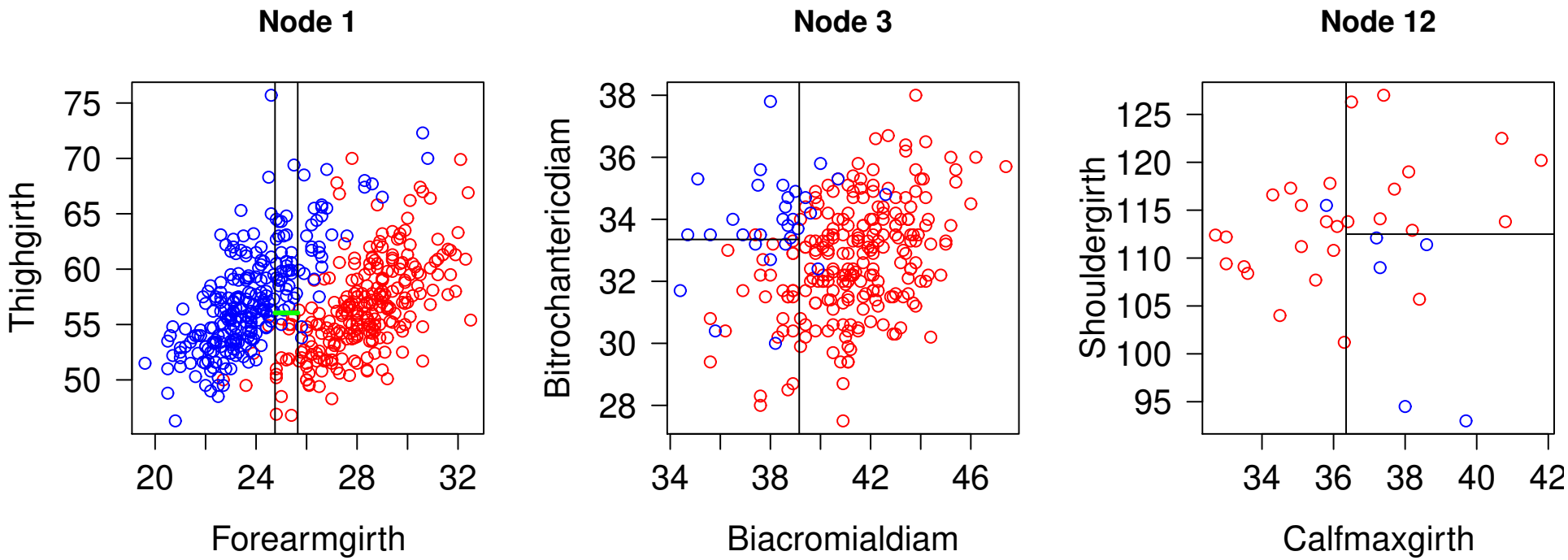
GUIDE default univariate splits



Chi-squared tests at root node

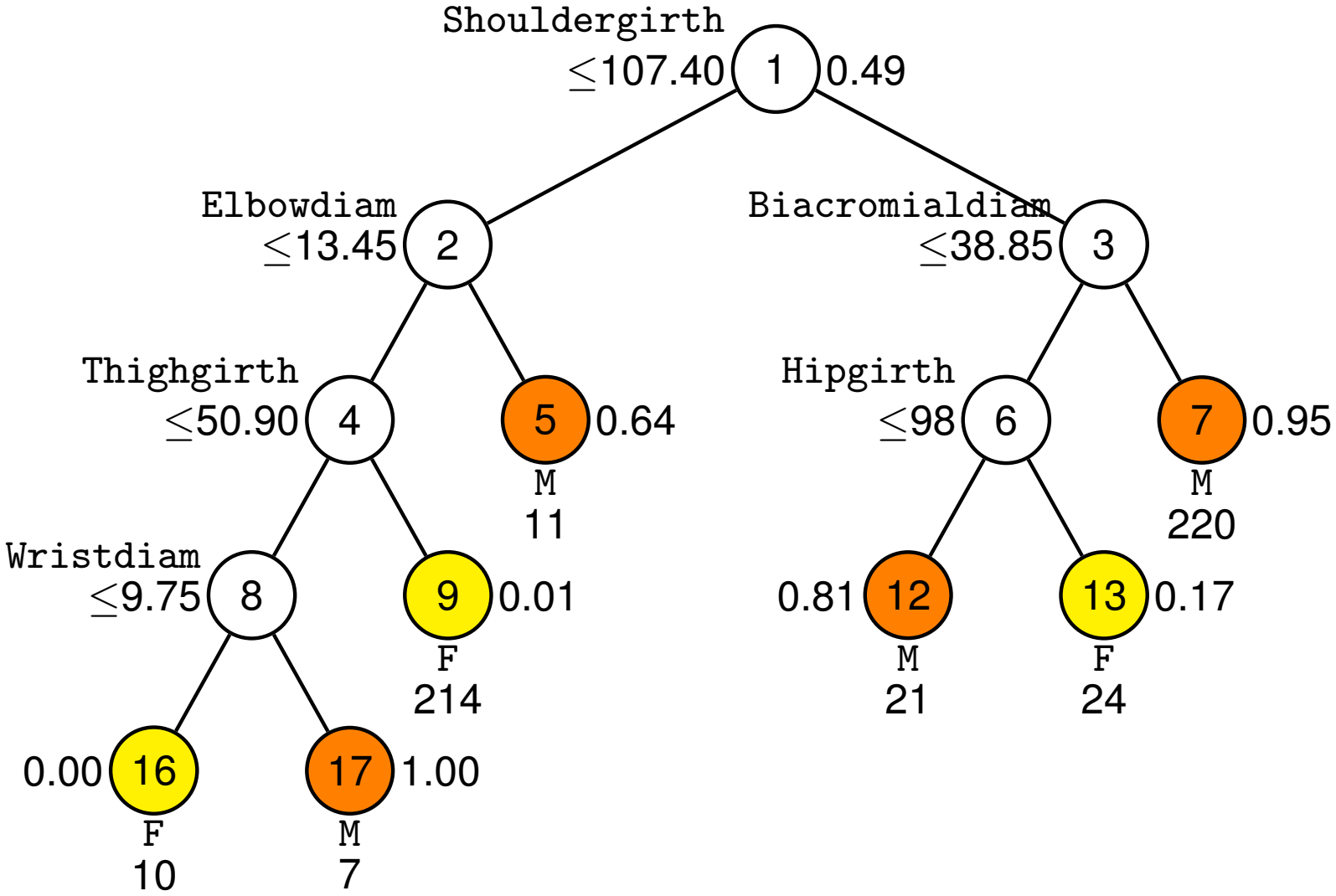
	Forearmgirth			
	$\leq$ 23.6	(23.6, 25.8]	(25.8, 28.4]	$>$ 28.4
F	133	100	23	4
M	2	20	106	119
	χ^2_3 = 341.3, p-value $<$ 2.2e-16, χ^2_1 = 309.7			

	Shouldergirth			
	$\leq$ 99.5	(99.5, 108]	(108, 117]	$>$ 117
F	127	103	26	4
M	0	25	99	123
	χ^2_3 = 328.6, p-value $<$ 2.2e-16, χ^2_1 = 297.6			

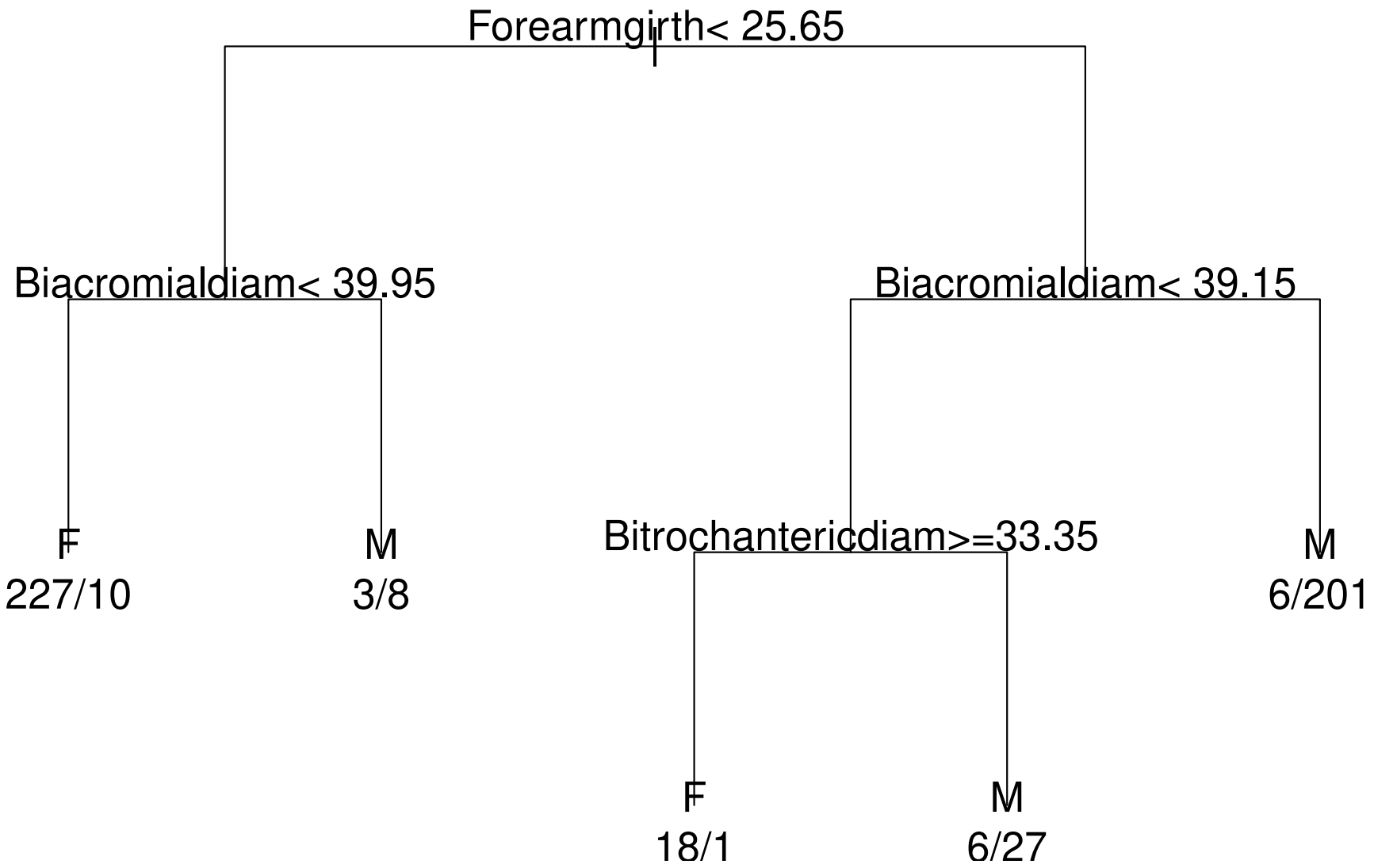

Node 1
Thighgirth
75
70
65
60
55
50
20
24
28
32
Forearmgirth
Node 3
Bitrochantericdiam
38
36
34
32
30
28
34
38
42
46
Biacromialdiam
Node 12
Shouldergirth
125
120
115
110
105
100
95
34
36
38
40
42
Calfmaxgirth

Popeye cartoon character

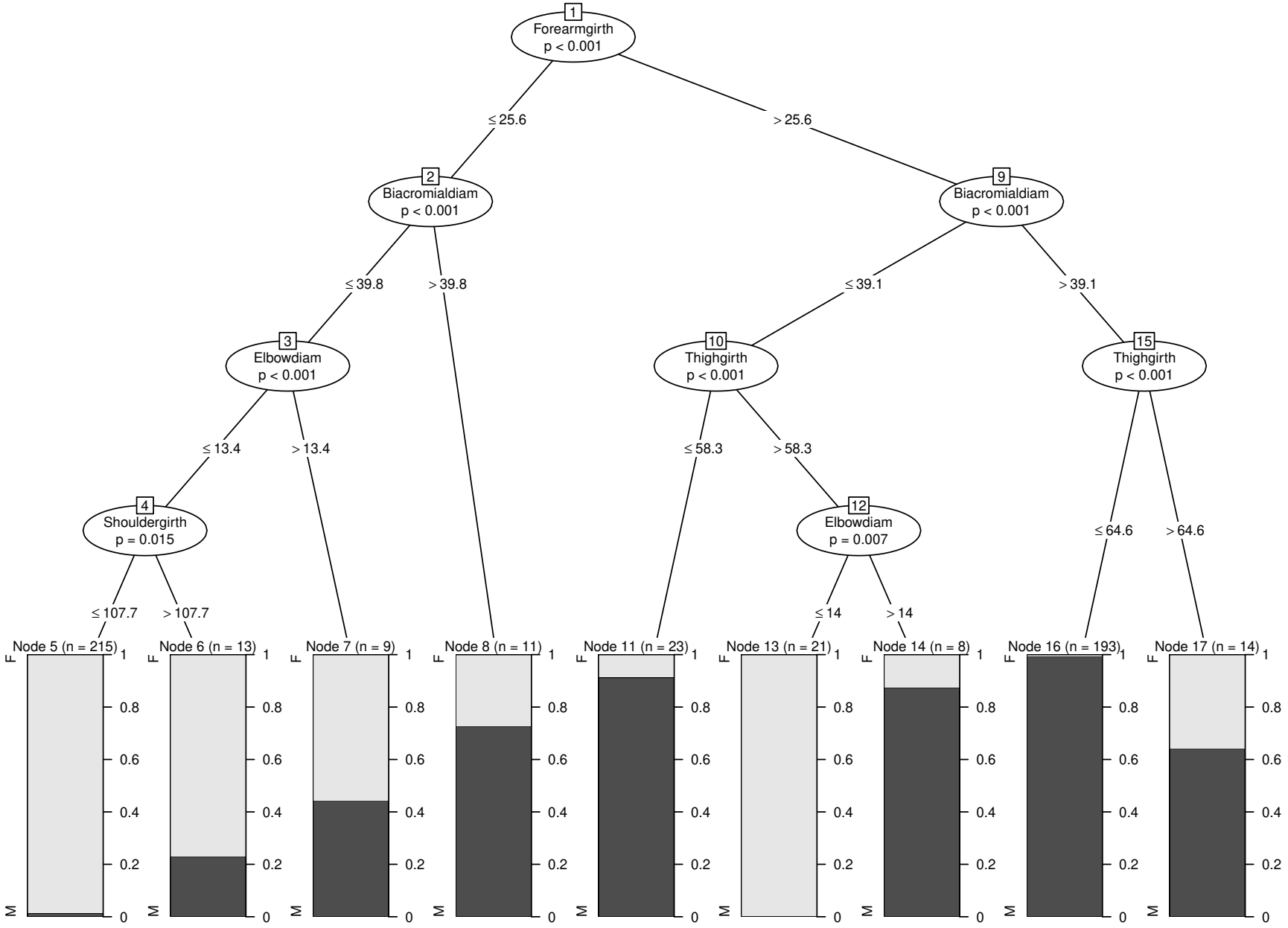
GUIDE univariate tree with 2nd best variable



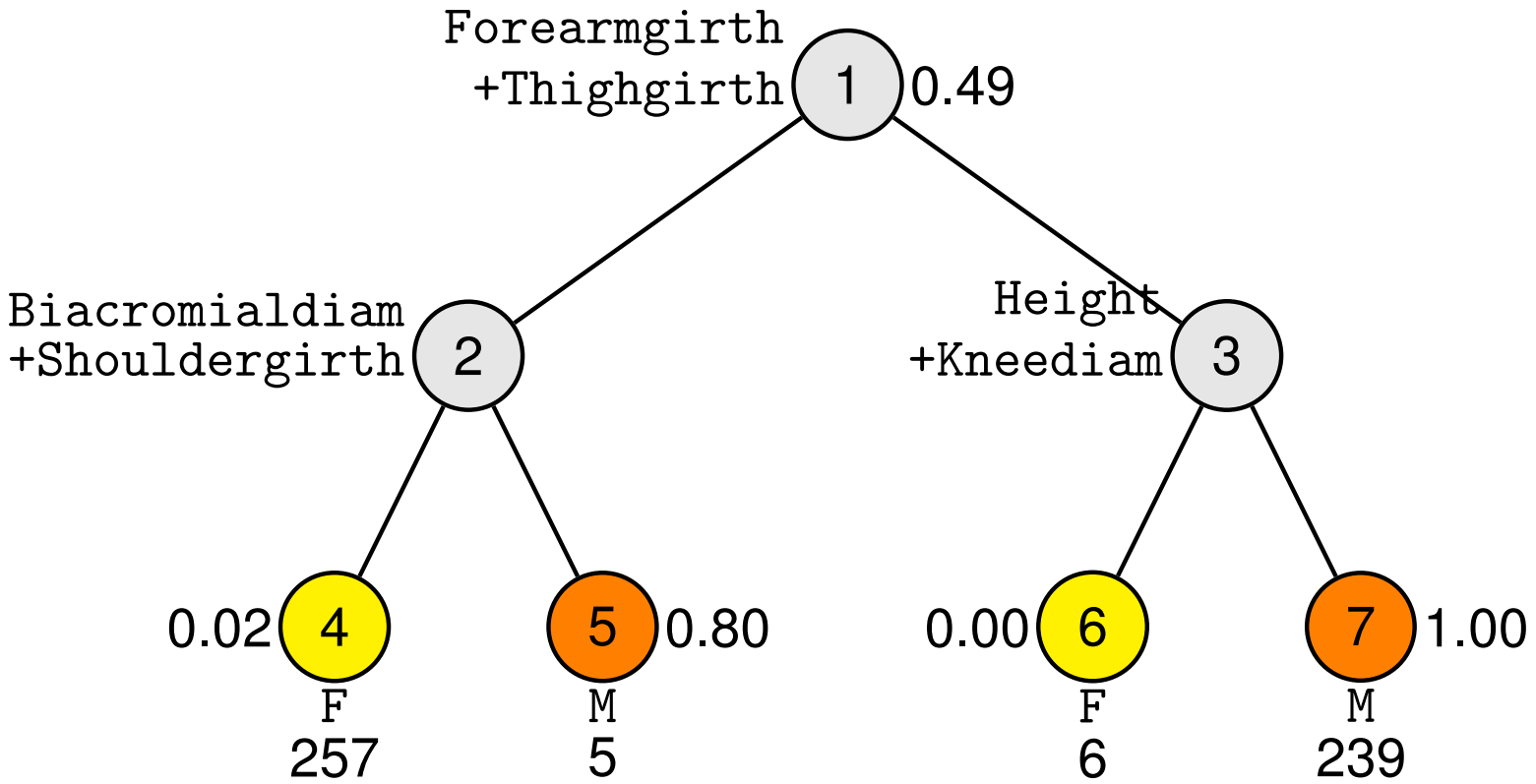
RPART tree for body data



Ctree tree for body data



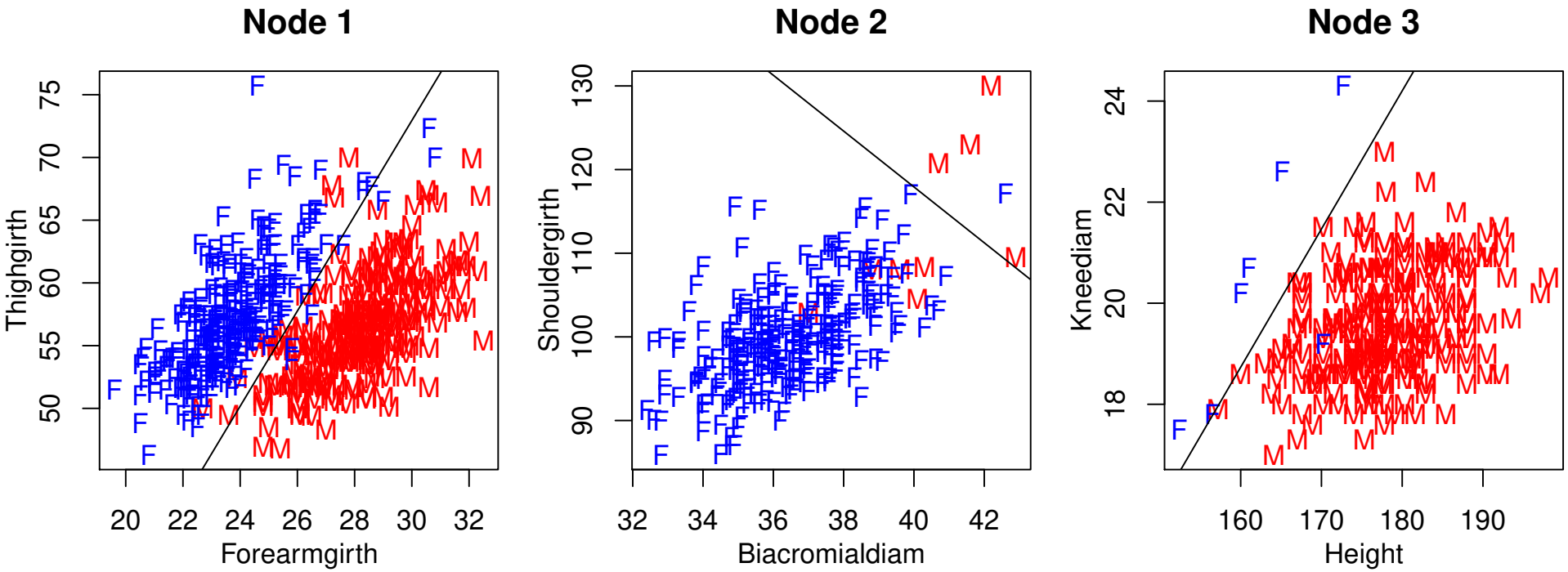
GUIDE tree with linear splits



Predicted classes and sample sizes printed below terminal nodes

Sample proportion of males beside nodes

Nodes of GUIDE linear split tree



Leave-one-out error counts for peptide data

Method	Errors[a]	Time[b]
GUIDE forest	20	5.353
RPART	33	0.009
RandomForest	34	0.078
GUIDE tree, 2nd best univ. split at root node	36	0.291
GUIDE tree, default univ. splits	45	0.284
GUIDE tree, bivariate NN	54	0.792
Logistic regression[c]	-	-
Ctree, Cforest, LDA[d]	-	-

[a] out of 310 observations
[b] average time (sec.) to fit one data set
[c] does not converge
[d] cannot predict observations with categorical values not in training data

Leave-one-out error counts for fish data

Method	Errors[a]	Time[b]
Multinomial logistic regression[c]	2	0.024
LDA	3	0.004
GUIDE tree, univ1	13	0.118
GUIDE tree, linear splits	14	0.152
GUIDE forest	15	6.031
GUIDE tree, univ2	18	0.128
GUIDE tree, bivariate NN	21	0.959
RPART	23	0.009
Cforest	24	12.344
GUIDE tree, bivariate kernel	25	0.951
Ctree	28	0.059
RandomForest	28	0.075

[a] out of 157 observations, excluding 2 with missing weight

[b] average time (sec.) to fit one data set

[c] multinom function in nnet package

Leave-one-out error counts for body data

Method	Count[a]	Time[b]
LDA	9	0.007
GUIDE forest	18	8.828
RandomForest	19	0.261
GUIDE tree, linear splits	20	1.361
Cforest	27	12.329
GUIDE tree, univ1 splits	33	0.914
RPART	35	0.021
Ctree	39	0.096
GUIDE tree, univ2 splits	46	0.816
Logistic regression[c]	-	-

[a] out of 507 observations
[b] average time (sec.) to fit one data set
[c] does not converge

Multiple imputation with MICE

```
> library(mice)
> z <- read.table(file="fish.dat",header=TRUE,stringsAsFactors=TRUE)
> names(z)
[1] "species" "weight" "length1" "length2" "length3" "heightpc" "widthpc"
[8] "sex"
> y <- z$species  ## save a copy of species variable
> z$sex[z$sex == "unknown"] <- NA  ## make "unknown" level NA
> z <- droplevels(z)  ## drop unused "unknown" factor level
> out <- mice(z[,-1])   ## exclude y variable
> c1 <- complete(out,1) ## 1st imputed data set
> c2 <- complete(out,2) ## 2nd imputed data set
> c3 <- complete(out,3) ## 3rd imputed data set
> c4 <- complete(out,4) ## 4th imputed data set
> c5 <- complete(out,5) ## 5th imputed data set
```

MICE (cont'd.)

```
> table(z$sex,y)  ## data before imputation
         y
          bream parkki perch pike roach smelt whitefish
  female      3       4    25    5     8     9          1
  male        6       3     2    1     0     5          0
  unknown    26       4    29   11    12     0          5
> table(c1$sex,y)
        y
         bream parkki perch pike roach smelt whitefish
  female    19       7    49    9    18     9          6
  male      16       4     7    8     2     5          0
> table(c2$sex,y)
        y
         bream parkki perch pike roach smelt whitefish
  female    20       4    46   16    17     9          6
  male      15       7    10    1     3     5          0
```

```
> table(c3$sex,y)
        y
         bream parkki perch pike roach smelt whitefish
  female    14      6    52   15    16     9          4
  male      21      5     4    2     4     5          2
> table(c4$sex,y)
        y
         bream parkki perch pike roach smelt whitefish
  female    13      5    46   12    18     9          4
  male      22      6    10    5     2     5          2
> table(c5$sex,y)
        y
         bream parkki perch pike roach smelt whitefish
  female    13      6    48   16    18     9          5
  male      22      5     8    1     2     5          1
```

Species prediction if sex is unknown

Sex	Bream	Parkki	Perch	Pike	Roach	Smelt	White	
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
			Imputed set 1					
female	19	7	49	9	18	9	6	117
male	16	4	7	8	2	5	0	42
			Imputed set 2					
female	20	4	46	16	17	9	6	118
male	15	7	10	1	3	5	0	41
			Imputed set 3					
female	14	6	52	15	16	9	4	116
male	21	5	4	2	4	5	2	43

Multiple imputation with AMELIA

```
> library(Amelia)
> z <- read.table(file="fish.dat",header=TRUE,stringsAsFactors=TRUE)
> y <- z$species  ## save a copy of y (species) variable
> z$sex[z$sex == "unknown"] <- NA  ## make "unknown" level NA
> z <- droplevels(z)  ## drop unused "unknown" factor level
> x <- z[,-1]  ## dataframe of predictor variables
> out <- amelia(x,noms=7)  ## 7th column of x is nominal
> c1 <- out$imputations$imp1  ## 1st imputed x matrix
> c2 <- out$imputations$imp2  ## 2nd imputed x matrix
> c3 <- out$imputations$imp3  ## 3rd imputed x matrix
> table(z$sex,y)
         y
          bream parkki perch pike roach smelt whitefish
  female      3      4    25    5     8     9          1
  male        6      3     2    1     0     5          0
  unknown    26      4    29   11    12     0          5
```

AMELIA (cont'd.)

```
> table(c1$sex,y)
        y
         bream parkki perch pike roach smelt whitefish
  female    13       7    47   13    16     9          3
  male      22       4     9    4     4     5          3
> table(c2$sex,y)
        y
         bream parkki perch pike roach smelt whitefish
  female    11       5    45   16    14     9          4
  male      24       6    11    1     6     5          2
> table(c3$sex,y)
        y
         bream parkki perch pike roach smelt whitefish
  female    21       8    44   12    17     9          3
  male      14       3    12    5     3     5          3
```

Multinomial logistic regression on fish data

```
> library(nnet)
> z <- na.omit(z)  ## remove observations with NAs
> mul <- multinom(species ~ ., data=z)
> predicted <- predict(mul,newdata=z,type="class")
> table(z$species,predicted)
            predicted
             bream parkki perch pike roach smelt whitefish
  bream         34      0     0    0     0     0          0
  parkki         0     11     0    0     0     0          0
  perch          0      0    56    0     0     0          0
  pike           0      0     0   17     0     0          0
  roach          0      0     0    0    19     0          0
  smelt          0      0     0    0     0    14          0
  whitefish      0      0     0    0     0     0          6
```

Homework 6 (due by 9:30am, April 8, 2021) (CV after imputation for fish data; see slide 173)

1. Use Amelia to impute missing values in sex and weight variables
2. Leaving one observation out at a time:
   (a) Apply LDA to remaining $(n-1)$ observations in each imputed data set
   (b) Obtain the average (over the 5 imputed data sets) prediction error for the left-out observation
3. Average the prediction errors over the n observations
4. Repeat steps 1–3 with multinomial logistic regression in place of LDA
5. Repeat steps 1–4 with mice in place of Amelia

ACS data: variables with no predictive power

Name	Definition (reason)
RT	Record type (constant)
SERIALNO	ID
SPORDER	SERIALNO and SPORDER uniquely identify subjects
DIVISION	Division of US (constant)
REGION	Region of US (constant)
ST	State code (constant)
ADJINC	Adjustment factor for dollar amounts (constant)
PWGTP1-80	Replicate weights for confidence interval estimation

Preparing ACS data for analysis

1. Remove variables with no predictive power from csv file
2. Remove all flag variables except FINTP
3. Import csv into R with `z = read.csv("psam_pxx.csv",na.strings="")`
4. Export cleaned data with
   `write.table(z,"cleandata.txt",row.names=FALSE,col.names=TRUE)`
5. Create GUIDE description file (see slide 57)
6. For later analysis with R methods, import cleaned data into R with
   `z = read.table("cleandata.txt",header=TRUE)`
   Important: remember to set categorical variables as factor variables in R

About the final project

- The project is equivalent to a final exam and should be treated as such
- Do not discuss the project with anyone
- You may ask general questions about the data and project in class
- If you seek help outside class from the instructor or TA, points will be deducted from your project grade

Node impurity measures for classification

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

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

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

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

- Examples are:

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

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

If $J = 2$, then $i(t) = 2p(1|t)p(2|t)$ = 2 $\times$ binomial variance

CART split set selection (Breiman et al., 1984)

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 set $\mathcal{S}$ of splits of the form $X \in A$, where X is a predictor and

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

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

Weaknesses and limitations of CART (and RPART)

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

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

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

Consequently CART has selection bias:

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

and these operational constraints:

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

Predicting drive train of cars

- 428 cars and 13 variables (2 categorical, 11 ordered)
- Drive train takes three values:
  - 94 (22%) four-wheel (4wd)
  - 224 (52%) front-wheel (Fwd)
  - 110 (25%) rear-wheel (Rwd)
- No missing values
- Only one Hummer

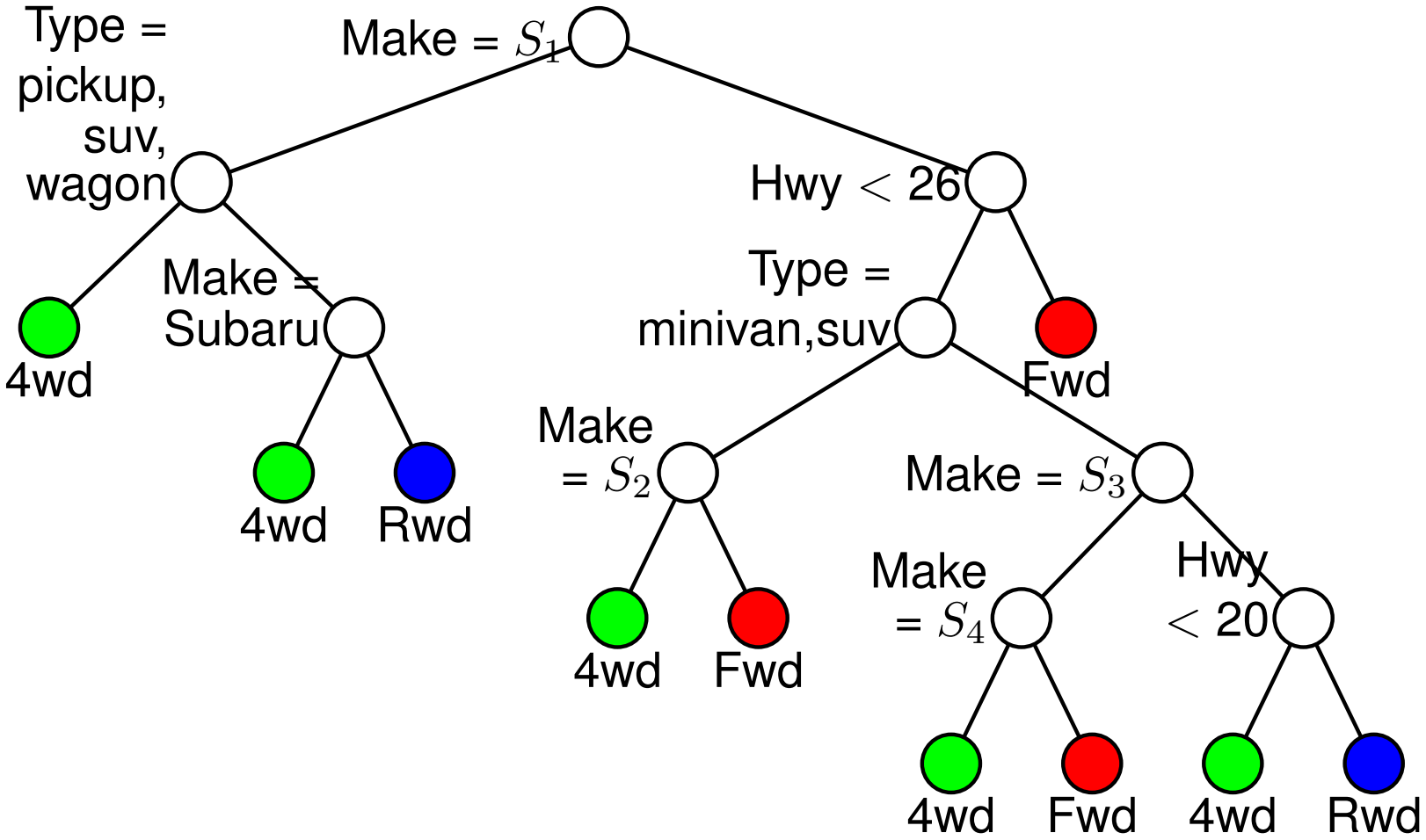
Predictor variables

Variable	Description	Variable	Description
Make	Make of car (38 values)	City	City miles/gallon
Type	Type of car (6 values)	Hwy	Highway miles/gallon
Rprice	Suggested retail price	Weight	Weight (pounds)
Dcost	Dealer cost	Whlbase	Wheel base (in.)
Enginsz	Engine size (liters)	Length	Length (in.)
Cylin	Number of cylinders	Width	Width (in.)
Hp	Horsepower		

Make has $(2^{38-1} - 1) \approx 10^{11}$ = 100 billion splits!

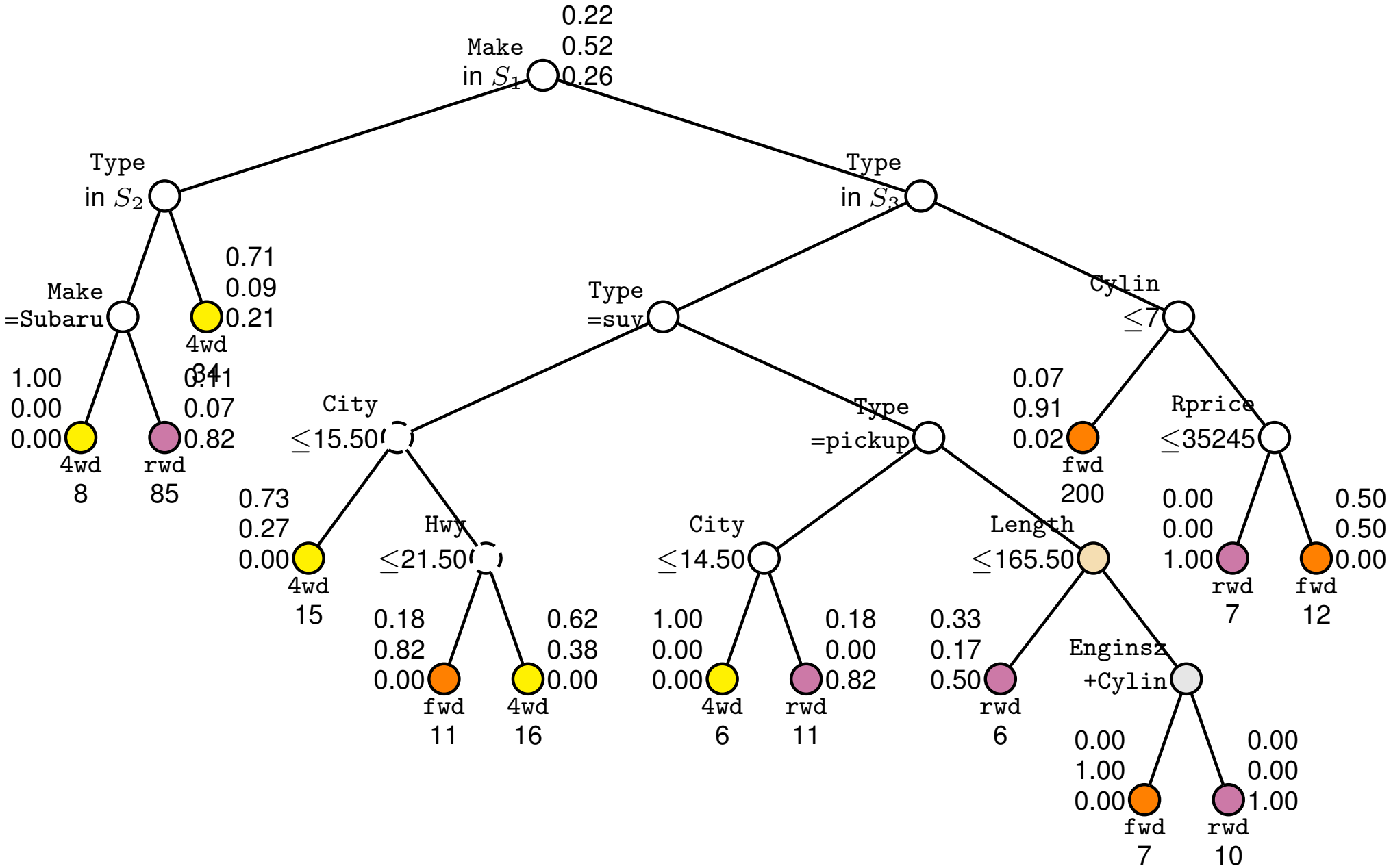
`party` accepts categorical variables with $>$ 31 levels, but not `partykit`

RPART tree for car data (36 cpu hrs!)



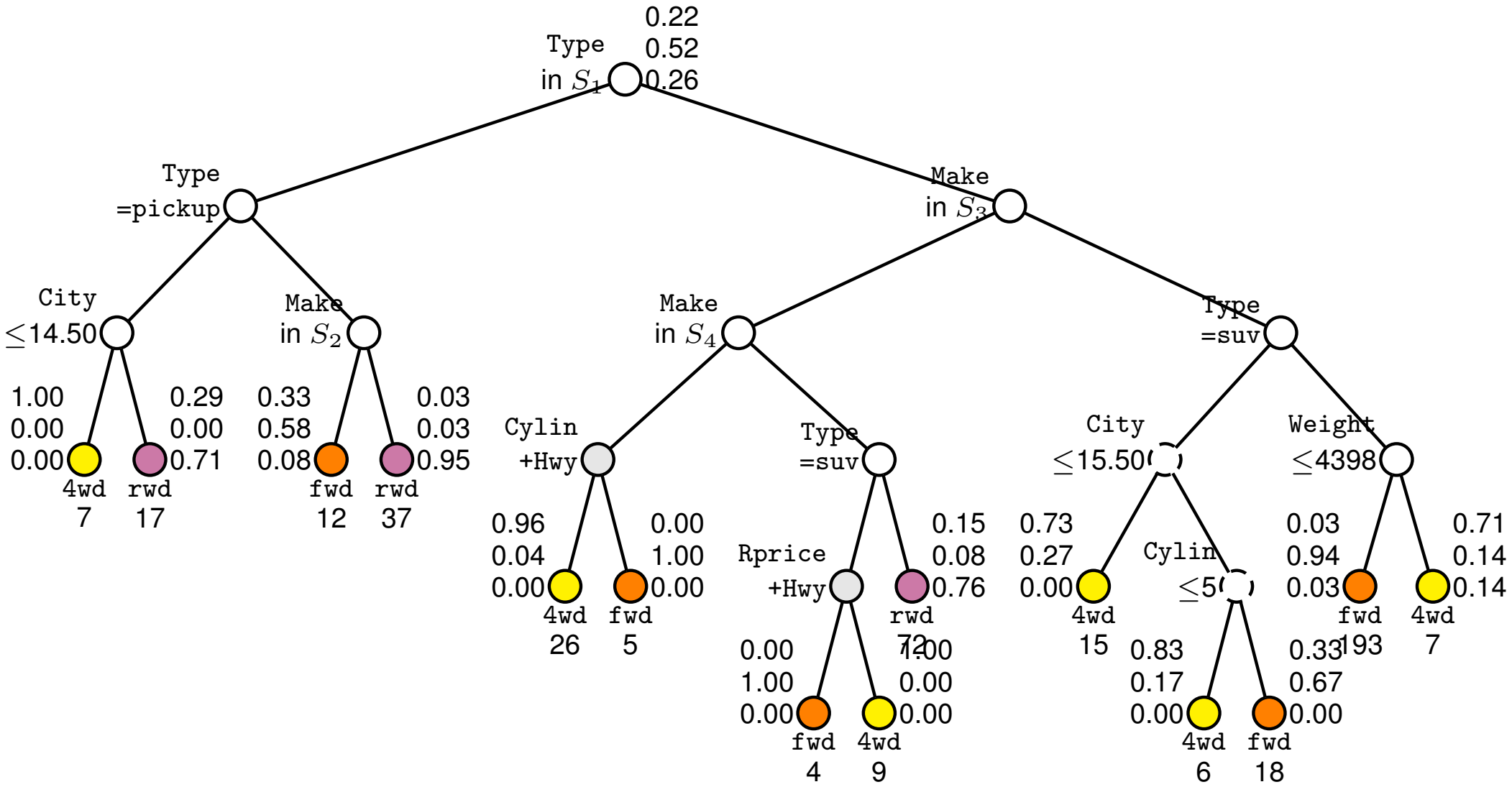
S_1 = {BMW, GMC, Hummer, Infiniti, Jaguar, Land-Rover, Lexus, Lincoln, Mazda, Mercedes, Porsche, Subaru}; S_2 = {Acura, Buick, Chevrolet, Dodge, Ford, Honda, Isuzu, Jeep, Mitsubishi, Pontiac, Suzuki, Toyota, Volkswagen, Volvo}; S_3 = {Audi, Kia, Mitsubishi, Nissan, Pontiac, Volkswagen, Volvo}; S_4 = {Audi, Nissan, Volvo}.

GUIDE tree for car data (0.5 sec.)



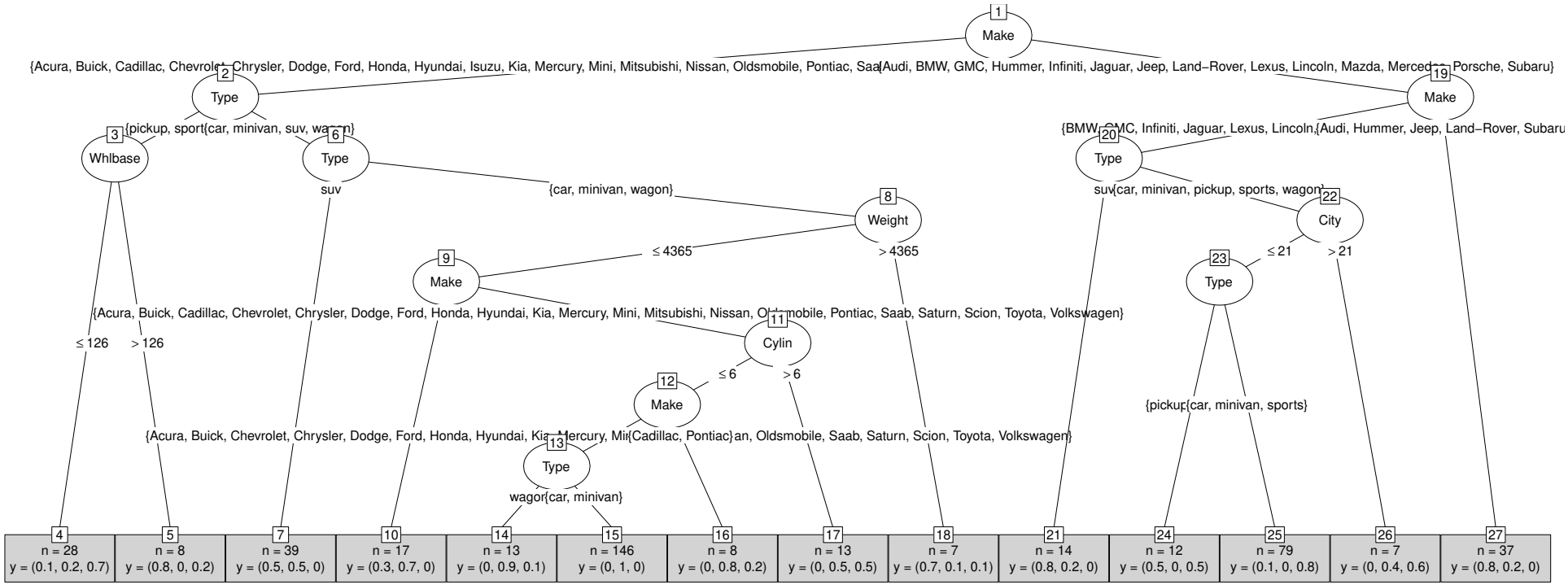
S_1 = {BMW, GMC, Hummer, Infiniti, Jaguar, Land-Rover, Lexus, Lincoln, Mazda, Mercedes, Porsche, Subaru}; S_2 = {car, minivan, sports}; S_3 = {pickup, sports, suv}

GUIDE tree using 2nd best variable at root node



Set S_1 = {pickup, sports}. Set S_2 = {Audi, Hyundai, Mitsubishi, Subaru, Toyota}. Set S_3 = {Audi, BMW, GMC, Hummer, Infiniti, Jaguar, Jeep, Land-Rover, Lexus, Lincoln, Mercedes, Mercury, Porsche, Subaru}. Set S_4 = {Audi, Hummer, Jeep, Land-Rover, Porsche, Subaru}.

Ctree (`party`) tree for car data



`partykit` inapplicable because `Make` has more than 31 categories

Ctree in text form

```
1) Make == {Acura, Buick, Cadillac, Chevrolet, Chrysler, Dodge, Ford, Honda,
            Hyundai, Isuzu, Kia, Mercury, Mini, Mitsubishi, Nissan, Oldsmobile,
            Pontiac, Saab, Saturn, Scion, Suzuki, Toyota, Volkswagen, Volvo};
            criterion = 1, statistic = 306.126
  2) Type == {pickup, sports}; criterion = 1, statistic = 192.925
    3) Whlbase <= 126; criterion = 0.997, statistic = 30.12
      4)*  weights = 28
    3) Whlbase > 126
      5)*  weights = 8
  2) Type == {car, minivan, suv, wagon}
    6) Type == {suv}; criterion = 1, statistic = 80.292
      7)*  weights = 39
    6) Type == {car, minivan, wagon}
      8) Weight <= 4365; criterion = 1, statistic = 87.261
        9) Make == {Suzuki, Volvo}; criterion = 1, statistic = 114.634
          10)*  weights = 17
        9) Make == {Acura, Buick, Cadillac, Chevrolet, Chrysler, Dodge, Ford,
                    Honda, Hyundai, Kia, Mercury, Mini, Mitsubishi, Nissan,
                    Oldsmobile, Pontiac, Saab, Saturn, Scion, Toyota, Volkswagen}
```

```
          11) Cylin <= 6; criterion = 1, statistic = 49.5
            12) Make == {Acura, Buick, Chevrolet, Chrysler, Dodge, Ford, Honda,
                         Hyundai, Kia, Mercury, Mini, Mitsubishi, Nissan,
                         Oldsmobile, Saab, Saturn, Scion, Toyota, Volkswagen};
                         criterion = 1, statistic = 66.993
              13) Type == {wagon}; criterion = 0.954, statistic = 14.9
                14)*  weights = 13
              13) Type == {car, minivan}
                15)*  weights = 146
            12) Make == {Cadillac, Pontiac}
              16)*  weights = 8
          11) Cylin > 6
            17)*  weights = 13
        8) Weight > 4365
          18)*  weights = 7
1) Make == {Audi, BMW, GMC, Hummer, Infiniti, Jaguar, Jeep, Land-Rover, Lexus,
            Lincoln, Mazda, Mercedes, Porsche, Subaru}
  19) Make == {BMW, GMC, Infiniti, Jaguar, Lexus, Lincoln, Mazda, Mercedes,
               Porsche}; criterion = 1, statistic = 84.578
    20) Type == {suv}; criterion = 1, statistic = 51.574
```

```
    21)*  weights = 14
  20) Type == {car, minivan, pickup, sports, wagon}
    22) City <= 21; criterion = 0.999, statistic = 31.116
      23) Type == {pickup, wagon}; criterion = 0.992, statistic = 27.477
        24)*  weights = 12
      23) Type == {car, minivan, sports}
        25)*  weights = 79
    22) City > 21
      26)*  weights = 7
19) Make == {Audi, Hummer, Jeep, Land-Rover, Subaru}
  27)*  weights = 37
```

Leave-one-out error counts for car data

Method	Errors[a]	Time[b]
GUIDE forest	75	23.20
Cforest (party)	80	12.83
LDA	84	0.02
GUIDE tree, linear splits	95	0.73
GUIDE tree, univ1 splits	99	0.62
GUIDE tree, univ2 splits	110	0.59
Ctree (party)	111	0.07
Multinomial logistic regression[c]	-	-
RPART[d]	-	-
RandomForest and partykit[e]	-	-

[a] out of 427 obs, excluding single Hummer observation which crashes LDA and Cforest
[b] average time (sec.) to fit one data set
[c] does not converge
[d] takes too long
[e] RandomForest and partykit cannot take categorical variables with more than 31 levels

GUIDE classification

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

GUIDE hierarchical split variable selection

Level 1: Marginal tests. Cross-tab each X with Y, including a level for NA if present in X. Select X with smallest p-value if its $p < 0.10/K$, where K is number of X variables. Otherwise, go to level 2.

Level 2: Interaction tests. For each pair (i, j), divide (X_i, X_j)-space into several regions. Cross-tab regions with Y. Select (X_i, X_j) with smallest p-value if its $p < 0.20/\{K(K-1)\}$. Otherwise go to level 3.

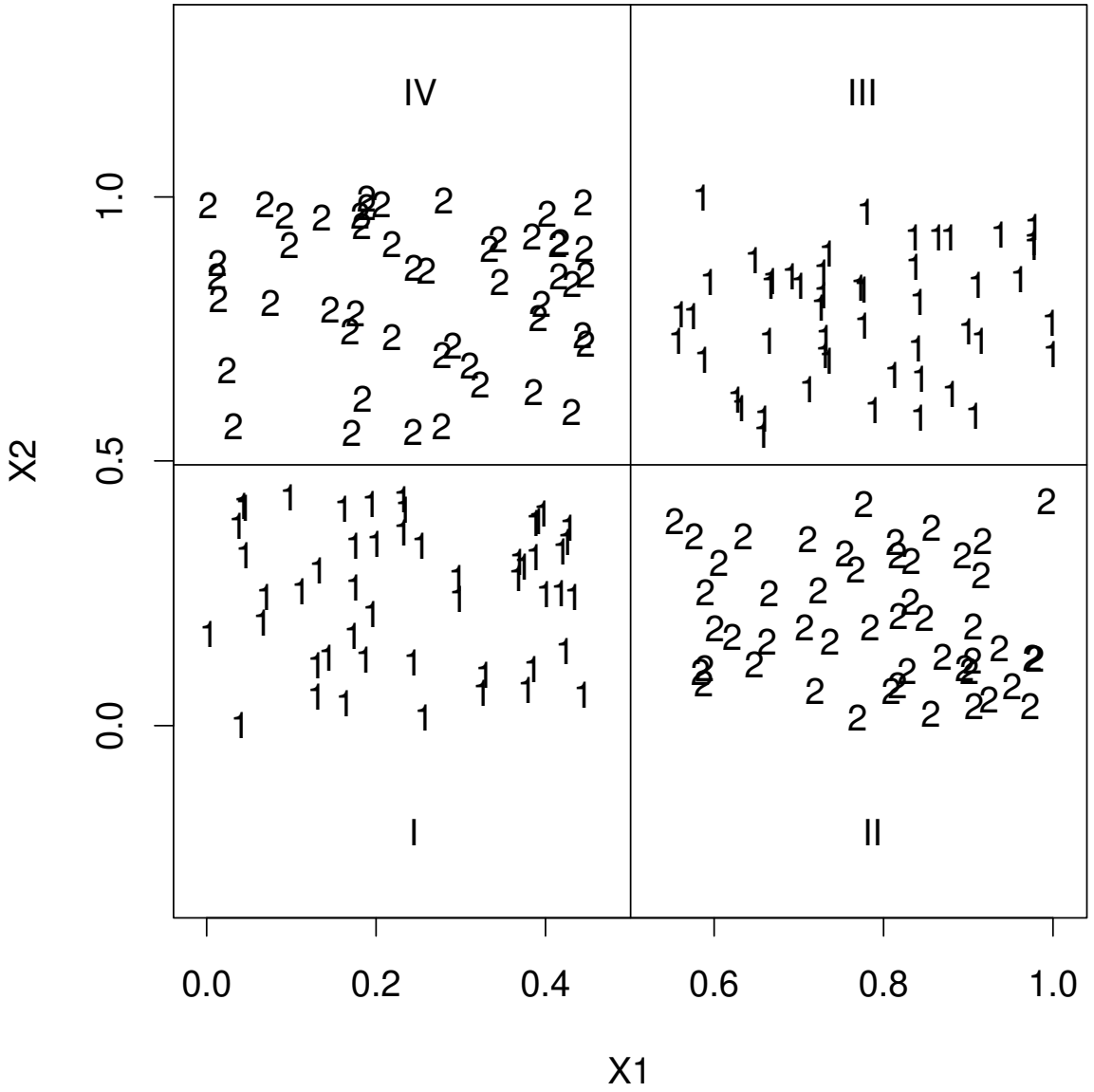
Level 3. Linear split. For each pair of ordinal variables $\{X_i, X_j\}$, apply marginal test to its largest linear discriminant coord. Select $\{X_i, X_j\}$ with smallest p-value if $p < 0.20/\{K'(K'-1)\}$, where K' is number of ordinal X variables. Otherwise, select most significant X from Level 1 tests.

Level 1 marginal tests for ordinal X

Let $N(t)$ = number of observations in node t and J_t = number of classes in t

1. Define $k = 3$ if $N(t) < 30J_t$ and $k = 4$ otherwise
2. If X has missing values, divide its range into $k - 2$ intervals at the $i/(k-1)$ quantiles, $i = 1, 2, \ldots, k-2$; add a "NA interval" for missing values
3. If X has no missing values, divide its range into $k - 1$ intervals at the i/k quantiles, $i = 1, 2, \ldots, k-1$
4. Form a table with Y values as rows and X intervals as columns.
5. Let ν be df of the table after deleting empty rows and columns
6. Compute chi-squared statistic χ^2_ν and p-value for testing independence
7. If $2 \times 10^{-6} < p < 1 - 2 \times 10^{-6}$, convert χ^2_ν to 1-df chi-squared W_M
8. Otherwise, use Wilson-Hilferty (1931) to convert χ^2_ν to W_M

X_1-X_2 interaction with insignificant marginal tests



Level 2 interaction tests for X_1, X_2

1. If X_i is ordinal, let $x_i^{(1)}, x_i^{(2)}, x_i^{(3)}$ be its 0.33, 0.50, and 0.67 quantiles

   No missing values in X_i:

   (a) If $N(t) < 50J_t$, split X_i into 2 intervals $A_{i1} = (-\infty, x_i^{(2)}]$, $A_{i2} = (x_i^{(2)}, \infty)$

   (b) If $N(t) \geq 50J_t$, split X_i 3 intervals $A_{i1} = (-\infty, x_i^{(1)}]$, $A_{i2} = (x_i^{(1)}, x_i^{(3)}]$, and $A_{i3} = (x_i^{(3)}, \infty)$

   Missing values in X_i: Split X_i into $A_{i1} = (-\infty, x_i^{(2)}]$, $A_{i2} = (x_i^{(2)}, \infty)$ and create a third "NA 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. Find chi-squared statistic and convert it to 1-df chi-squared $W_I(X_1, X_2)$

GUIDE split selection

If selected X is ordinal: Find best split over all c from the sets $\{X = \texttt{NA}\}$, $\{X \leq c \text{ or } X = \texttt{NA}\}$, and $\{X \leq c \text{ and } X \neq \texttt{NA}\}$

If selected X is categorical with c levels:

- If $c \leq 11$, search over all $(2^{c-1} - 1)$ splits of the form $\{X \in S\}$
- If $c > 11$, transform X to dummy vector and find best split on largest discriminant coord of dummy vectors; then convert it to form $\{X \in S\}$ — method originally proposed in Loh and Vanichsetakul (1988)

GUIDE variable importance scores

- Let $W(X_i, t)$ be the Wilson-Hilferty chi-squared value of X_i at node t if X_i is not constant in t
- If X_i is constant at the root node, define $W(X_i, t) = 0$ for all t
- Otherwise, if X_i is constant at t but not at the root node, define $W(X_i, t) = 0.5 \max_j W(X_j, t)$
- Define the unadjusted scores

$$v(X_i) = \sum_t \sqrt{n(t)} W(X_i, t)$$

where $n(t)$ is sample size in t and the sum is over the intermediate nodes of a tree with 4 levels of splits

Bias adjustment of $v(X_i)$

- For $b = 1, 2 \ldots, B$ (default is $B = 300$),
  1. Randomly permute the Y values keeping X values fixed
  2. Grow a tree with 4 levels of splits and let $v_b^*(X_i)$ be the unadjusted scores
- Define $\bar{v}(X_i) = B^{-1} \sum_b v_b^*(X_i)$
- The *bias-adjusted* importance scores are

$$\texttt{IMP}(X_i) = v(X_i)/\bar{v}(X_i)$$

Threshold for important variables in GUIDE

- X_i is said to be "unimportant" if it is independent of Y
- An "important" variable is one that is not "unimportant"
- To find a cut-off score for the important variables,
  1. Randomly permute the Y values keeping the X values fixed B times (default is $B = 300$)
  2. Let $u_b = \max_i \texttt{IMP}(X_i)$ for permutation $b = 1, 2, \ldots, B$
  3. Let $u^*(\alpha)$ be the $(1-\alpha)$-quantile of $u_1, u_2, \ldots, u_B$
  4. Under the hypothesis H_0 that all variables are unimportant,

$$P(\text{at least one } \texttt{IMP}(X_i) \text{ exceeds } u^*(\alpha)) \approx \alpha$$

- Hence all X_i such that $\texttt{IMP}(X_i) > u^*(\alpha)$ are considered important

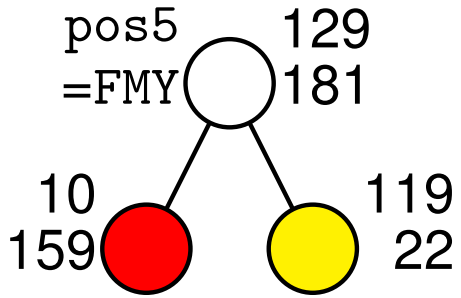
Importance scoring of GUIDE vs Random forest

- RF (as implemented in R) requires prior imputation of missing values
- GUIDE does not require missing value imputation
- GUIDE gives a threshold score for identifying noise variables
- RF does not have a threshold score for identifying noise variables
- RF scores are differences in accuracy between permuting and not permuting variables one at a time; this reduces importance of correlated variables

CART pruning

1. Let $R(T)$ be the misclassification cost of tree T and $\tilde{T}$ be its terminal nodes
2. Given α, define the cost-complexity function $R_\alpha(T) = R(T) + \alpha|\tilde{T}|$
3. For each α, there is a tree T that minimizes the cost-complexity
4. Let t be any node and T_t be the branch of T with root node t. Then

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



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

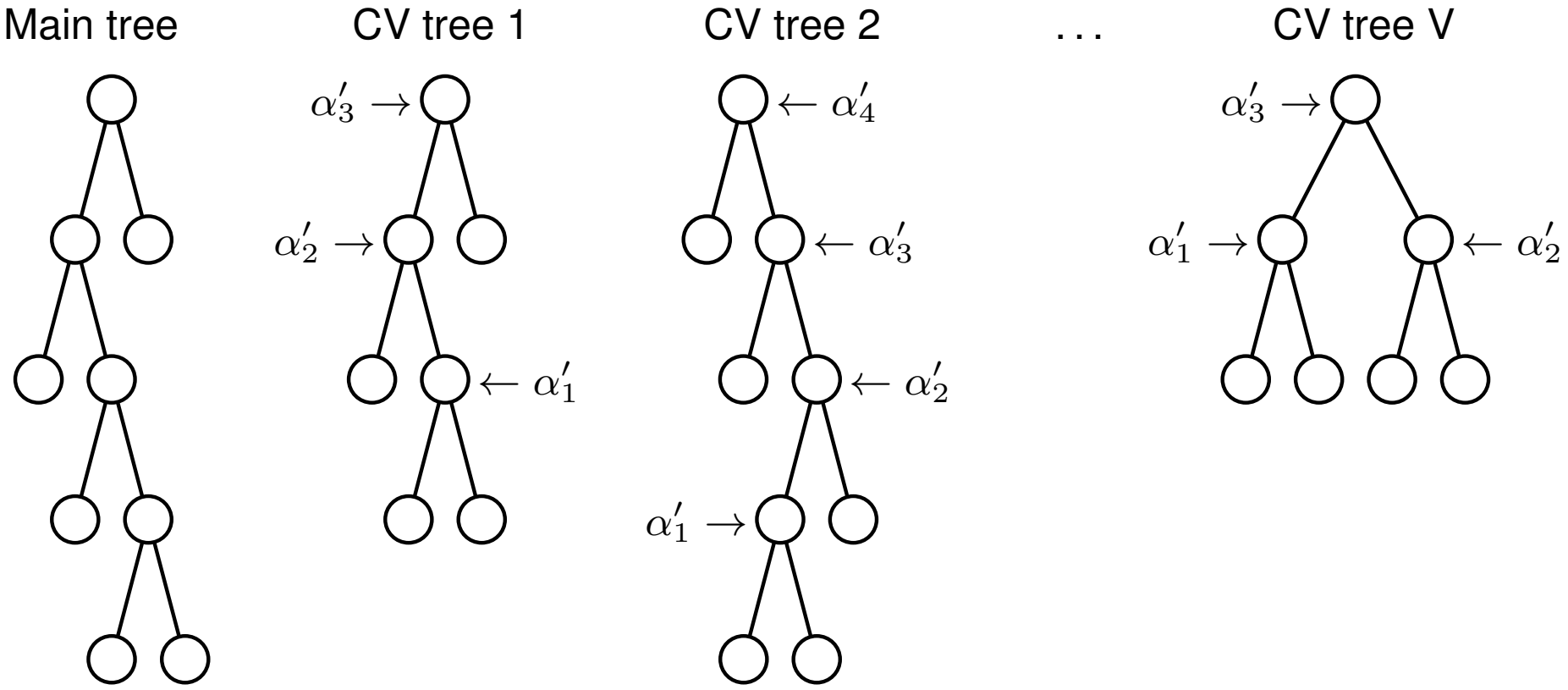
$$u(t) = [R(t) - R(T_t)]/[|\tilde{T}_t| - 1] = (129 - 32)/(310(2 - 1)) = 97/310$$

6. Prune branches at nodes t_1 for which $u(t_1) = \min\{u(t) \, : \, t \in T - \tilde{T}\}$
7. Define $\alpha_1 = u(t_1)$ and iterate to obtain a nested sequence of trees

Subtree selection by test-sample estimation

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

V-fold cross-validation



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

Subtree selection by V-fold cross-validation

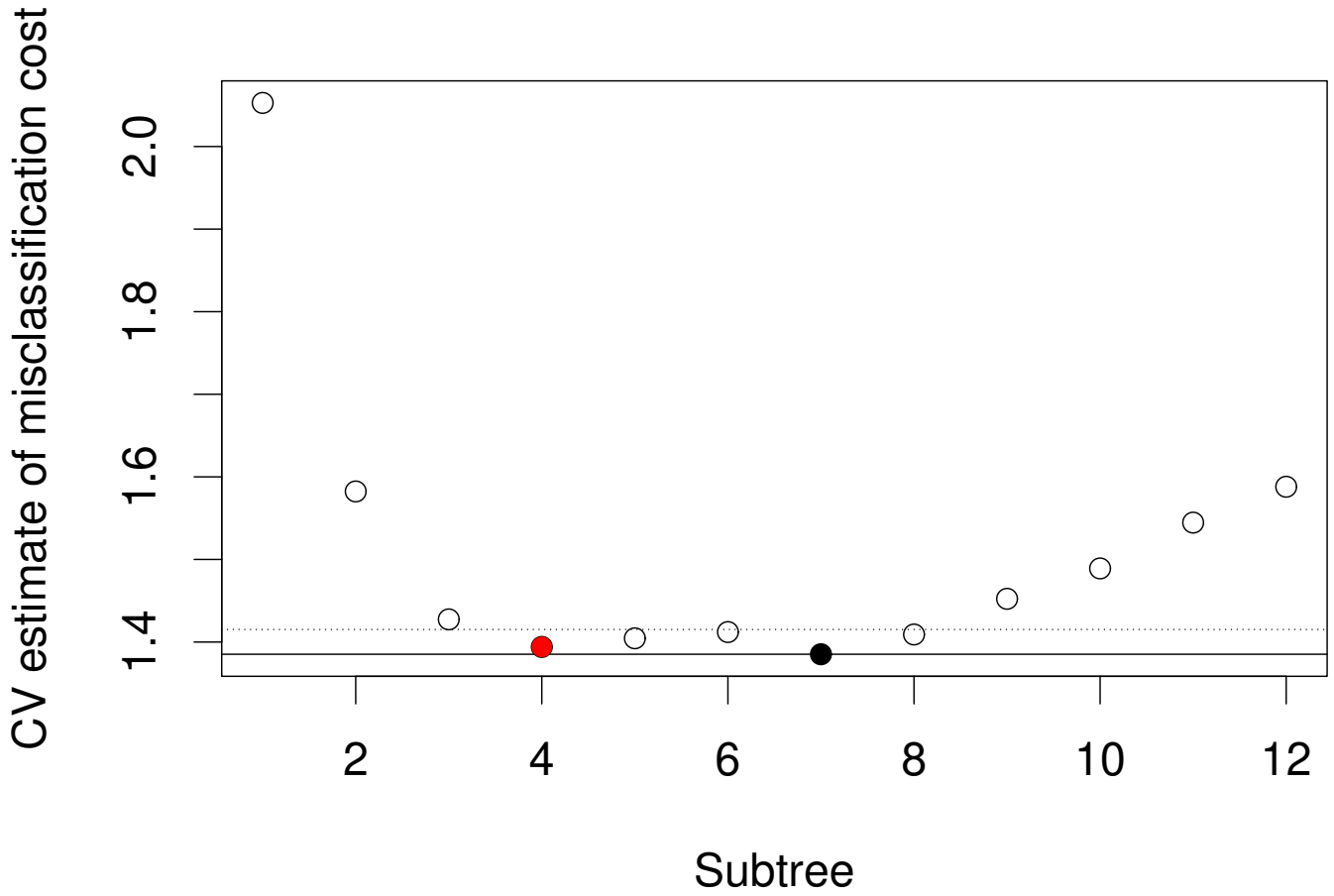
1. Let $\alpha_1 < \alpha_2 < \ldots$ be the α-values associated with the pruned sequence of subtrees $T_1 \succ T_2 \succ \ldots$. Define $\alpha'_k = \sqrt{\alpha_k \alpha_{k+1}}$
2. Divide $\mathcal{L}$ into V subsets $\mathcal{L}_1, \ldots, \mathcal{L}_V$
3. Let $T^{(v)}(\alpha'_k)$ be the minimal 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

k-SE rule

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



Note about ACS data

- INTP is blank in the csv data file if the respondent is less than 15 years old
- PAP, SEMP, etc., are also blank if respondent is less than 15 years old
- Because there is no information on INTP for people less than 15 years old, it is best to omit them from the analysis
- Allocation flag FINTP = 1 if INTP variable is originally missing or incorrect and has been imputed by the Census Bureau
- You can revert their values to originally missing with these 3 steps:

```
z <- read.csv("psam_p55.csv",na.strings="")
z <- z[!is.na(z$INTP),]
z$INTP[z$FINTP == 1] <- NA
```

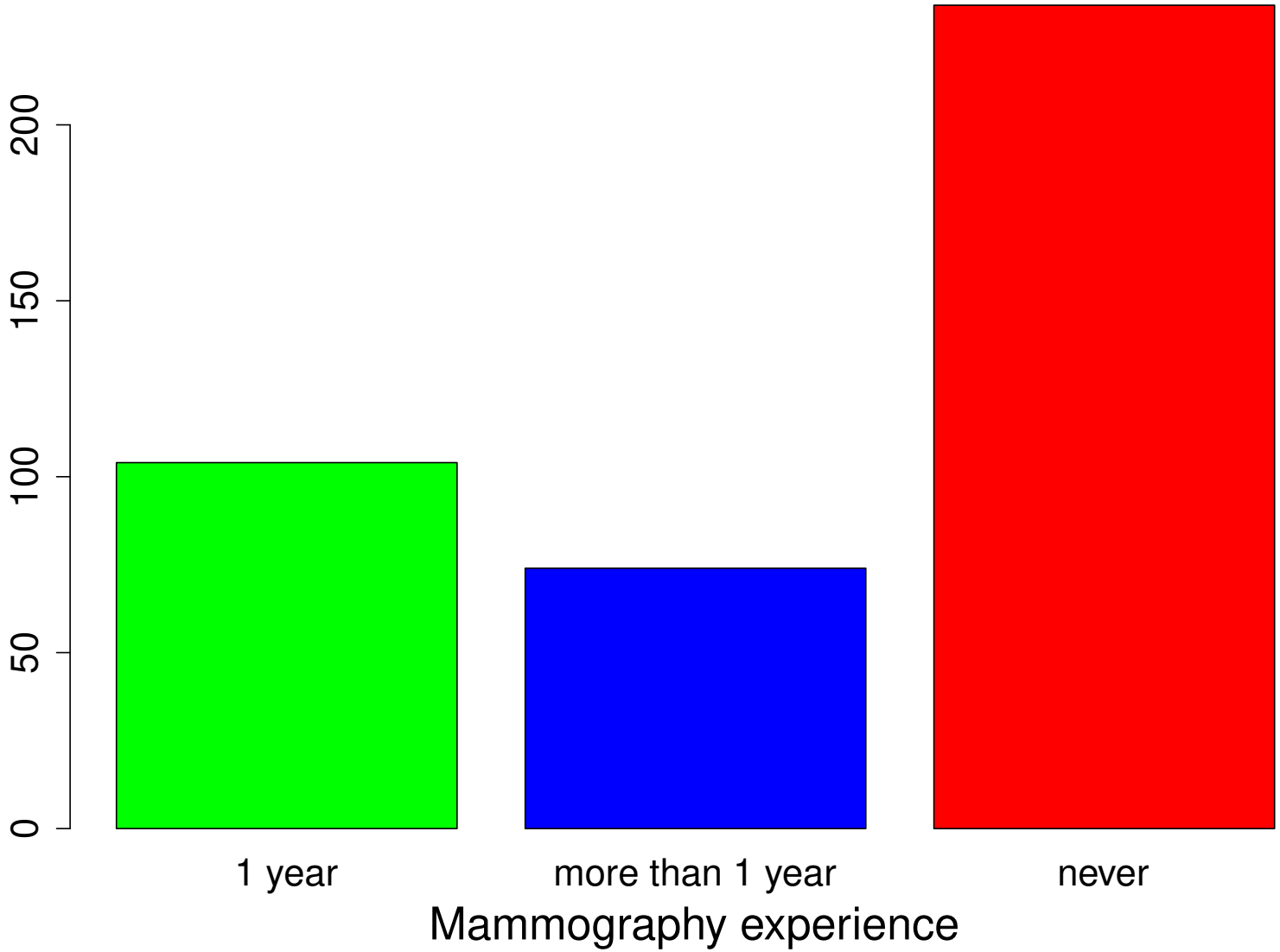
Ordinal response and misclassification costs: Knowledge, attitude, and behavior toward mammography (Hosmer and Lemeshow, 2000)

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

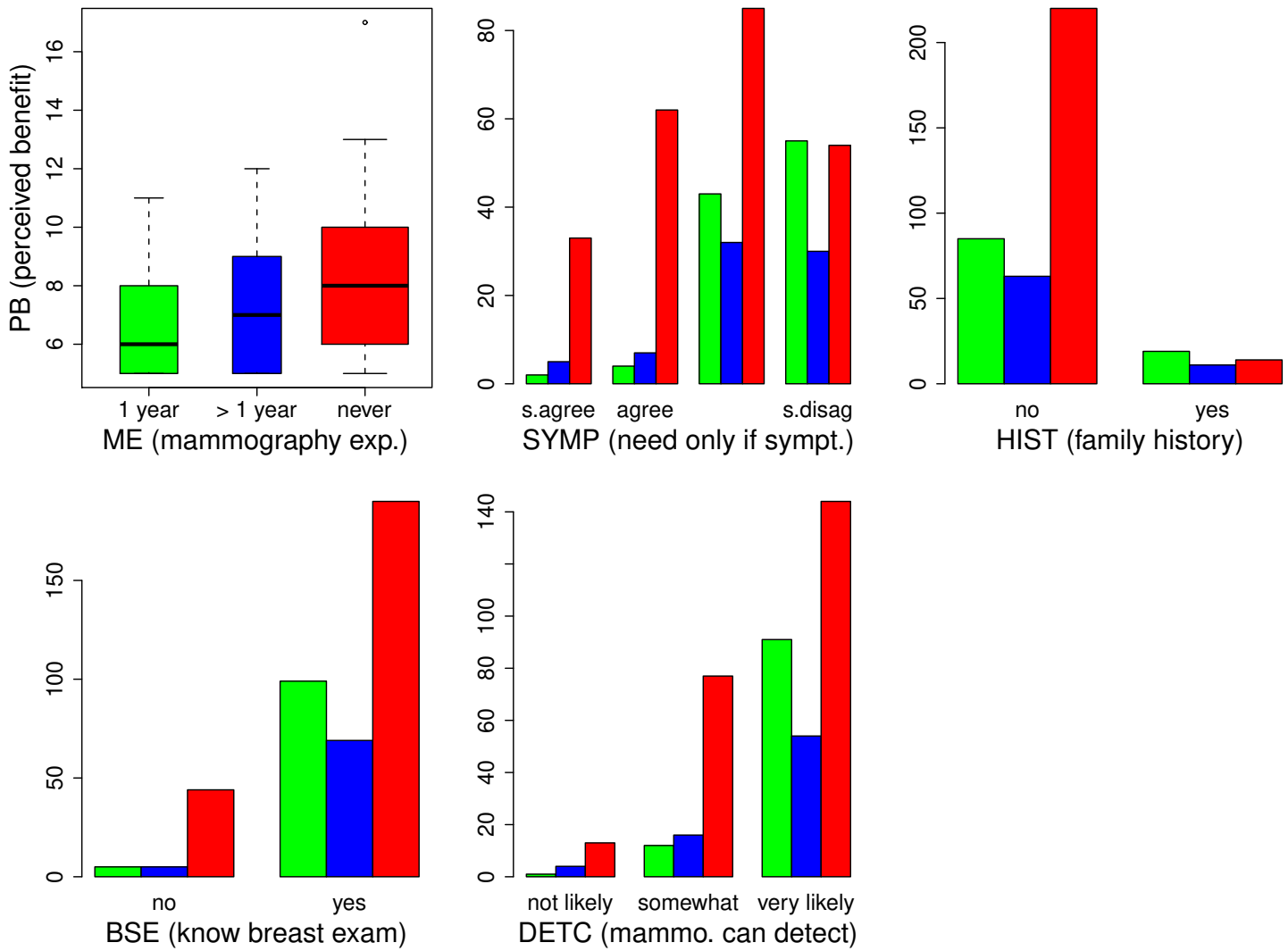
Mammography variables

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

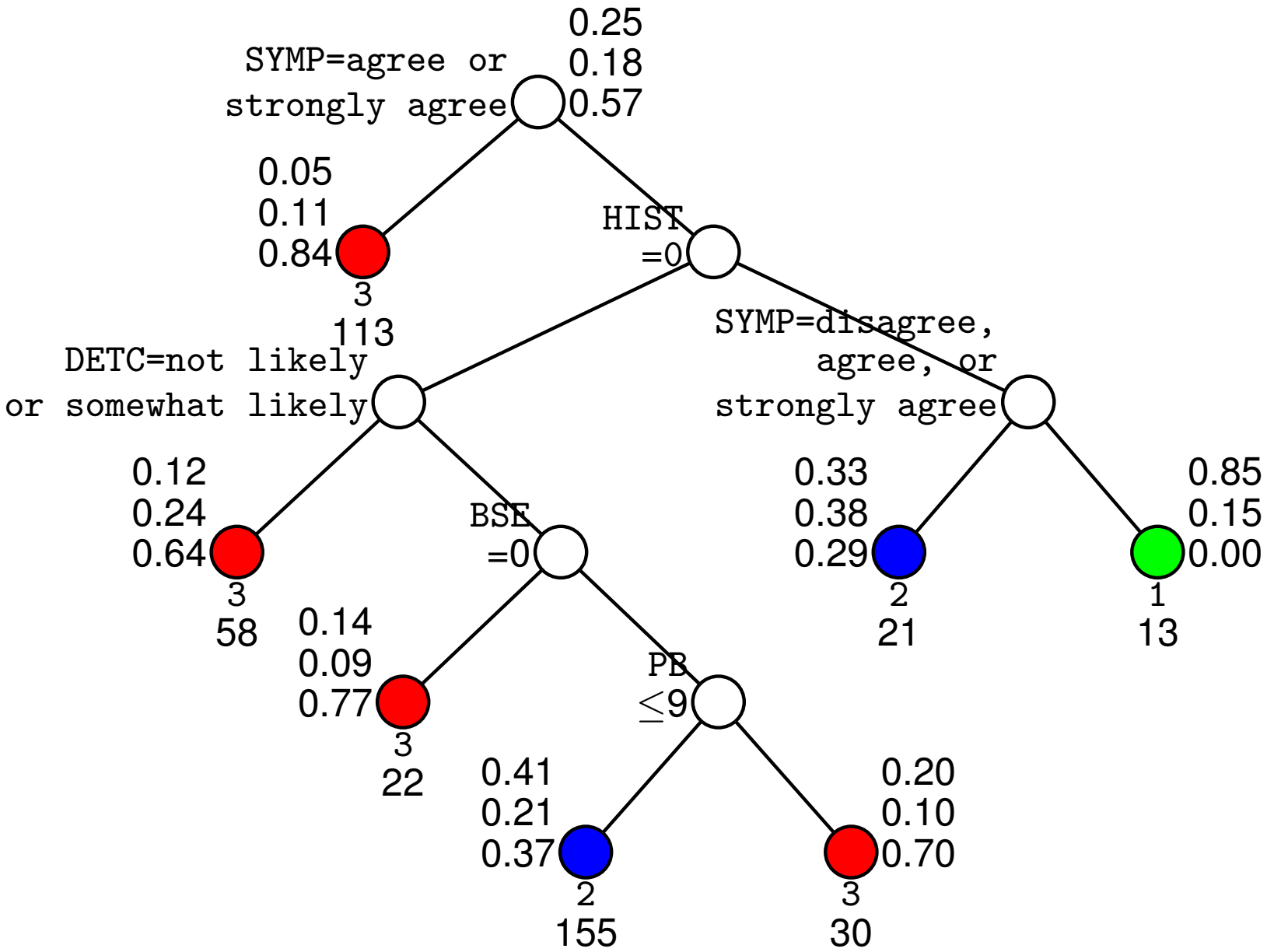
** DETCD = 1 if DETC = "very likely", DETCD = 0 otherwise

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

Ctree and Cforest do not allow unequal misclassification costs

GUIDE tree for mammography data



Chi-squared tests

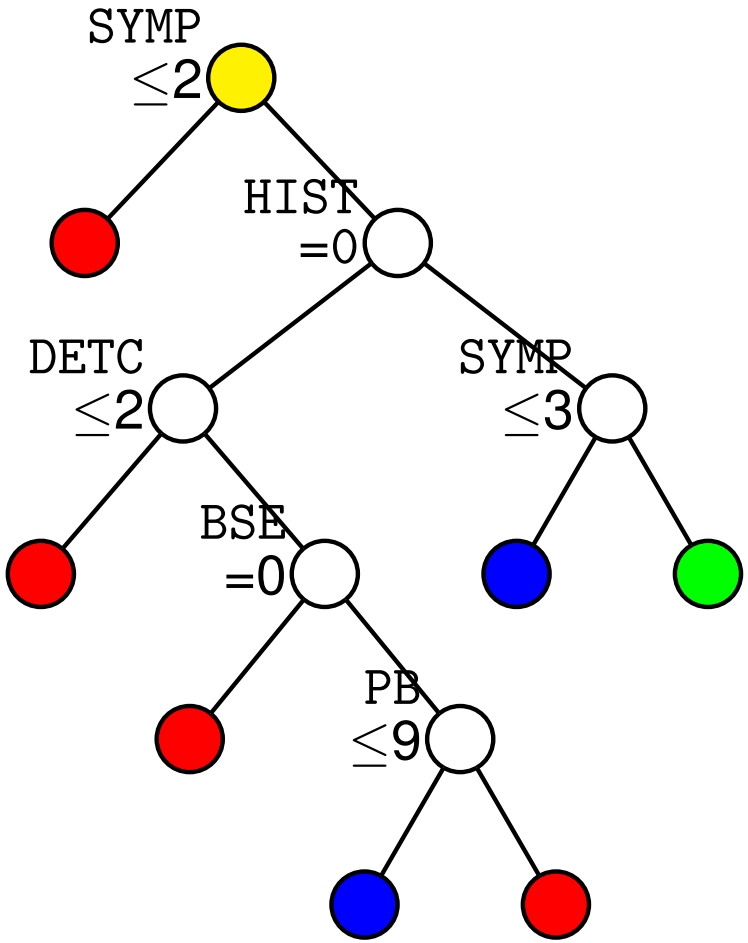
	SYMP (χ^2_6 = 57.2; $\chi^2_1 \approx 47$)			
ME	strongly agree	agree	disagree	strongly disagree
Never	33	62	85	54
1 year	2	4	43	55
$>$ 1 yr	5	7	32	30

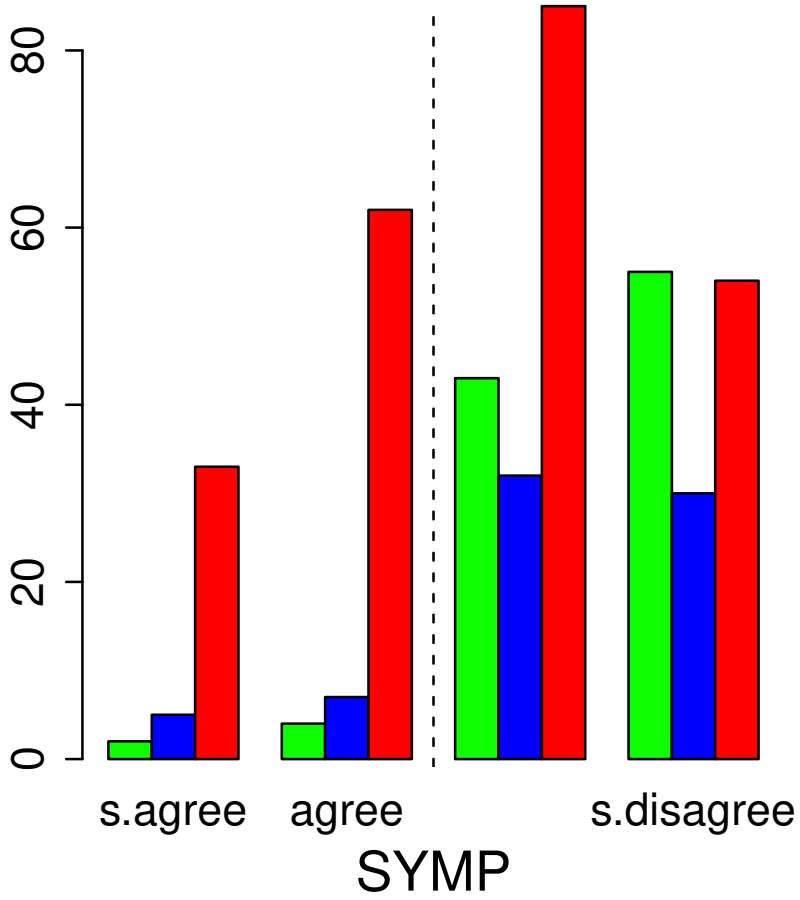
	PB (χ^2_6 = 31.3; $\chi^2_1 \approx 19$)			
ME	$\leq$ 5.7	(5.7, 7.6]	(7.6, 9.4]	$>$ 9.4
Never	33	68	65	68
1 year	31	43	22	8
$>$ 1 yr	19	25	18	12

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

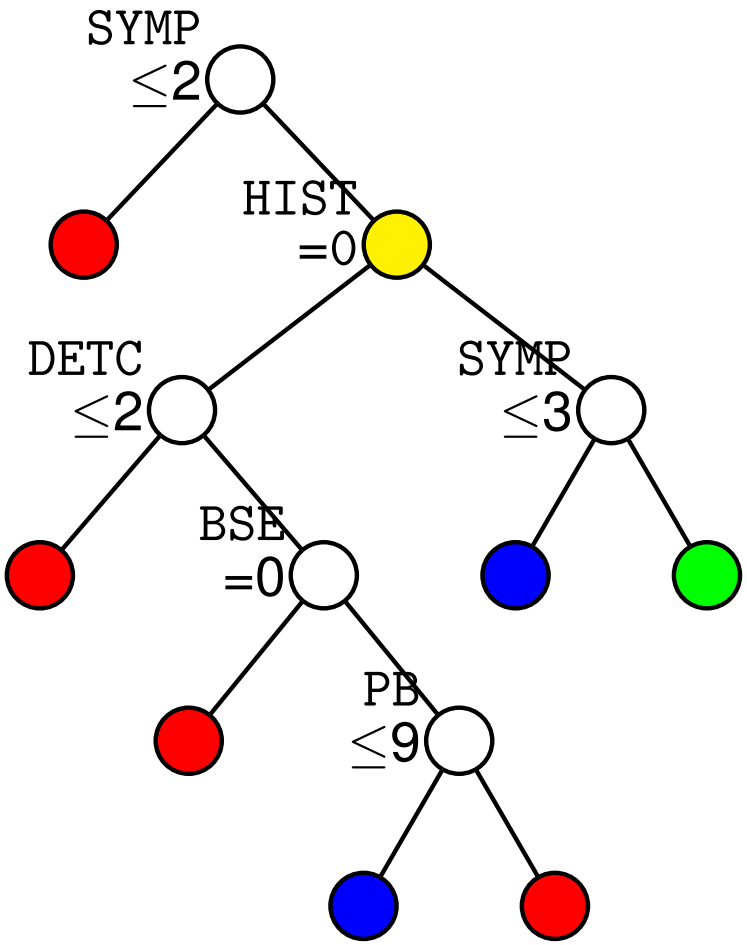
	BSE (χ^2_2 = 15.6, $\chi^2_1 \approx$ 13)		HIST (χ^2_2 = 13.1, $\chi^2_1 \approx$ 10)	
ME	no	yes	no	yes
Never	44	190	220	14
1 year	5	99	85	19
$>$ 1 yr	5	69	63	11

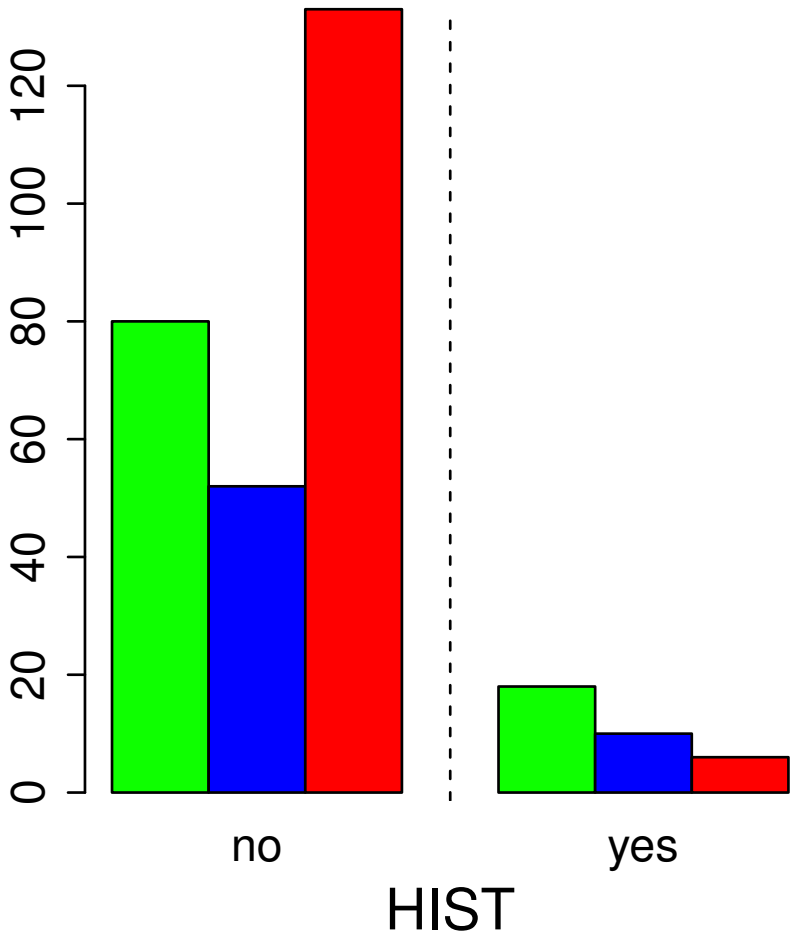
1st split


SYMP
≤2
HIST
=0
DETC
≤2
SYMP
≤3
BSE
=0
PB
≤9

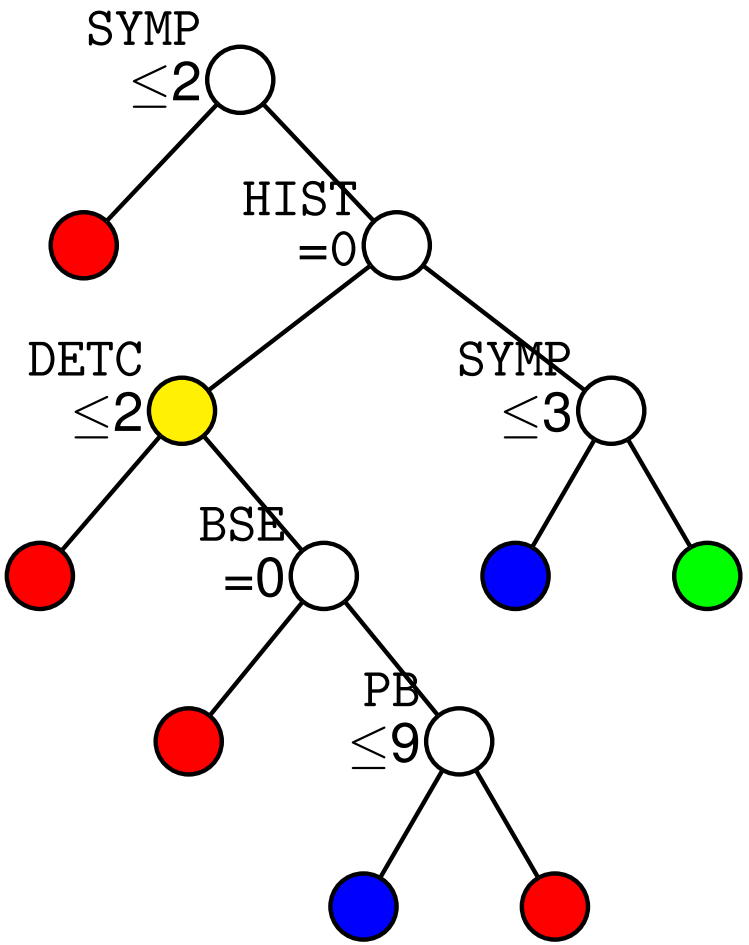

0
20
40
60
80
s.agree
agree
s.disagree
SYMP

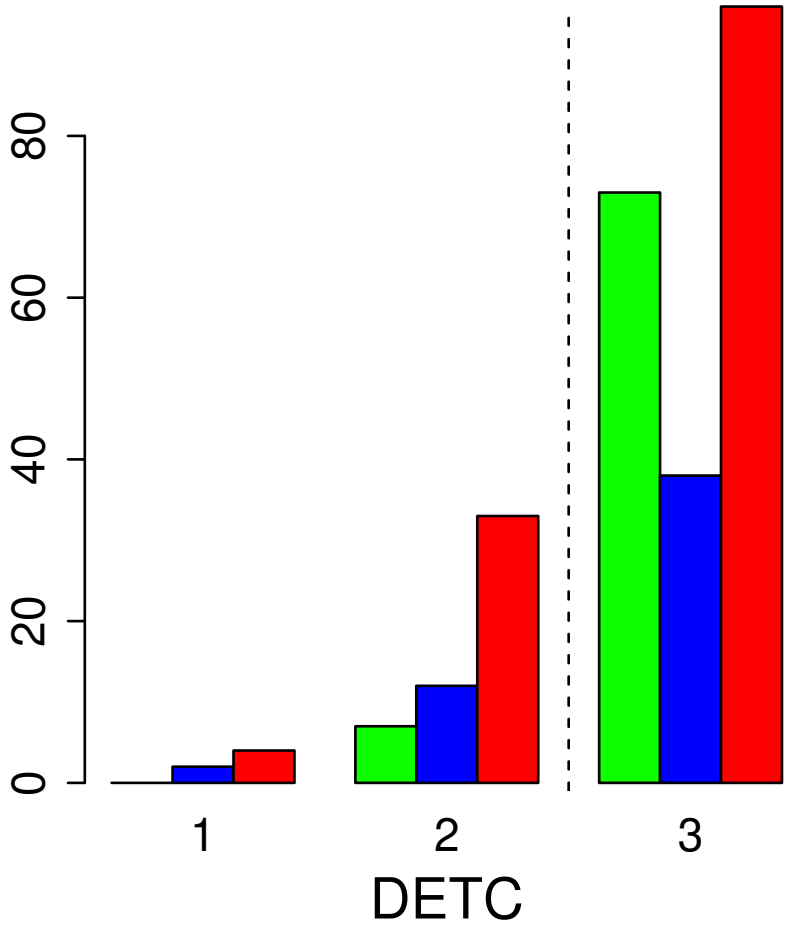
2nd split


SYMP
≤2
HIST
=0
DETC
≤2
SYMP
≤3
BSE
=0
PB
≤9

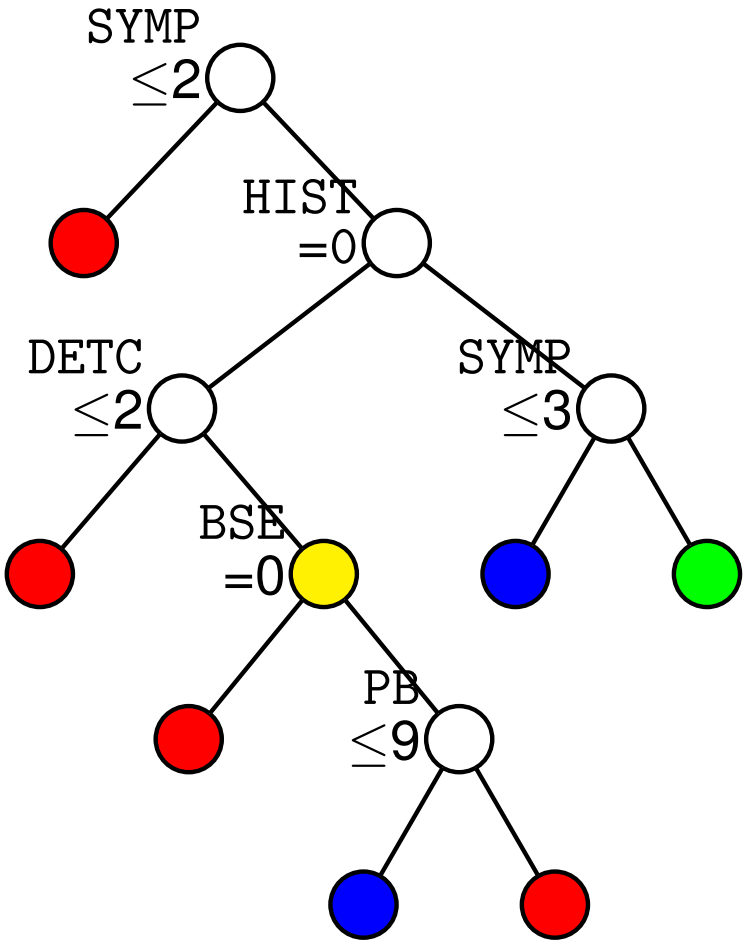

0
20
40
60
80
100
120
no
yes
HIST

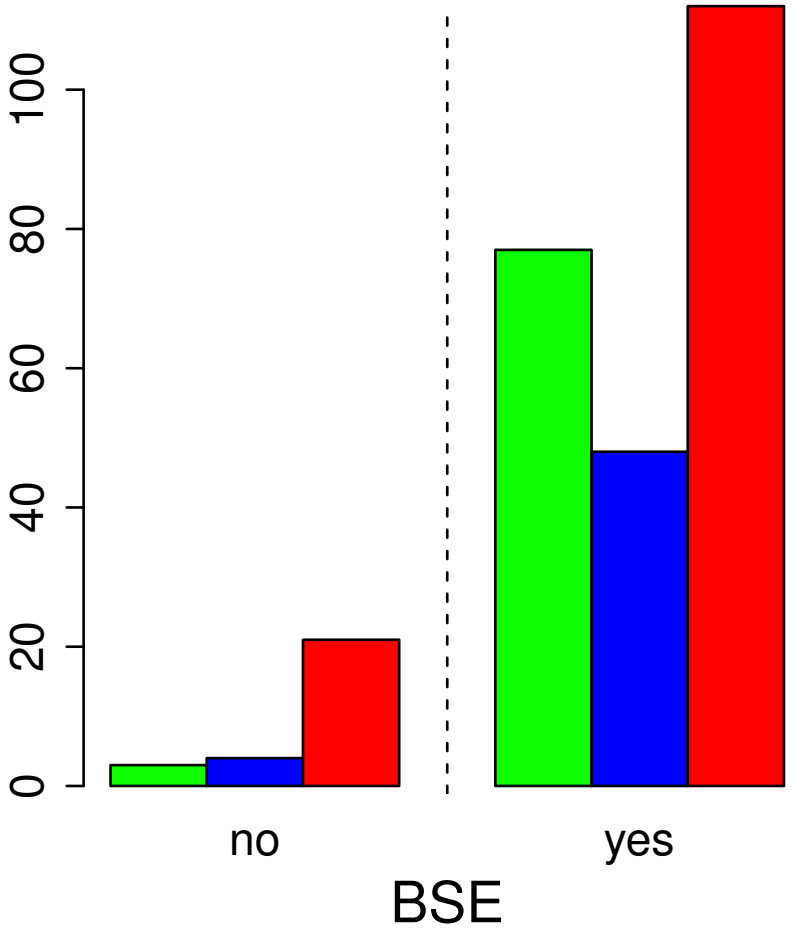
3rd split



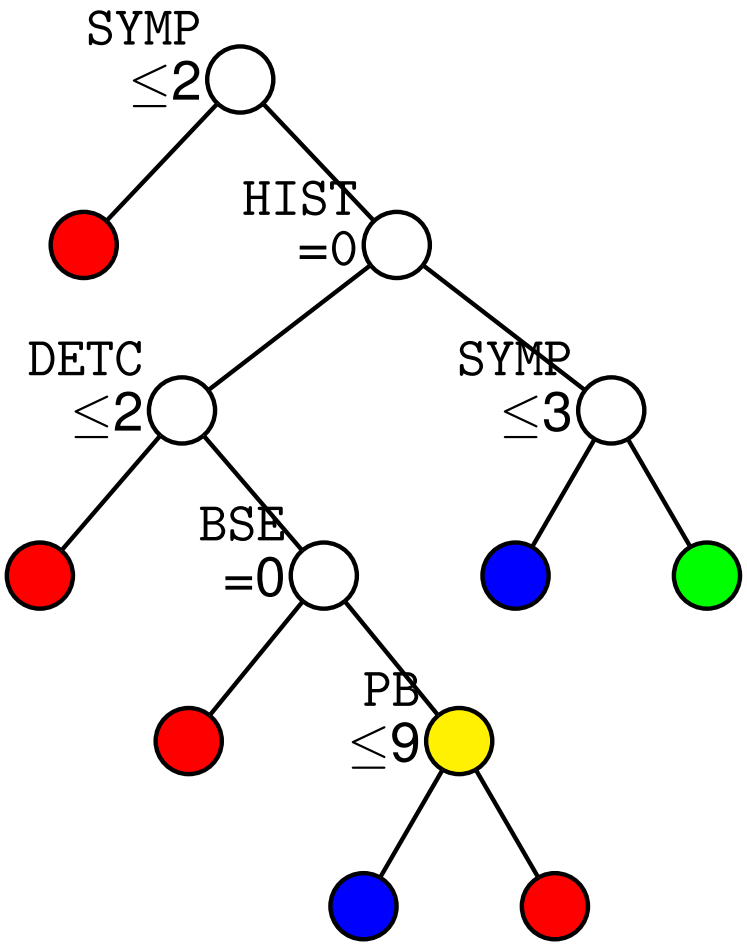


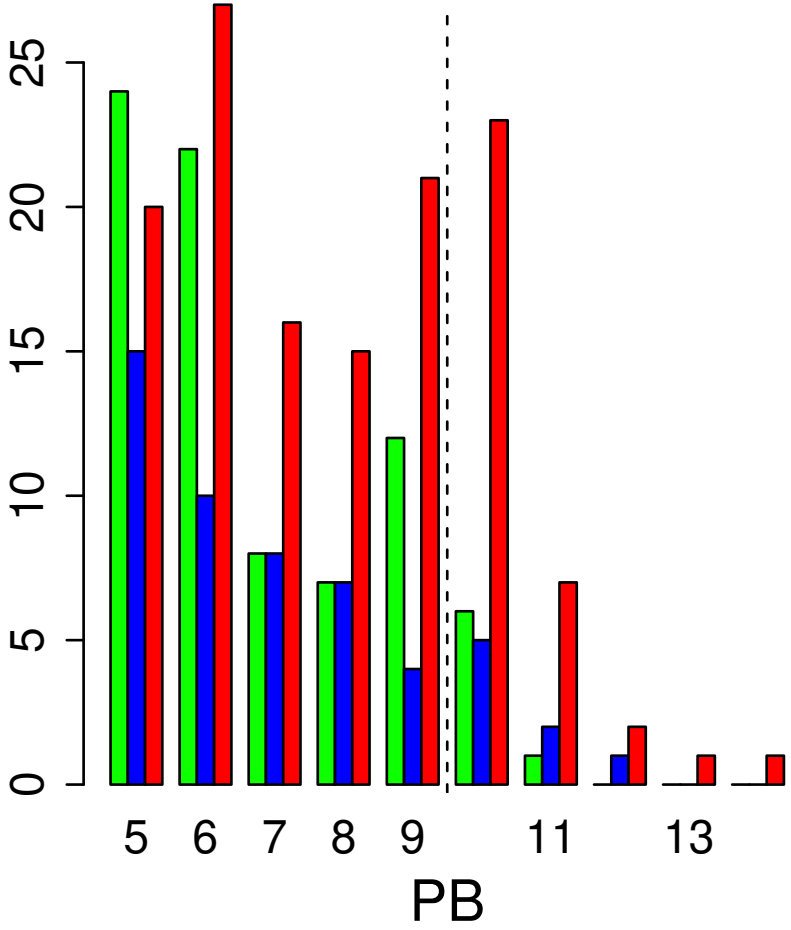
4th split



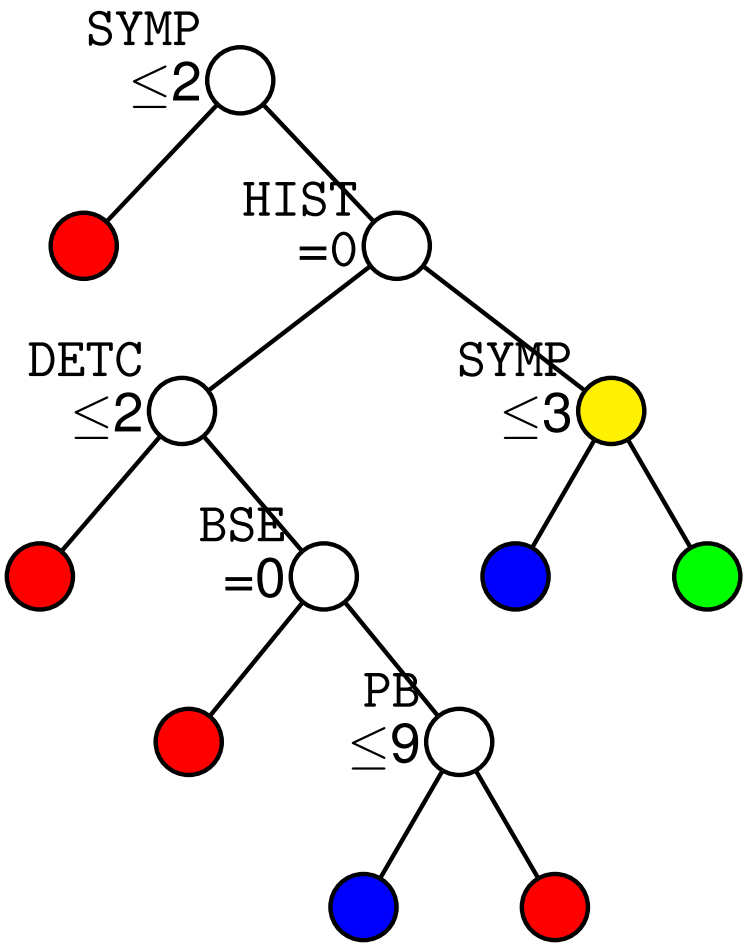


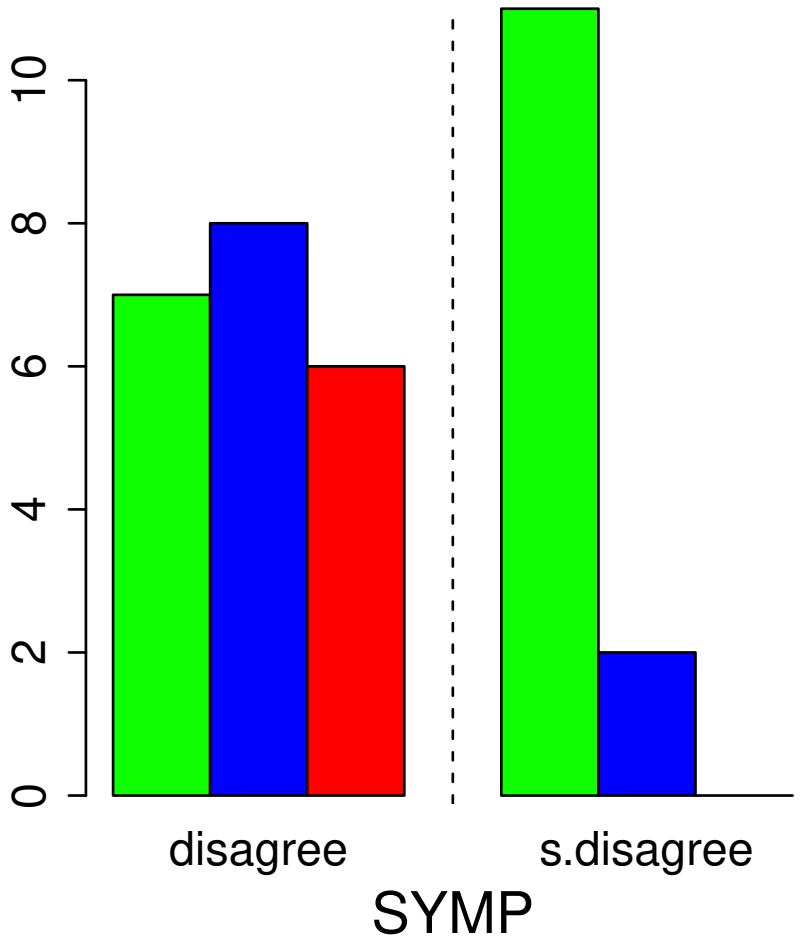
5th split





6th split


SYMP
≤2
HIST
=0
DETC
≤2
SYMP
≤3
BSE
=0
PB
≤9


0
2
4
6
8
10
disagree
s.disagree
SYMP

Ctree and Cforest (Hothorn et al., 2006)

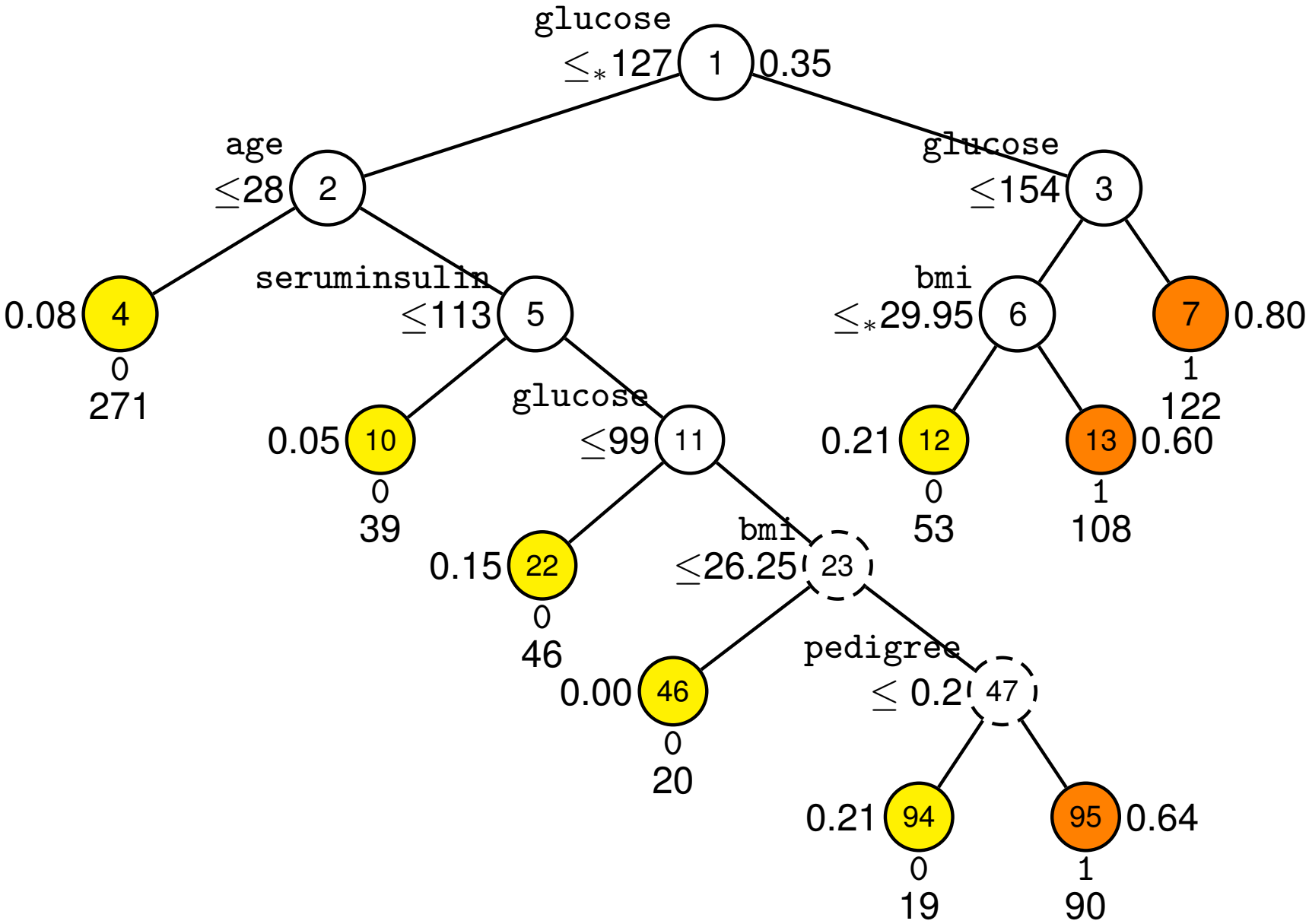
- Implemented in `party` and `partykit` packages; no class priors, no unequal misclassification costs, and no real weights (only replicate counts)
- `partykit` inapplicable to categorical variables with $>$ 31 levels
- ctree uses permutation tests of independence to select split variables; observations with missing values are omitted from the tests
- ctree has three methods to pass missing values through a split
  - Default randomly sends missing values to the child nodes, with probabilities proportional to node sample sizes
  - 2nd alternative uses surrogate splits with maximum number of such splits set by `maxsurrogate` parameter
  - 3rd alternative (for `partykit`) sends missing values to the larger child node (using `ctree_control` parameter `majority=TRUE`)
- cforest cannot predict cases with categorical values not in training data

Missing values in predictor variables: Pima Indian diabetes data

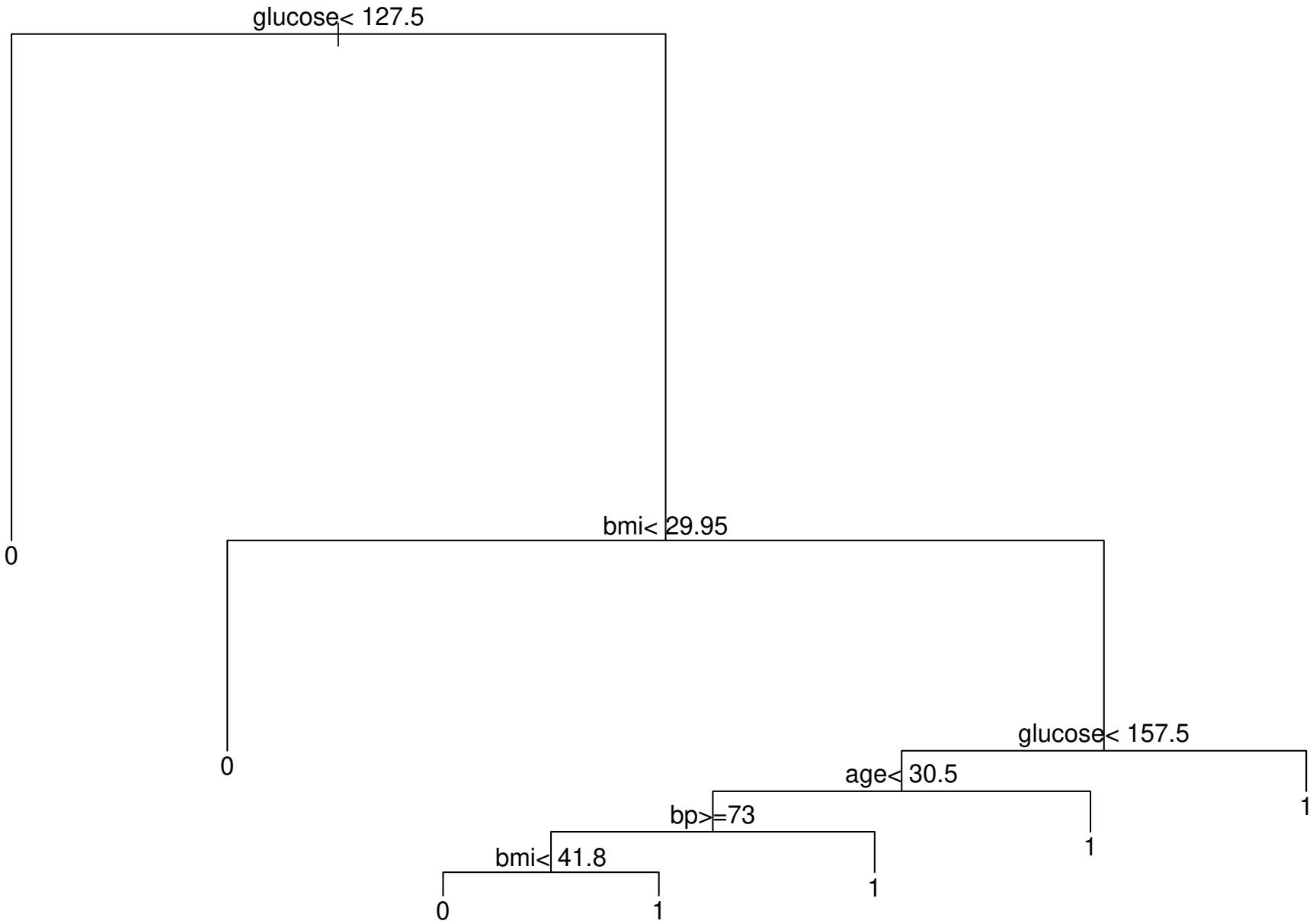
- 768 observations; 268 positive (class `1`) and 500 negative (class `0`) for diabetes
- 8 ordinal predictor variables with missing values

Variable	#Miss
pregnant (# times pregnant)	111
glucose (plasma glucose conc.)	5
bp (diastolic blood pressure)	35
skinfold (skin fold thickness)	227
seruminsulin (serum insulin)	374
bmi (body mass index)	11
pedigree (diabetes pedigree function)	0
age (age in years)	0

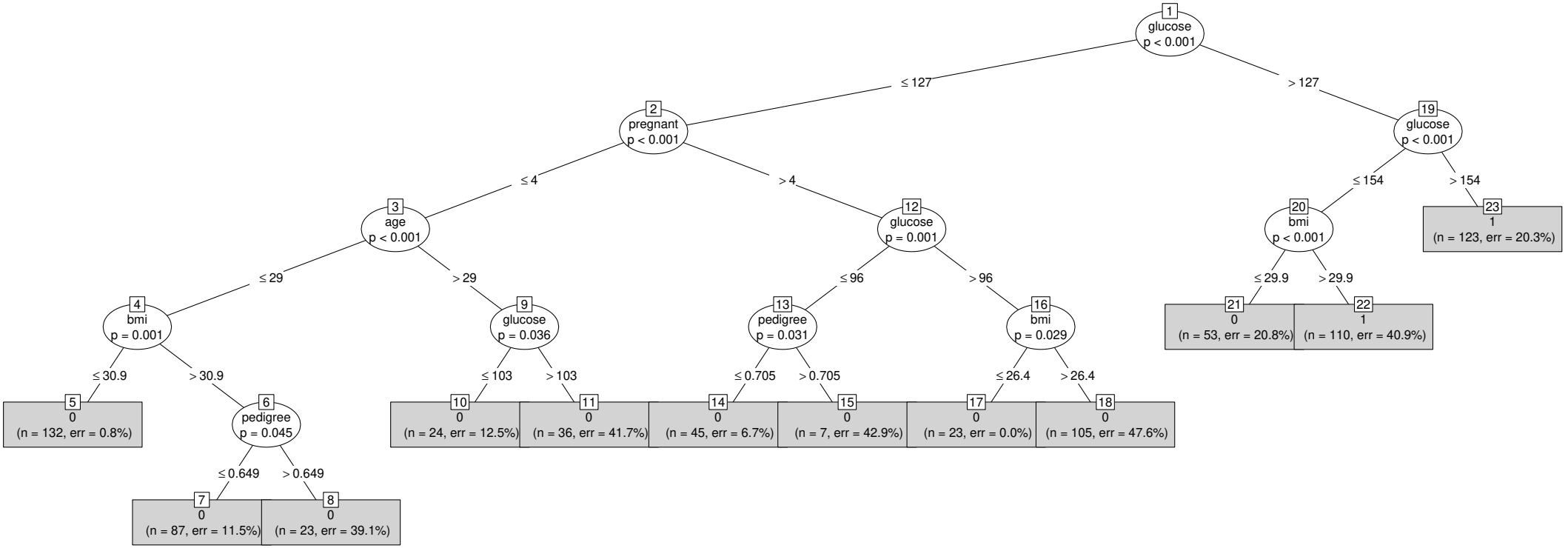
GUIDE tree for Pima Indian data



RPART tree for Pima Indian data



Ctree (partykit) tree for Pima Indian data



Leave-one-out error counts for Pima Indian data

Method	Errors[a]	Time[b]
Ctree (partykit)	173	0.08
RPART	178	0.02
GUIDE forest	182	15.24
Cforest (partykit)	188	37.88
GUIDE tree, univ1 splits	193	1.00
GUIDE tree, linear splits	214	1.92
GUIDE tree, bivariate kernel	215	4.37
GUIDE tree, bivariate NN	216	4.16
GUIDE tree, univ2 splits	225	0.96

[a] out of 768 observations

[b] average time (sec.) to fit one data set

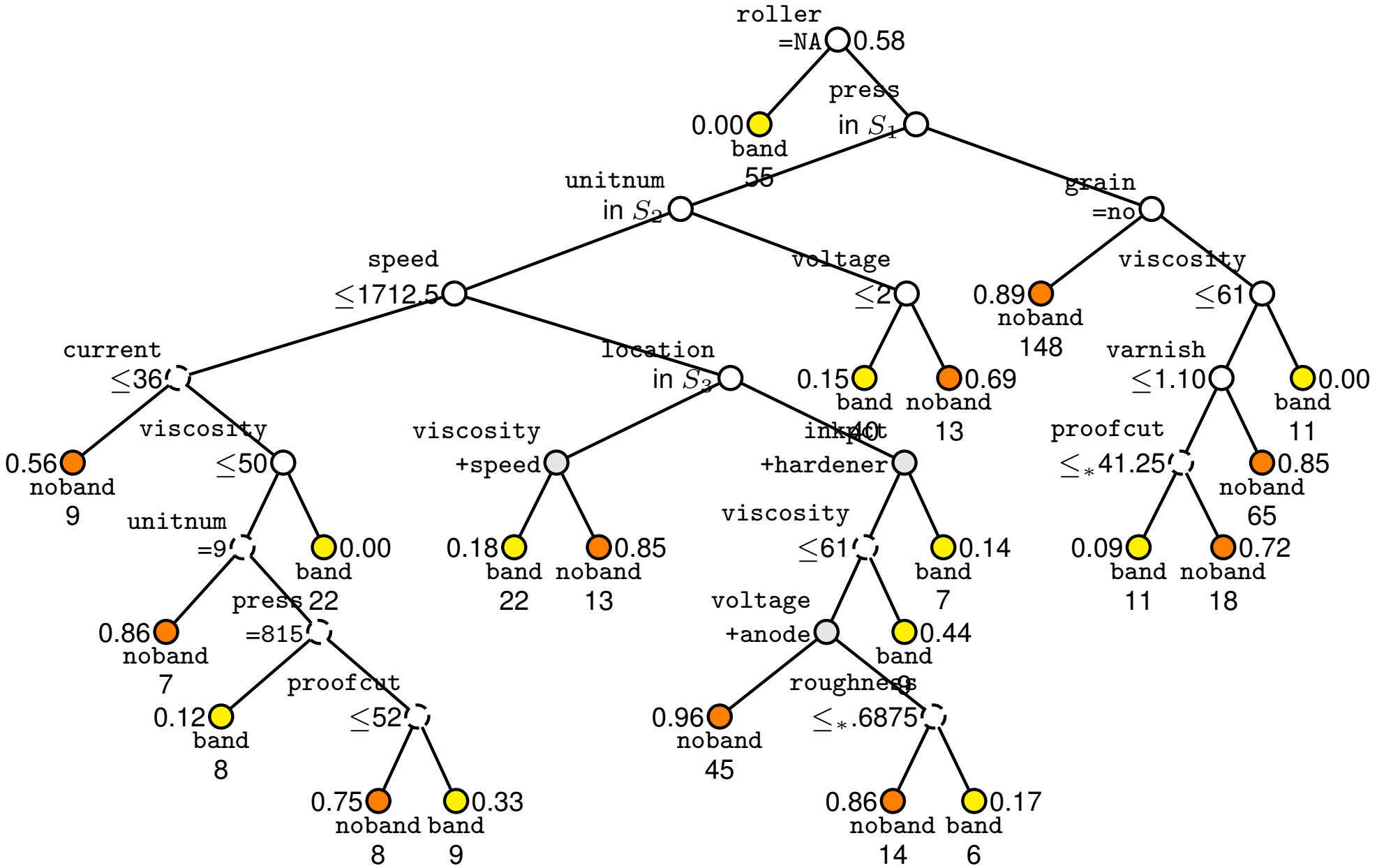
Missing values in predictor variables: cylinder banding in rotogravure printing

- 540 observations on 19 continuous and 14 categorical variables
- Response variable is binary: band (228) vs noband (312)
- Missing values in 175 observations

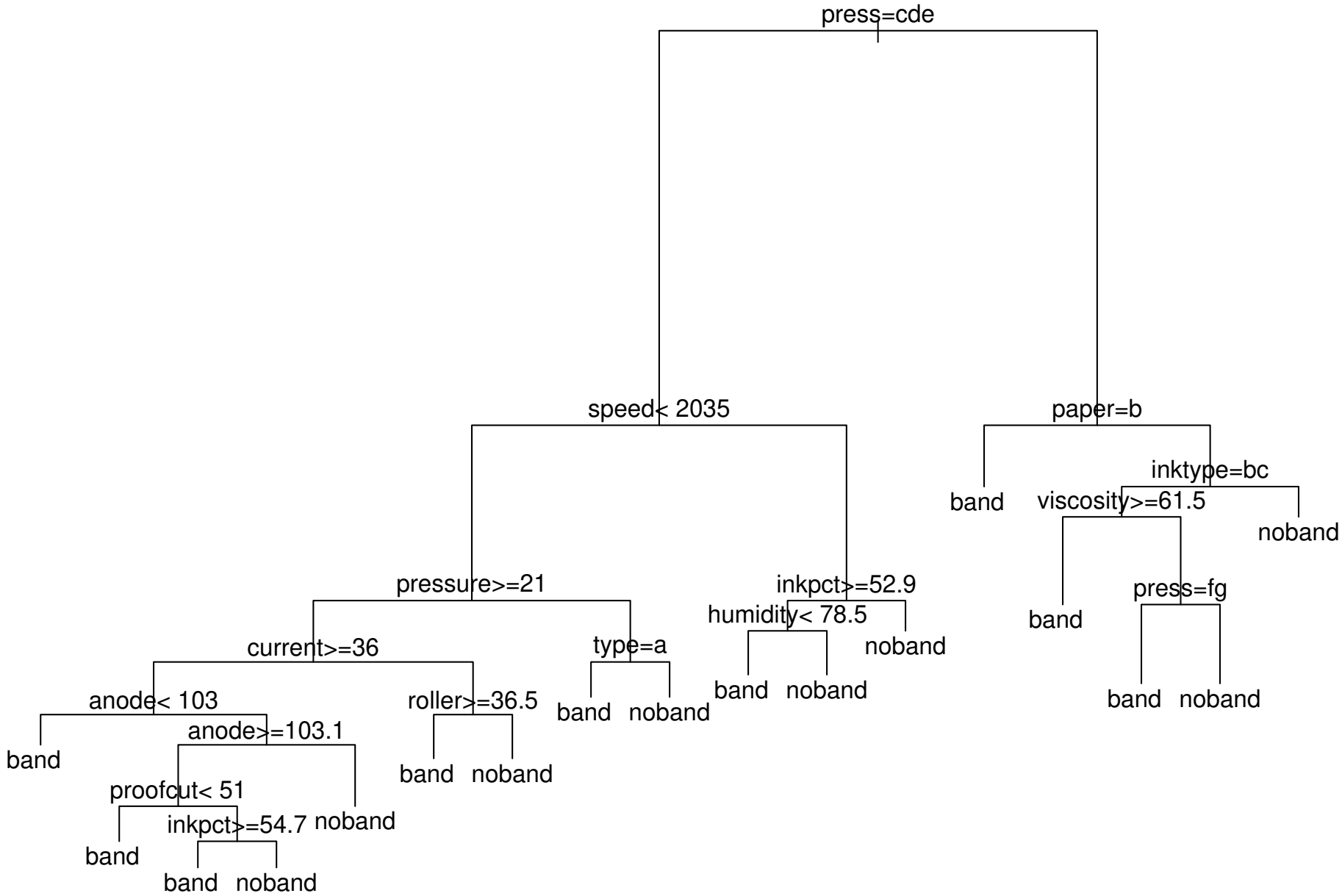
Predictor variables
(number of categorical levels in parentheses)

Variable	#Miss	Variable	#Miss	Variable	#Miss
grain (2)	49	size (3)	3	speed	10
proof (2)	57	location (5)	156	inkpct	56
blademfg (2)	60	tank (2)	18	solventpct	56
paper (3)		proofcut	55	voltage	57
inktype (3)		viscosity	5	amperage	55
direct (2)	25	caliper	27	wax	6
solvent (3)	55	temperature	2	hardener	7
type (2)	18	humidity	1	roller	55
presstype (4)		roughness	30	current	7
press (8)		pressure	63	anode	7
unitnum (7)		varnish	56	chrome	3

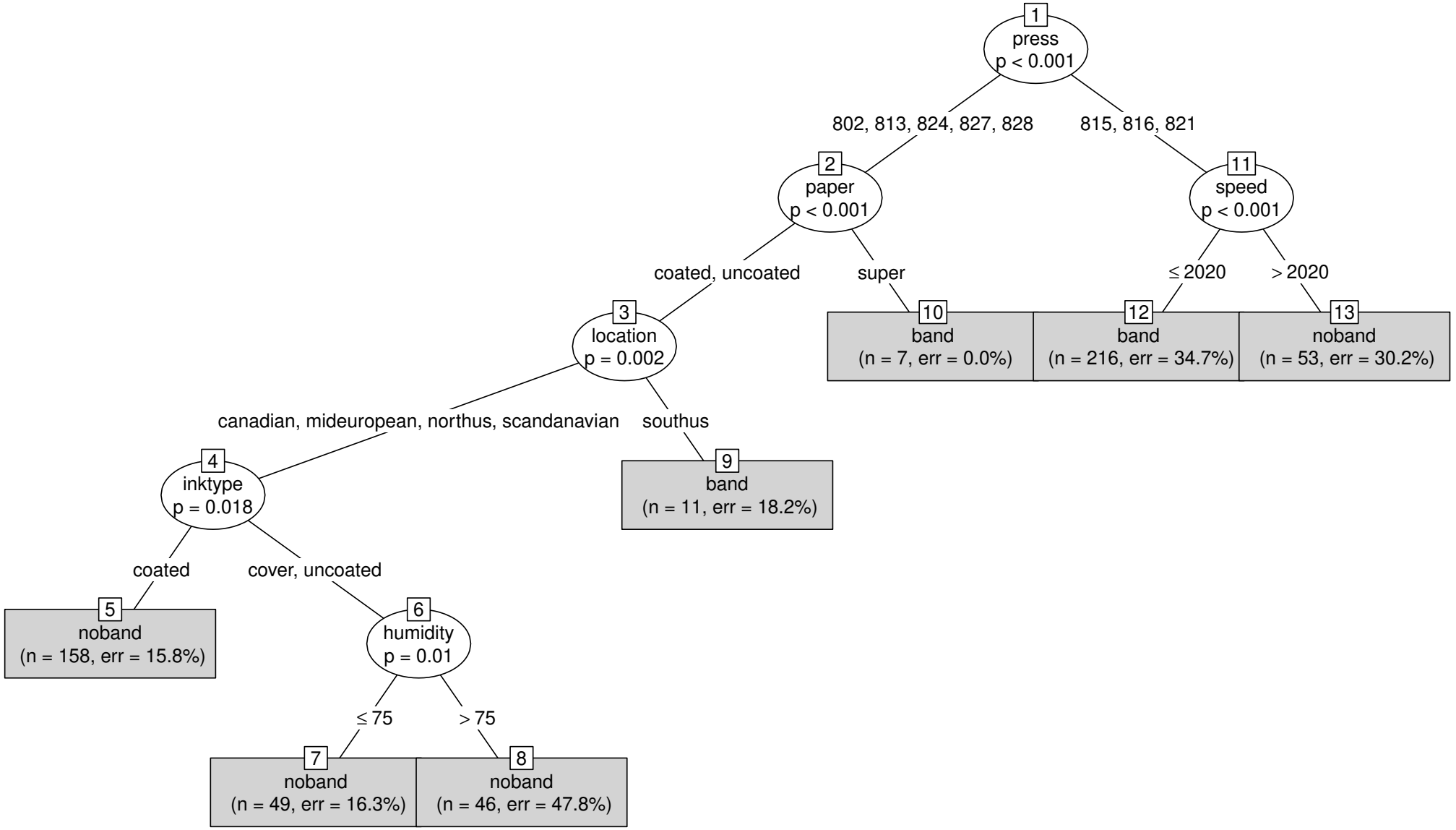
GUIDE tree for cylinder data



RPART tree for cylinder data



Ctree (partykit) tree for cylinder data



Leave-one-out error counts for cylinder data

Method	Errors[a]	Time[b]
GUIDE forest	100	15.88
GUIDE tree, univ2 splits	124	2.10
Cforest (partykit)	126	44.11
GUIDE tree, univ1 splits	136	2.09
GUIDE tree, linear splits	143	2.46
Ctree (partykit)	165	0.16
GUIDE tree, bivariate kernel	182	1.12
GUIDE tree, bivariate NN	183	1.11
RPART	190	0.05

[a] out of 540 observations, without variable blademfg

[b] average time (sec.) to fit one data set

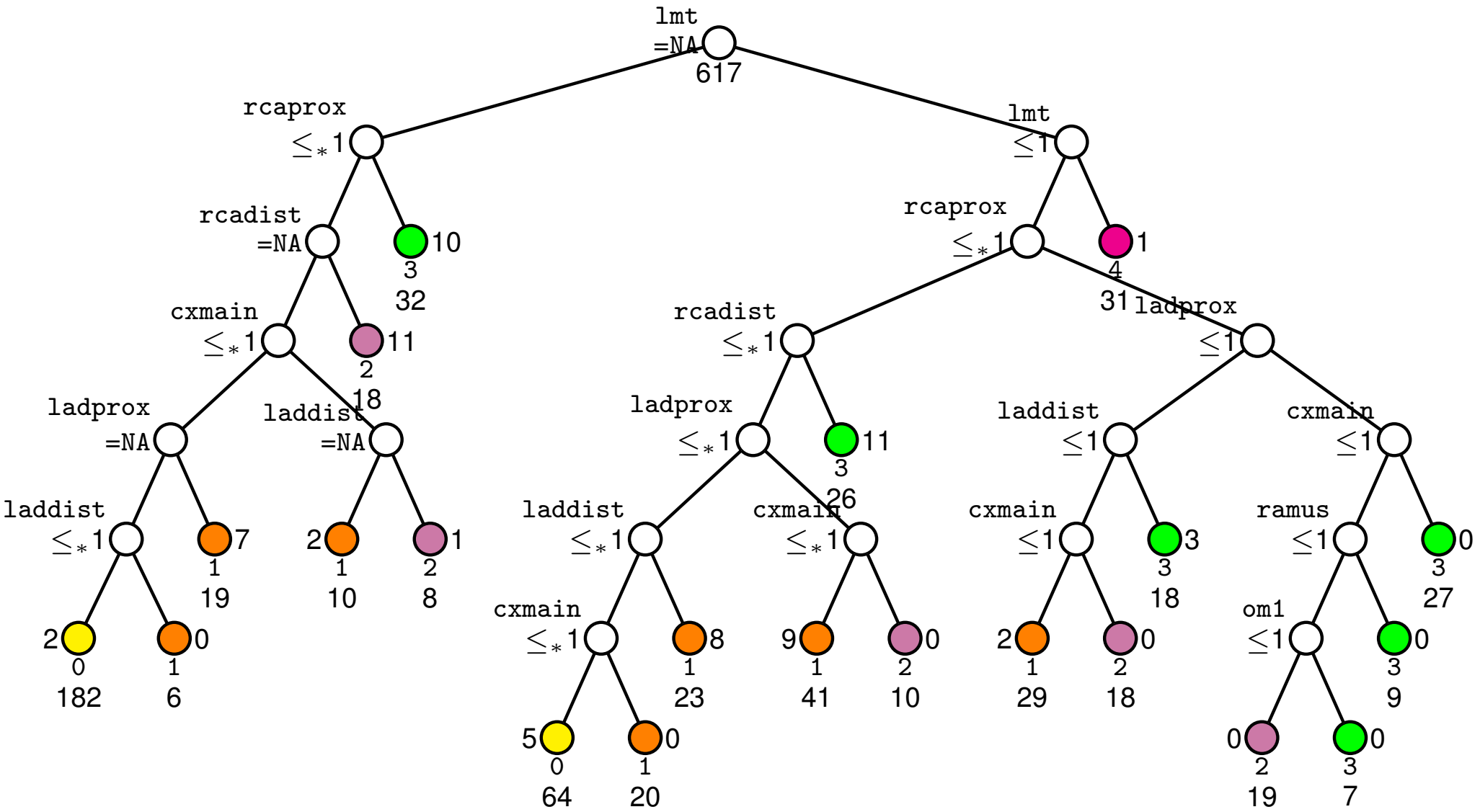
Heart disease data

- 617 observations on 29 ordinal and 22 categorical variables
- Response variable (`num`) has 5 classes (0, 1, . . . , 4)
- Missing values in 615 observations
- `http://archive.ics.uci.edu/ml/datasets/Heart+Disease`

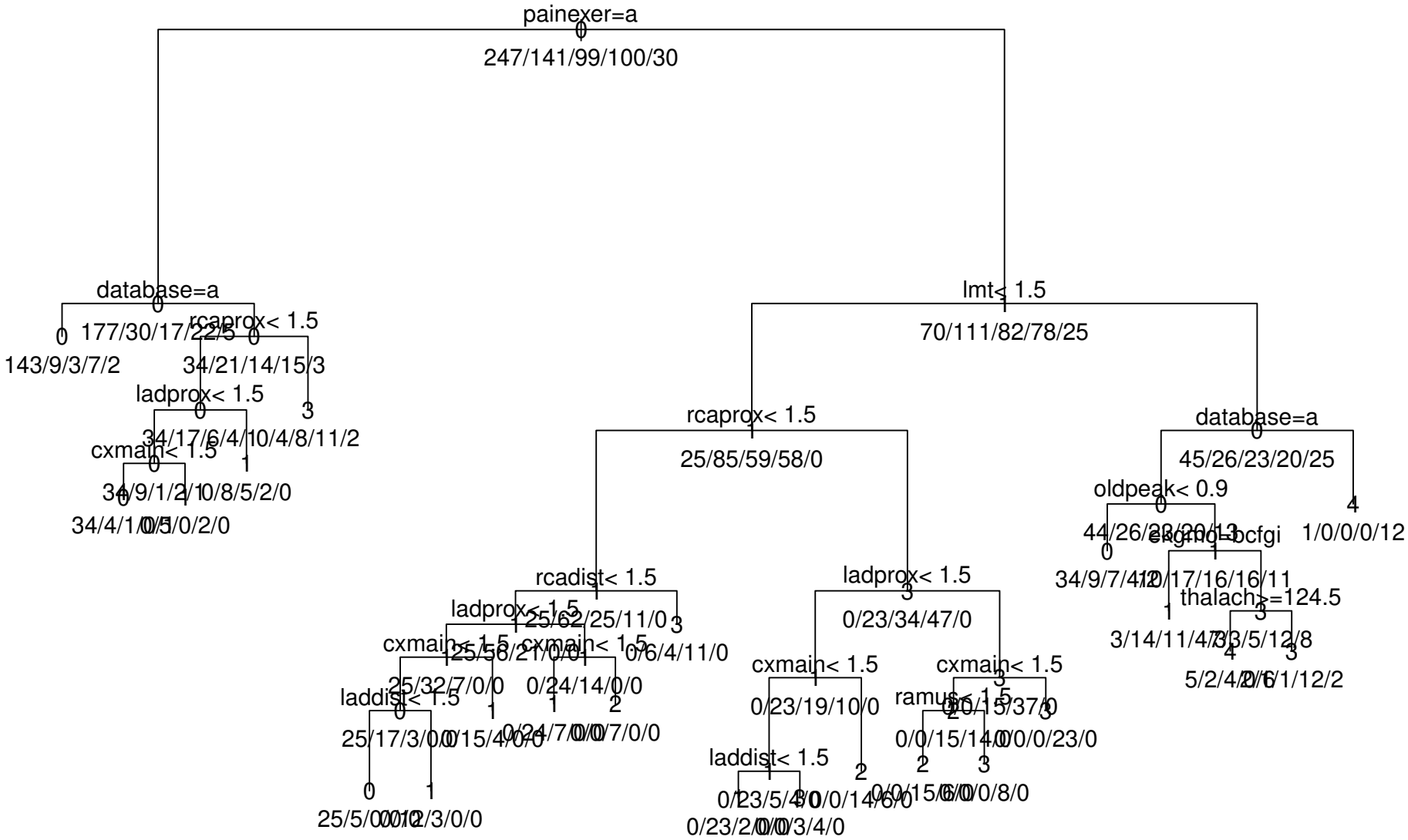
Predictors (categorical levels in parentheses)

Variable	#Miss	Variable	#Miss	Variable	#Miss	Variable	#Miss
age		famhist (2)	422	thalach	55	cyr	9
sex (2)		restecg (3)	2	thalrest	56	lmt	275
painloc (2)		ekgmo (12)	53	tpeakbps	63	ladprox	236
painexer (2)		ekgyr	53	tpeakbpd	63	laddist	246
relrest (2)	4	dig (2)	66	trestbpd	59	diag	276
cp (4)		prop (3)	64	exang (2)	55	cxmain	235
trestbps	59	nitr (2)	63	xhypo (2)	58	ramus	285
chol	30	pro (2)	61	oldpeak	62	om1	271
smoke (2)	387	diuretic (2)	80	slope (4)	308	om2	290
cigs	415	proto (14)	112	rldv5	143	rcaprox	245
years	427	thaldur	56	rldv5e	142	rcadist	270
fbs (2)	90	thaltime	384	ca	606	database (3)	
dm (2)	545	met	105	thal (7)	475		

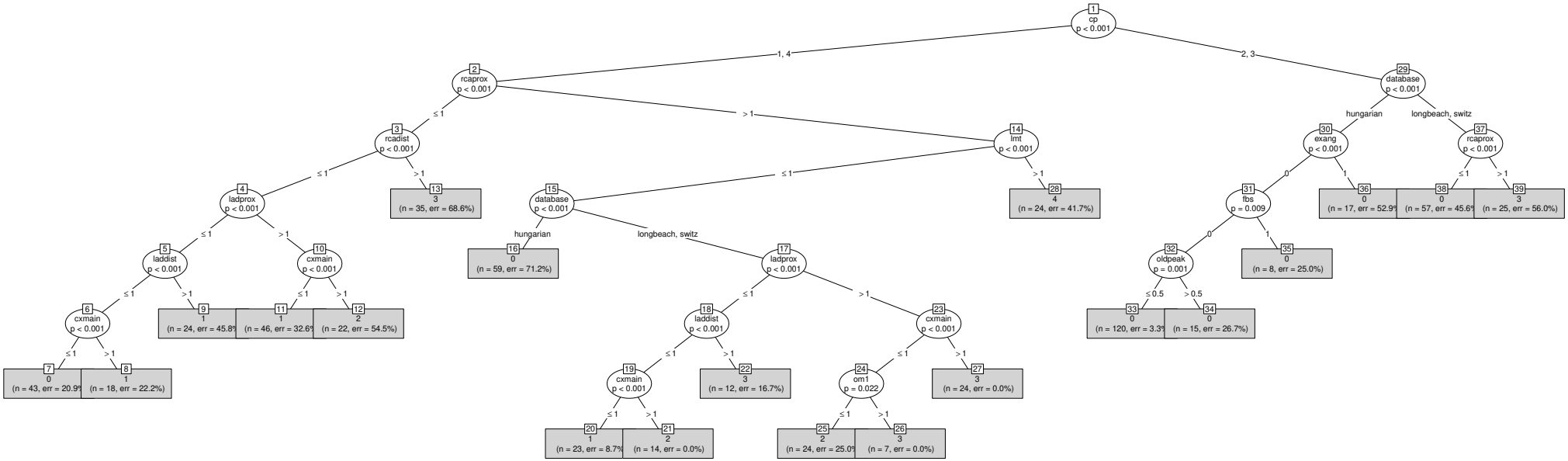
GUIDE tree


lmt
=NA
617
rcaprox
≤∗1
lmt
≤1
rcadist
=NA
10
3
32
rcaprox
≤∗1
1
4
31
cxmain
≤∗1
11
2
18
rcadist
≤∗1
ladprox
≤1
ladprox
=NA
laddist
=NA
ladprox
≤∗1
11
3
26
laddist
≤1
cxmain
≤1
laddist
≤∗1
7
1
19
2
1
10
1
2
8
laddist
≤∗1
cxmain
≤∗1
cxmain
≤1
3
3
18
ramus
≤1
0
3
27
2
0
0
1
182
6
cxmain
≤∗1
8
1
23
9
1
41
0
2
10
2
1
29
0
2
18
om1
≤1
0
3
9
5
0
0
1
64
20
0
2
19
0
3
7

RPART tree for heart disease data



Ctree (partykit) tree for heart disease data



Leave-one-out error counts for heart disease data

Method	Errors[a]	Time[b]
GUIDE forest	69	13.50
GUIDE tree, linear splits	84	1.13
GUIDE tree, univ1 splits	93	0.90
GUIDE tree, univ2 splits	94	0.87
Ctree (partykit)	212	0.50
RPART	214	0.08
Cforest (partykit)	220	43.56
GUIDE tree, bivariate NN	332	1.25
GUIDE tree, bivariate kernel	336	1.30

[a] out of 611 observations, to enable cforest

[b] average time (sec.) to fit one data set

Low birth weight data: Missing values and highly unbalanced classes

- Data from 2016 CDC Natality Public Use File
- Birth weight and 121 other variables for 3,956,112 births in U.S. in 2016
- 8.15% have low birth weight (less than 2500 gm $\approx$ 5.5 lbs)
- 99.6% of subjects have missing values
- Goal: what factors and how are they predictive of low birth weight?

Some variables and numbers of missing values

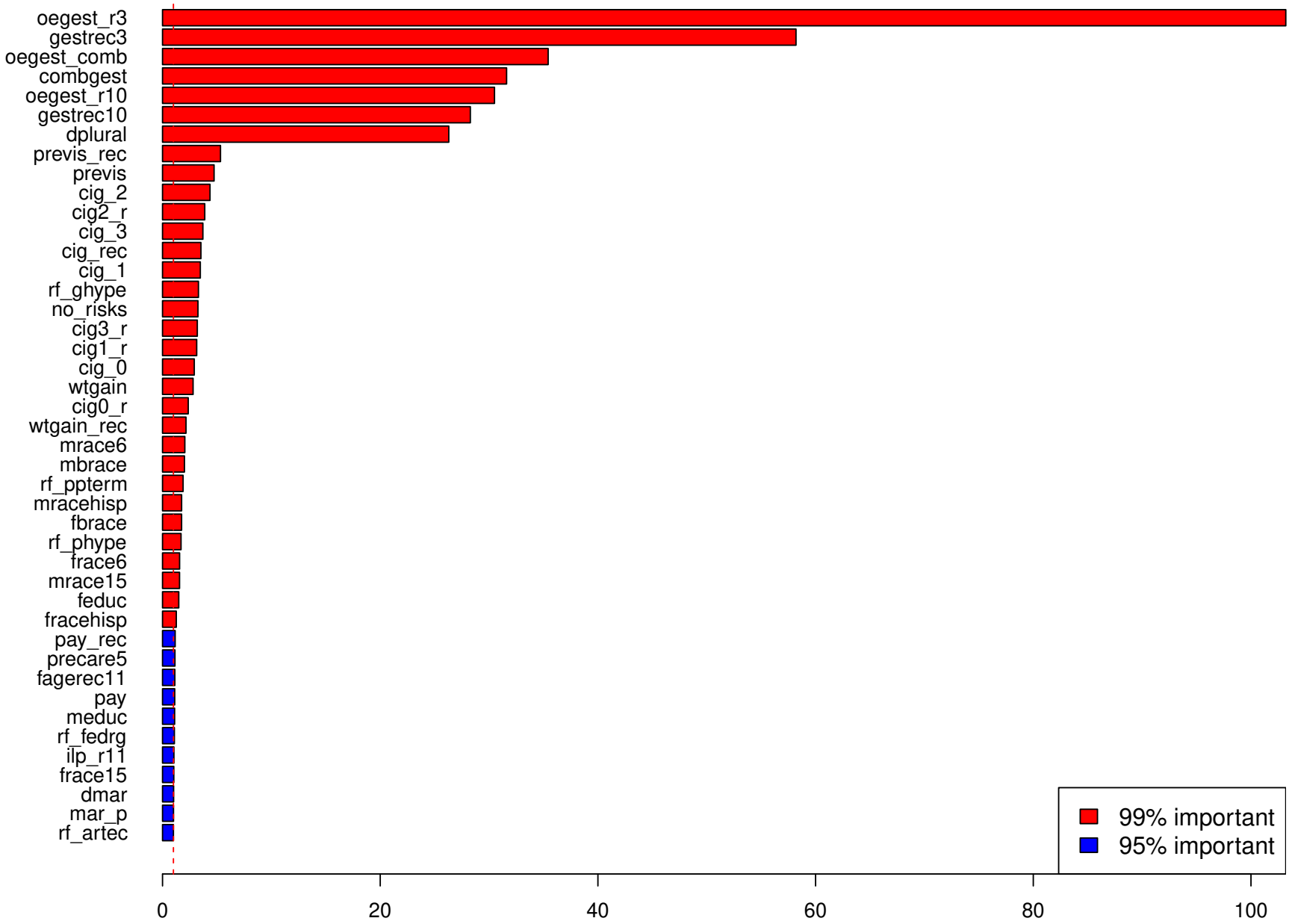
dmar	Marital status [**0**]
dplural	Plurality: 1=single, 2=twin, 3=triplet, 4=quadruplet, 5=quintuplet or higher [**0**]
dwgt_r	Mother's delivery weight in pounds [**70,304**]
m_ht_in	Mother's height in inches [**28,356**]
wtgain	Weight gain in lbs [**143,049**]
mbrace	Mother's race: 1=White, 2=Black, 3=American Indian or Alaskan Native, 4=Asian or Pacific Islander [**0**]
fbrace	Father's race [**0**]
meduc	Mother's education [**51,721**]
feduc	Father's education: 1=8th grade or less, 2=9–12th grade with no diploma, . . . , 8=doctorate or professional degree [**555,897**]
mar_p	Paternity acknowledged [**0**]

combgest	Combined gestation in weeks [**3,516**]
gestrec3	combgest recode: 1=under 37 weeks, 2=37 weeks or more [**0**]
gestrec10	combgest recode: 1=under 20 weeks, 2=20-27, 3=28-31, 4=32-33, 5=34-36, 6=37-38, 7=39, 8=40, 9=41, 10=42 or more [**3,516**]
oegest_comb	Obstetric estimate of combined gestation in weeks [**3,513**]
oegest_r10	oegest_comb recode: 1=under 20 wks, 2=20-27, 3=28-31, 4=32-33, 5=34-36, 6=37-38, 7=39, 8=40, 9=41, 10=42 and over [**3,513**]
oegest_r3	oegest_comb recode: 1=under 37 weeks, 2=37 weeks and over, 3=not stated [**0**]
ilp_r11	Interval since last pregnancy [**2,215,402**]
ilop_r	Interval since last other pregnancy [**3,293,032**]
illb_r	Interval since last live birth: 0-3=plural delivery, 4-300=months since last live birth [**1,649,343**]

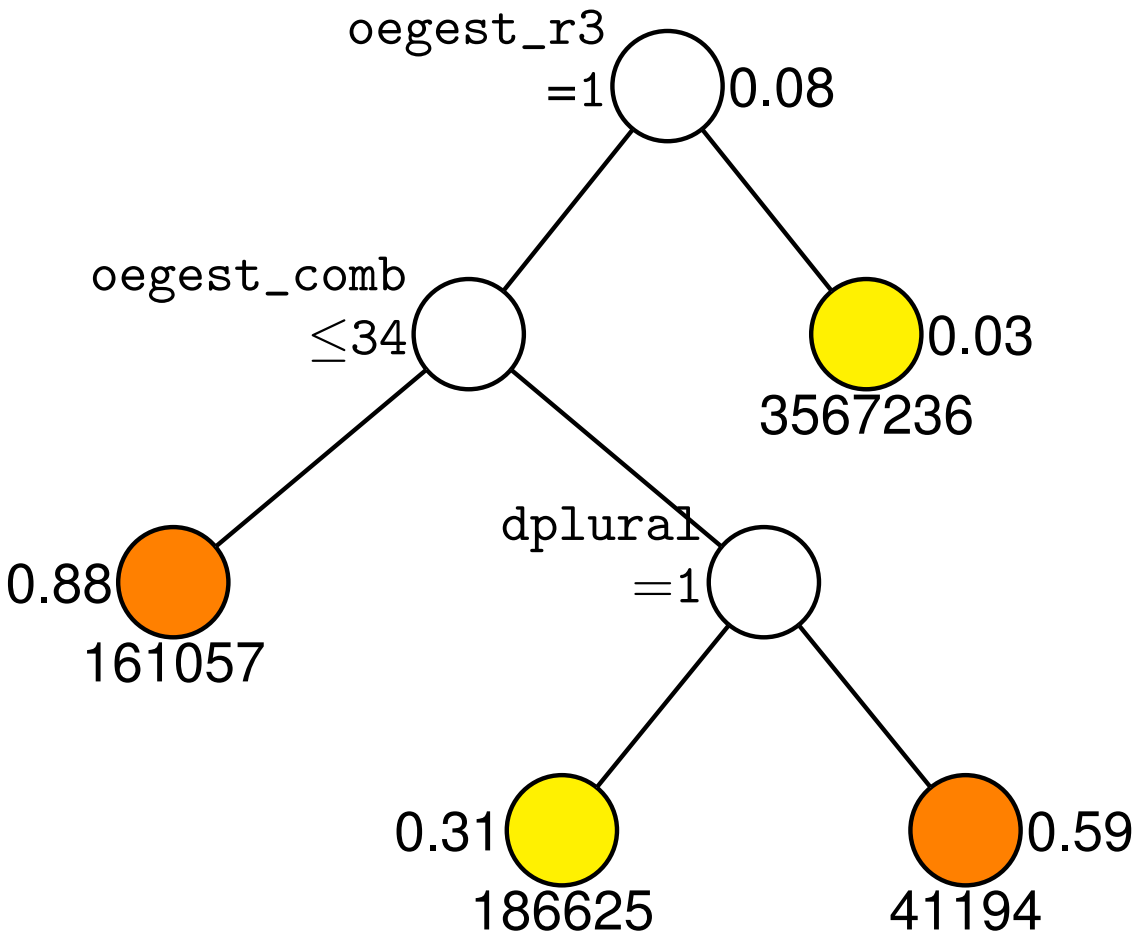
precare5	Month prenatal care began: 1=1st-3rd month, 2=4th-6th , 3=7th to final month, 4=no prenatal care [**113,394**]
previs	Number of prenatal visits [**112,704**]
previs_rec	previs recode: 1=no visits, 2=1-2 visits, 3=3-4, 4=5-6, 5=7-8, 6=9-10, 7=11-12, 8=13-14, 9=15-16, 10=17-18, 11=19 or more [**112,704**]
cig_0	Daily number of cigarettes before pregnancy [**19,350**]
cig_1	Daily number of cigarettes during 1st trimester [**19,719**]
cig_2	Daily number of cigarettes during 2nd trimester [**19,985**]
cig_3	Daily number of cigarettes during 3rd trimester [**20,035**]

rf_fedrg	Fertility enhancing drugs [**0**]
rf_ghype	Risk factor for gestational hypertension [**0**]
rf_phype	Risk factor for pre-pregnancy hypertension [**0**]
rf_ppterm	Previous preterm birth [**0**]
rf_artec	Assistive reproductive technology [**0**]
no_risks	No risk factors reported: 1=true, 0=false [**0**]
pay	Payment source for delivery: 1=medicaid, 2=private insurance, 3=self pay, 4=Indian health service, 5=CHAMPUS/TRICARE, 6=other government, 8=other [**0**]
pay_rec	pay recode: 1=medicaid, 2=private insurance, 3=self pay, 4=other [**0**]

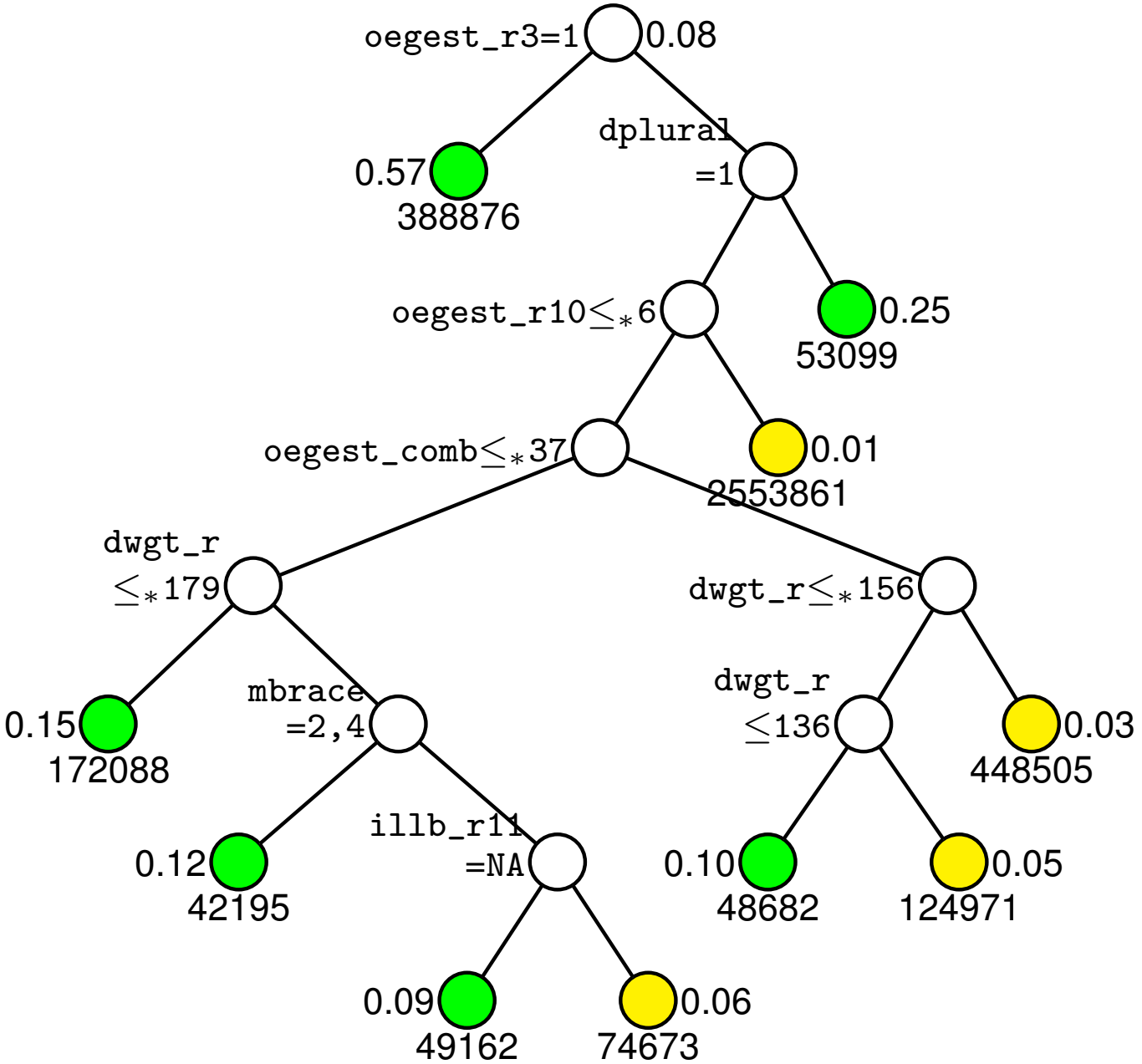
Importance scores (based on subsample of 10,000)



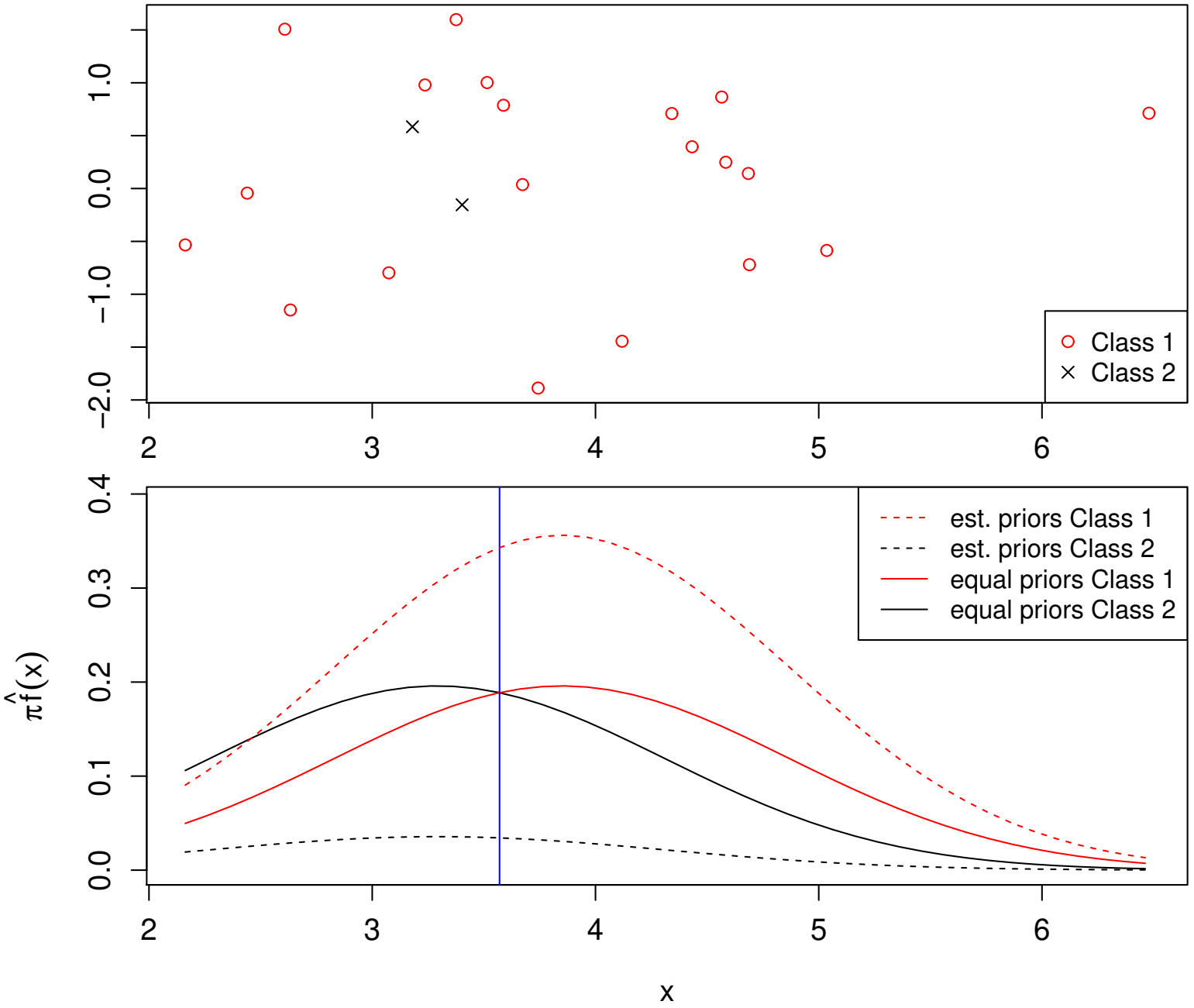
GUIDE classification tree (estimated priors)



GUIDE classification tree (equal priors)



Estimated vs equal priors



Estimated vs equal priors (cont'd.)

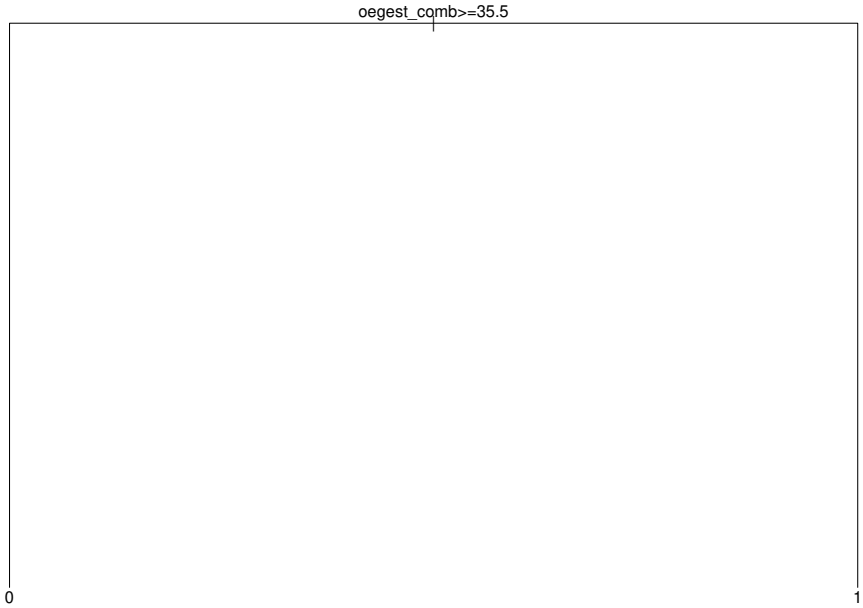
- Let $f_j(x)$ be the pop. density and π_j the prior prob. of Class j ($j = 1, 2$)
- If π_j and f_j are given, x is classified in Class 1 if $\pi_1 f_1(x) > \pi_2 f_2(x)$
- If π_j and f_j are estimated, x is classified in Class 1 if $\hat{\pi}_1 \hat{f}_1(x) > \hat{\pi}_2 \hat{f}_2(x)$
- Let N_j and n_j be the Class j sample sizes in nodes 1 and t, respectively
- Then $\hat{f}_j(x) = n_j/N_j$ for $x \in t$
- If π_j is estimated from the sample, then $\hat{\pi}_j = N_j/(N_1 + N_2)$ and x is predicted to be in Class 1 if

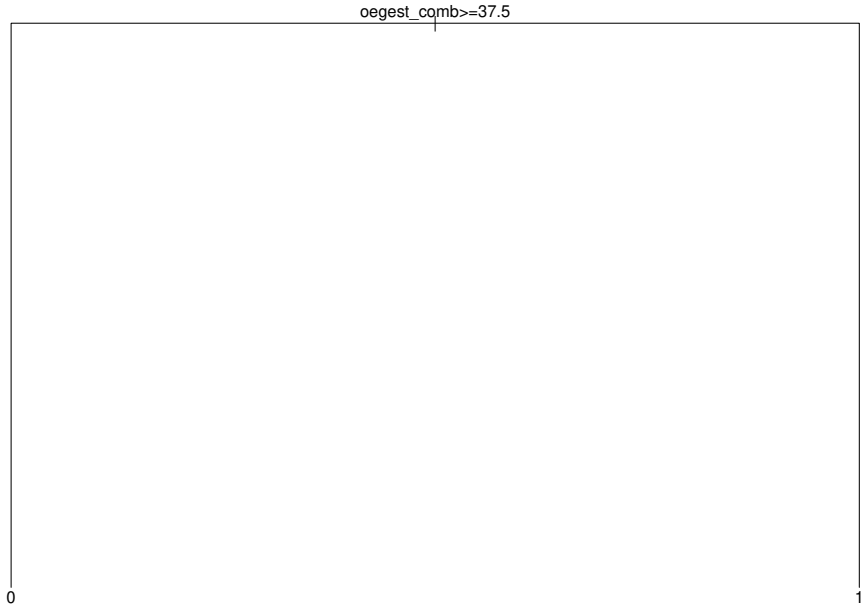
$$\frac{N_1}{N_1 + N_2} \times \frac{n_1}{N_1} > \frac{N_2}{N_1 + N_2} \times \frac{n_2}{N_2} \Longrightarrow n_1 > n_2$$

- If priors are equal, $\pi_j = 0.5$ and x is predicted to be in Class 1 if

$$0.5 \times \frac{n_1}{N_1} > 0.5 \times \frac{n_2}{N_2} \Longrightarrow \frac{n_1}{n_2} > \frac{N_1}{N_2} \Longrightarrow \frac{n_1}{n_1 + n_2} > \frac{N_1}{N_1 + N_2}$$

RPART: estimated (left) and equal priors (right)





Ctree (partykit) tree

- Estimated priors tree has 738 terminal nodes, splitting root node with oegest_comb ≤ 35
- Tree is large because ctree uses p-value rules to decide when to stop splitting—these rules tend to produce large trees when sample size is large
- Ctree does not allow priors to be specified

Logistic regression (based on 1,442,726 complete observations from 43 important predictors)

```
Coefficients: (23 not defined because of singularities)
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 30.9079717  0.2537067 121.826  < 2e-16 ***
oegest_r32   0.1159092  0.0212835   5.446 5.15e-08 ***
gestrec32   -0.1730464  0.0189170  -9.148  < 2e-16 ***
oegest_comb -0.8481357  0.0090998 -93.204  < 2e-16 ***
combgest    -0.0579274  0.0076654  -7.557 4.12e-14 ***
oegest_r10  -0.1888983  0.0183217 -10.310  < 2e-16 ***
gestrec10    0.0093710  0.0130807   0.716 0.473745
dplural      1.7077334  0.0143844 118.721  < 2e-16 ***
previs_rec  -0.0418056  0.0072728  -5.748 9.02e-09 ***
previs       0.0163597  0.0029962   5.460 4.76e-08 ***
cig_2       -0.0191566  0.0082940  -2.310 0.020904 *
cig2_r       0.1475135  0.0583691   2.527 0.011496 *
cig_3        0.0072475  0.0072619   0.998 0.318272
cig_recY     0.3566415  0.0479863   7.432 1.07e-13 ***
```

```
cig_1         0.0052303  0.0060693   0.862 0.388821
cig3_r        0.0553452  0.0525822   1.053 0.292549
cig1_r        0.0171305  0.0505280   0.339 0.734587
cig_0        -0.0024292  0.0035464  -0.685 0.493363
cig0_r        0.0975391  0.0290366   3.359 0.000782 ***
rf_ghypeU    -0.1121082  0.2390046  -0.469 0.639025
rf_ghypeY     0.4793263  0.0171238  27.992  < 2e-16 ***
no_risks1     0.2012862  0.0119289  16.874  < 2e-16 ***
no_risks9            NA         NA      NA       NA
wtgain       -0.0143635  0.0008803 -16.317  < 2e-16 ***
wtgain_rec   -0.0316959  0.0105389  -3.008 0.002634 **
mrace620      0.0218481  0.0831539   0.263 0.792749
mrace630     -0.1089724  0.1832601  -0.595 0.552089
mrace640      0.1088125  0.1470724   0.740 0.459387
mrace641      0.0591567  0.1318474   0.449 0.653665
mrace651      0.0901516  0.2732902   0.330 0.741494
mrace661      0.1061560  0.0861545   1.232 0.217890
mbrace2       0.1196336  0.0716339   1.670 0.094905 .
mbrace3      -0.0265665  0.1270763  -0.209 0.834402
```

```
mbrace4     -0.0600090  0.0910975  -0.659 0.510066
rf_pptermU          NA         NA      NA       NA
rf_pptermY   0.2258868  0.0183812  12.289  < 2e-16 ***
mracehisp2   0.3002853  0.0473262   6.345 2.22e-10 ***
mracehisp3  -0.1279763  0.1457613  -0.878 0.379952
mracehisp4   0.1838108  0.0870212   2.112 0.034665 *
mracehisp5  -0.3173365  0.2601310  -1.220 0.222499
mracehisp6   0.0206340  0.0859601   0.240 0.810298
mracehisp7  -0.0253430  0.0194193  -1.305 0.191878
mracehisp8   0.0340556  0.0961416   0.354 0.723171
fbrace2      0.1321014  0.0799259   1.653 0.098372 .
fbrace3     -0.1099508  0.1045130  -1.052 0.292786
fbrace4      0.2712343  0.0892163   3.040 0.002364 **
fbrace9     -0.0205748  0.0251942  -0.817 0.414131
rf_phypeU           NA         NA      NA       NA
rf_phypeY    0.2721160  0.0291897   9.322  < 2e-16 ***
frace62      0.0086537  0.1042047   0.083 0.933816
frace63     -0.0097828  0.1876956  -0.052 0.958433
frace64      0.0903098  0.2027387   0.445 0.655995
```

```
frace65     -0.5127737  0.3262788  -1.572 0.116048
frace66     -0.0347022  0.0916966  -0.378 0.705099
frace69             NA         NA      NA       NA
mrace1510           NA         NA      NA       NA
mrace1511    0.2269873  0.3311048   0.686 0.493000
mrace1512    0.1509662  0.3255528   0.464 0.642847
mrace1513   -0.2874930  0.2941522  -0.977 0.328390
mrace1514           NA         NA      NA       NA
mrace1515           NA         NA      NA       NA
mrace152            NA         NA      NA       NA
mrace153            NA         NA      NA       NA
mrace154     0.2890144  0.1085393   2.663 0.007750 **
mrace155    -0.1976183  0.1022356  -1.933 0.053240 .
mrace156    -0.1121010  0.1013933  -1.106 0.268898
mrace157     0.1694560  0.1378367   1.229 0.218923
mrace158    -0.0798455  0.1248463  -0.640 0.522465
mrace159            NA         NA      NA       NA
feduc        0.0065593  0.0040658   1.613 0.106684
fracehisp2   0.0258961  0.0691855   0.374 0.708182
```

```
fracehisp3  -0.0064213  0.1675668  -0.038 0.969432
fracehisp4  -0.0428494  0.1592642  -0.269 0.787894
fracehisp5   0.1101917  0.3085560   0.357 0.721001
fracehisp6   0.1149577  0.0906743   1.268 0.204866
fracehisp7   0.0970886  0.0201733   4.813 1.49e-06 ***
fracehisp8   0.3334949  0.0814594   4.094 4.24e-05 ***
fracehisp9   0.1887047  0.0897306   2.103 0.035465 *
pay_rec2    -0.0554420  0.0128347  -4.320 1.56e-05 ***
pay_rec3    -0.0937277  0.0269719  -3.475 0.000511 ***
pay_rec4     0.0166405  0.0340255   0.489 0.624800
pay_rec9     0.0123680  0.0758327   0.163 0.870443
precare5     0.0357404  0.0093247   3.833 0.000127 ***
fagerec11   -0.0120939  0.0037743  -3.204 0.001354 **
pay2                NA         NA      NA       NA
pay3                NA         NA      NA       NA
pay4         0.0581175  0.1819609   0.319 0.749426
pay5        -0.0133710  0.0552914  -0.242 0.808913
pay6         0.0210235  0.0625980   0.336 0.736985
pay8                NA         NA      NA       NA
```

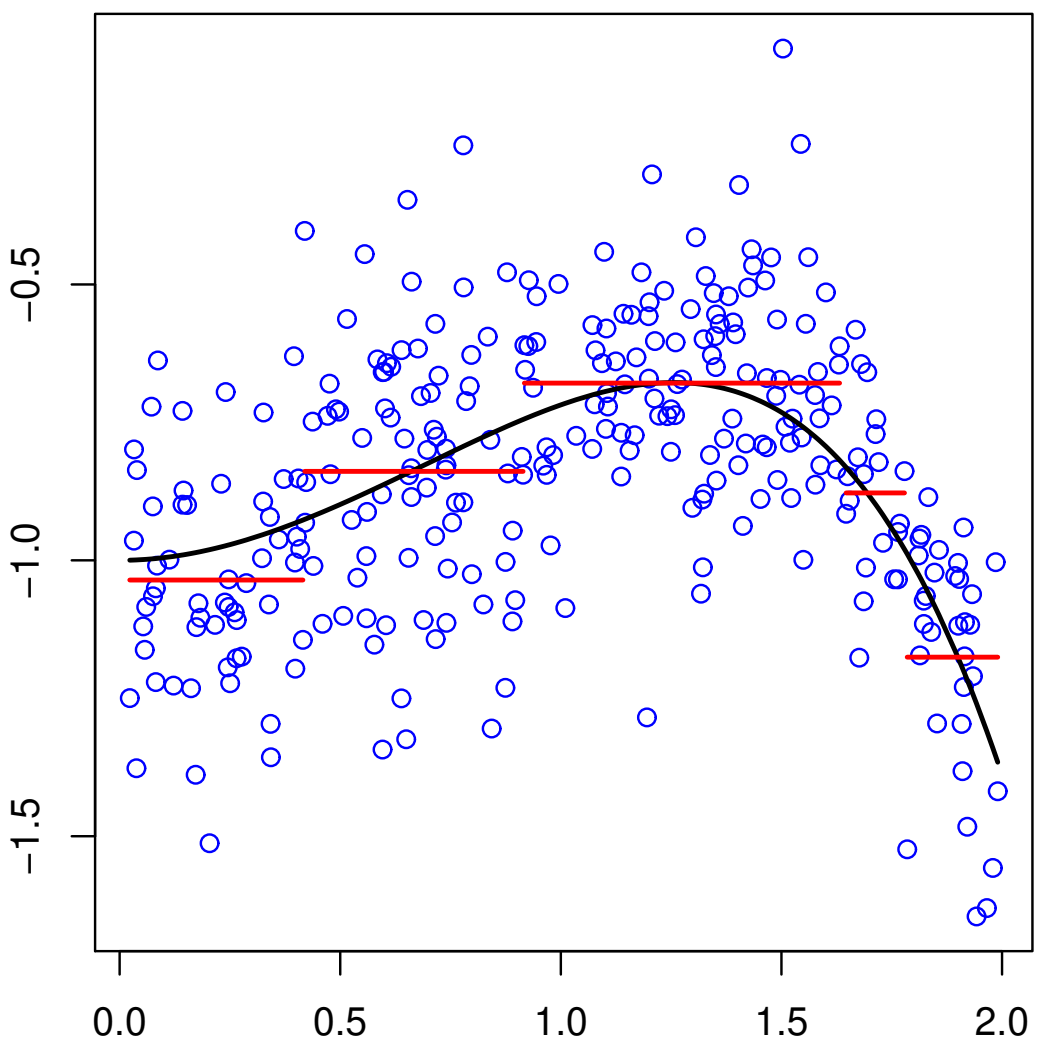
```
pay9              NA          NA      NA       NA
meduc       -0.0248809   0.0041822  -5.949 2.69e-09 ***
rf_fedrgU    0.0217534   0.1366780   0.159 0.873544
rf_fedrgX   -0.1637146   0.0988164  -1.657 0.097569 .
rf_fedrgY    0.1329961   0.0892414   1.490 0.136146
ilp_r11     -0.0351318   0.0029695 -11.831  < 2e-16 ***
frace1510    0.1361641   0.1094445   1.244 0.213449
frace1511    0.3924095   0.3103599   1.264 0.206098
frace1512    0.1997478   0.3280464   0.609 0.542590
frace1513    0.0356708   0.2730314   0.131 0.896054
frace1514         NA          NA      NA       NA
frace1515         NA          NA      NA       NA
frace152          NA          NA      NA       NA
frace153          NA          NA      NA       NA
frace154     0.1768129   0.1171161   1.510 0.131114
frace155     0.0914581   0.1120570   0.816 0.414400
frace156     0.1905527   0.1140431   1.671 0.094745 .
frace157     0.1763113   0.1666884   1.058 0.290179
frace158    -0.0094585   0.1409423  -0.067 0.946495
```

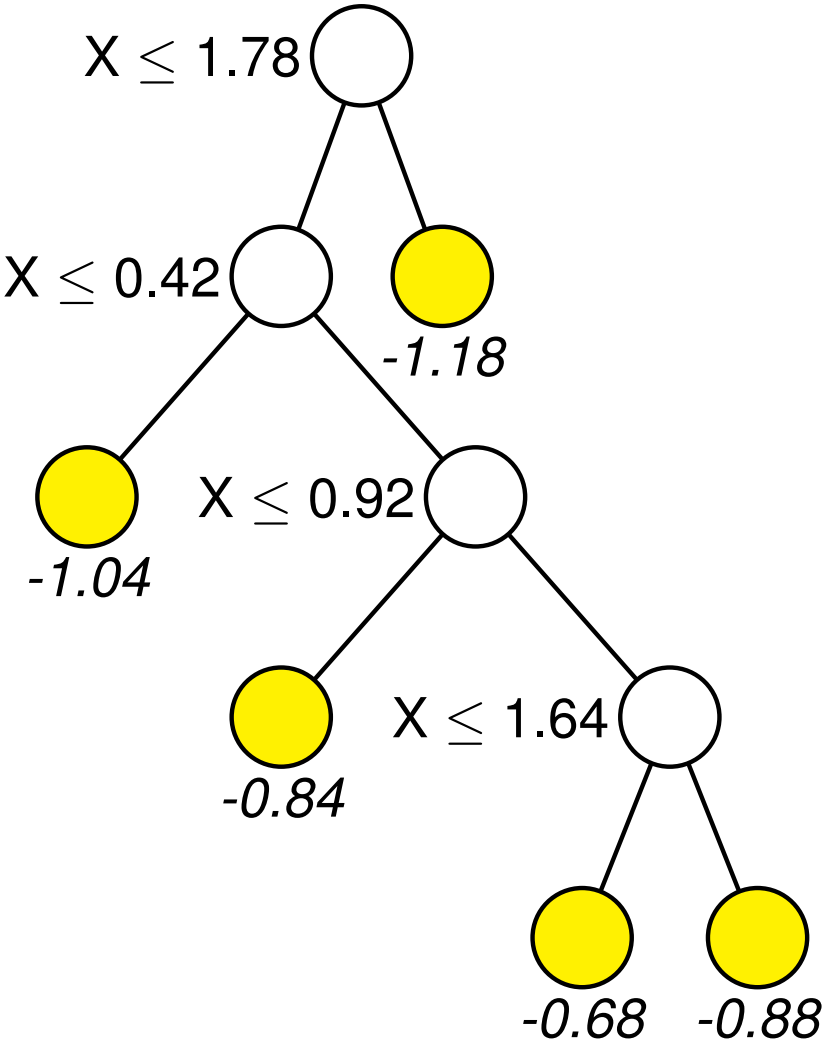
```
frace159            NA          NA      NA       NA
frace1599           NA          NA      NA       NA
dmar2        0.2149034   0.1039195   2.068 0.038642 *
mar_pU      -1.2340646   1.1466784  -1.076 0.281834
mar_pX              NA          NA      NA       NA
mar_pY      -0.0559210   0.1038528  -0.538 0.590257
rf_artecU           NA          NA      NA       NA
rf_artecX           NA          NA      NA       NA
rf_artecY    0.0162279   0.0926477   0.175 0.860956
```

CART regression

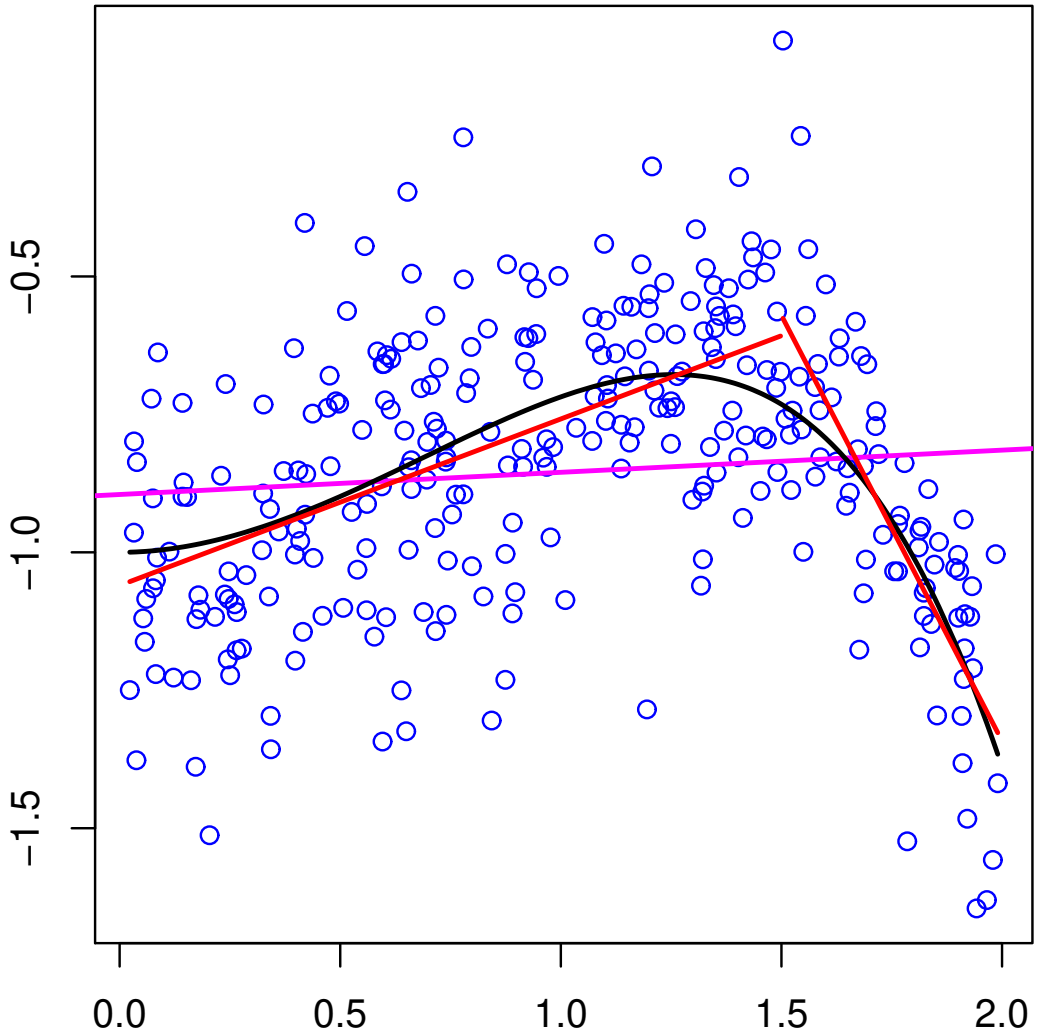
- Fit a constant $\bar{y}$ to each node
- Use residual sum of squares as node impurity and error measure
- Everything else the same as in CART classification

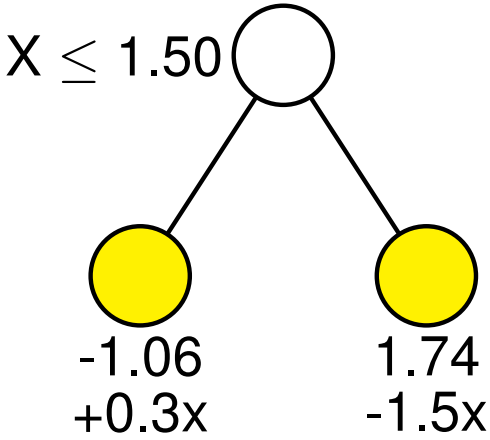
Piecewise-constant regression model





Piecewise-linear regression model





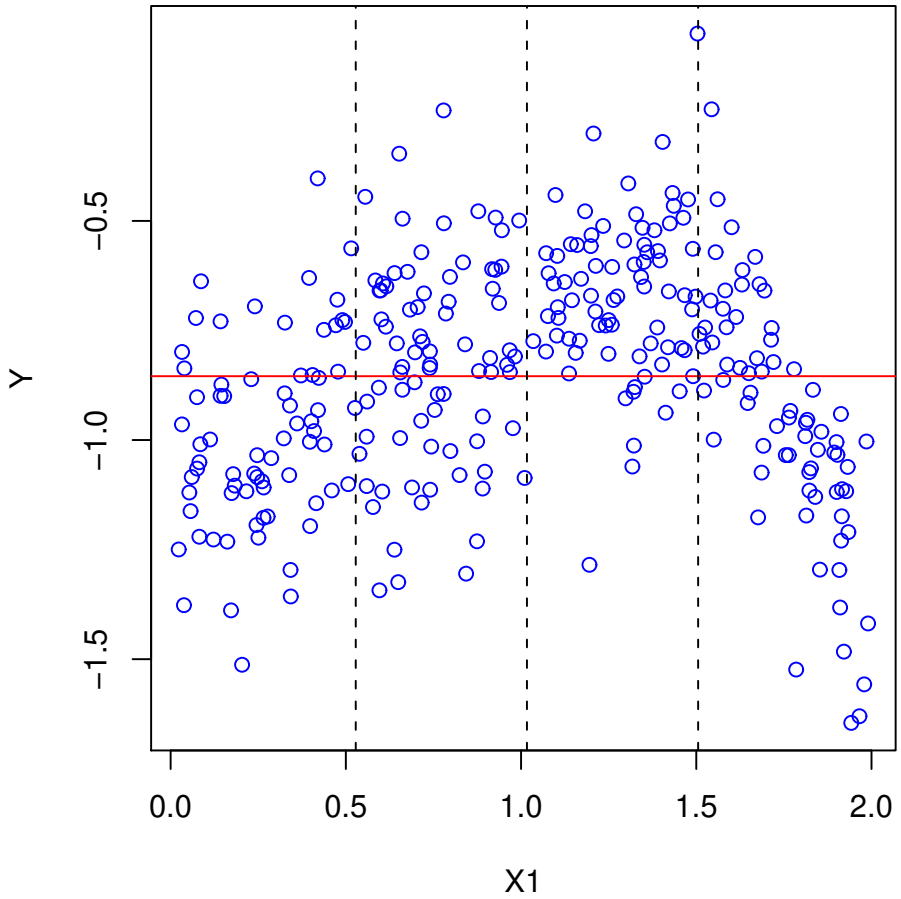
GUIDE regression tree models

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

GUIDE variable selection for regression

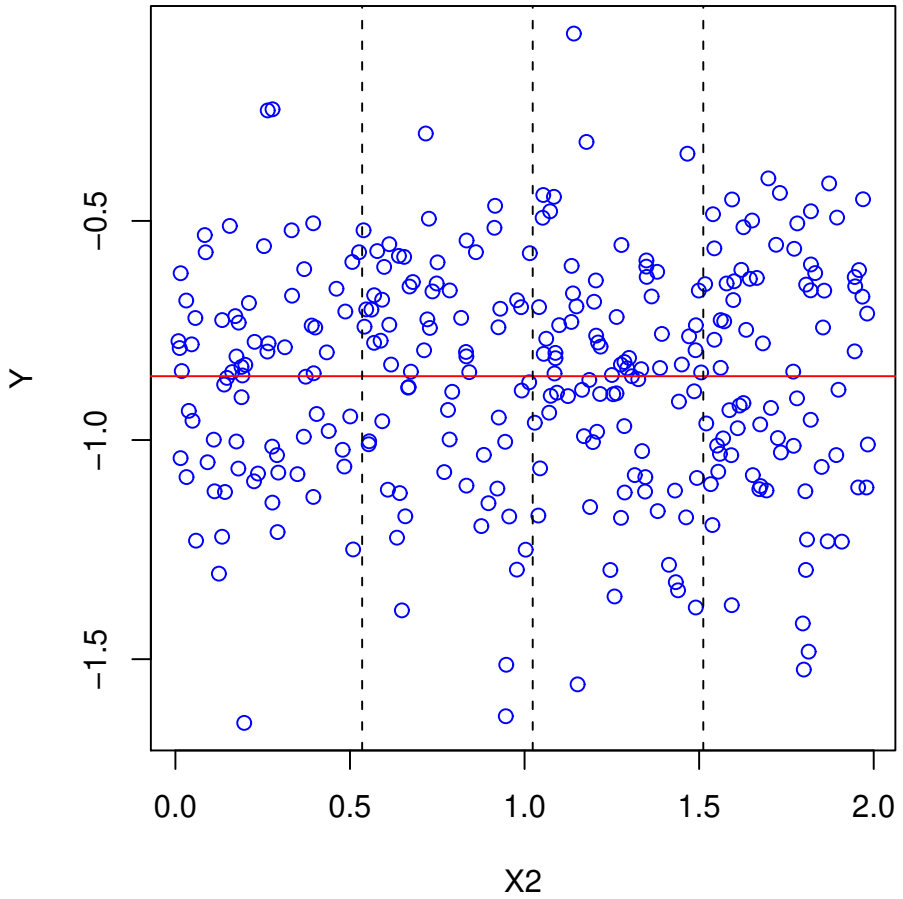
1. Fit a model to the data in the node and obtain the residuals
2. Define a “class” variable that equals +1 if residual is positive, -1 otherwise
3. Follow GUIDE classification procedure to select a variable to split node

Split variable selection based on residual patterns



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

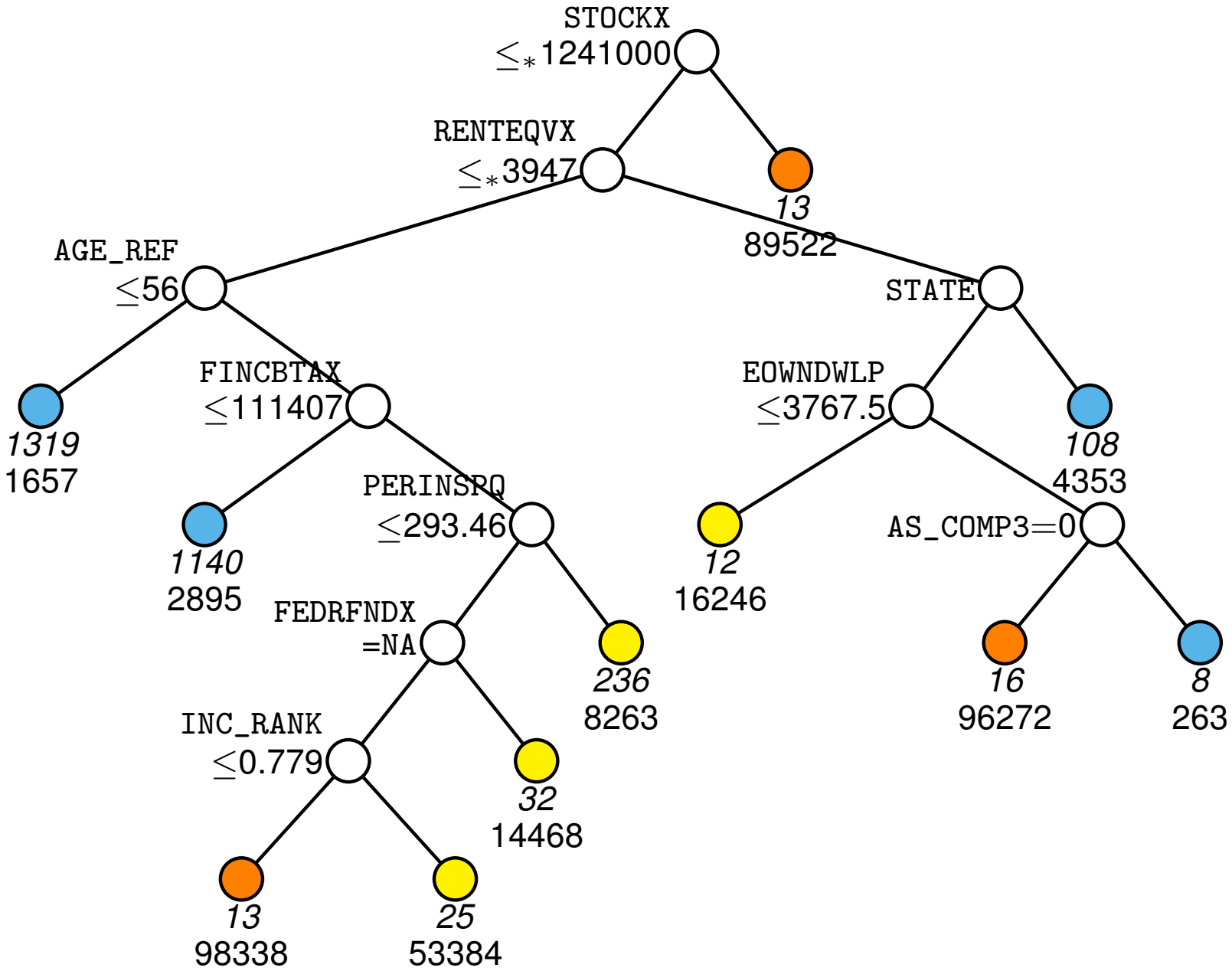
χ^2_3 = 66.7, p = 2×10^{-14}



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

χ^2_3 = 1.14, p = 0.77

GUIDE tree for predicting INTRDVX

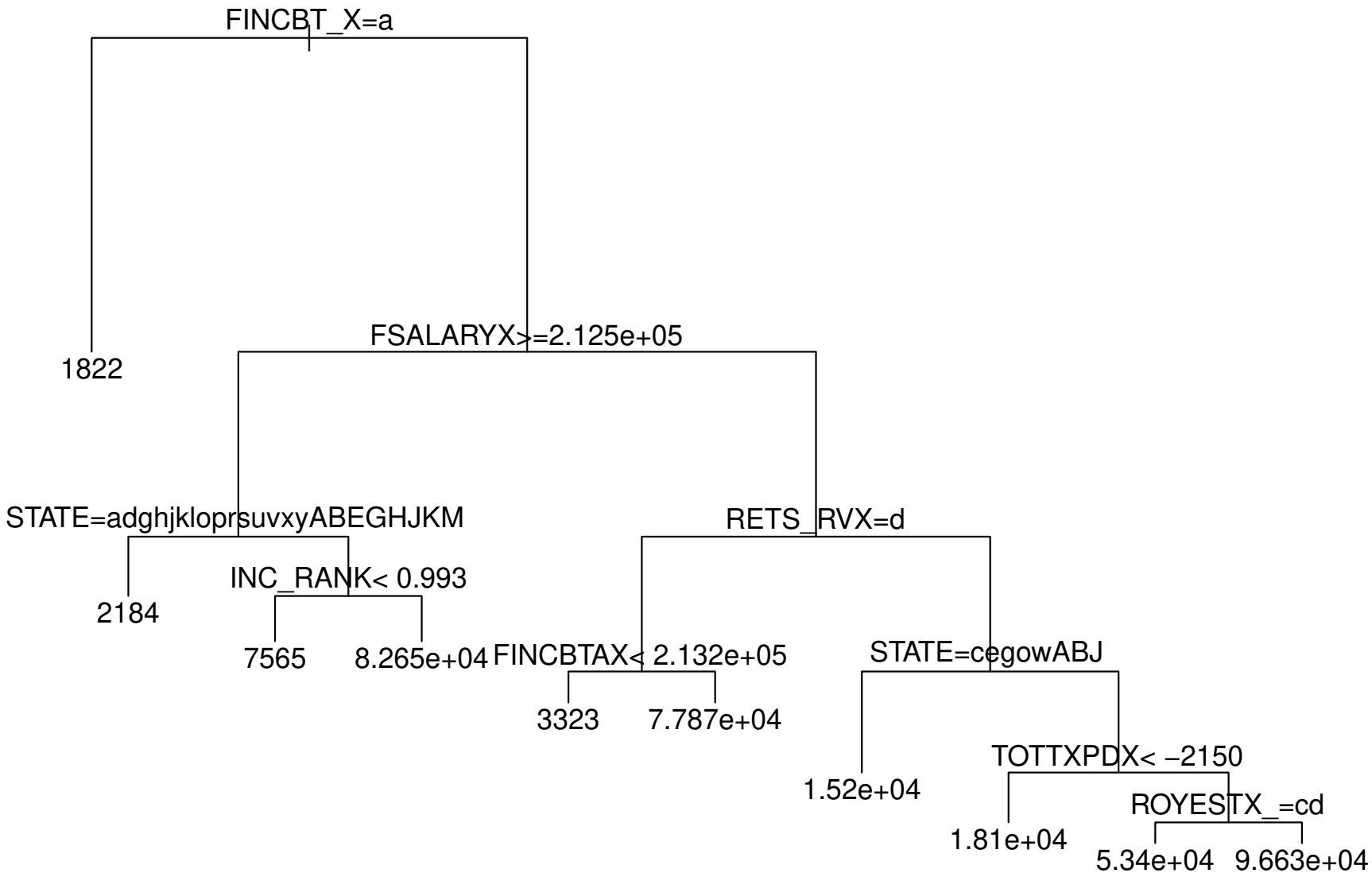


BLS data (weighted mean INTRDVX = $4778)

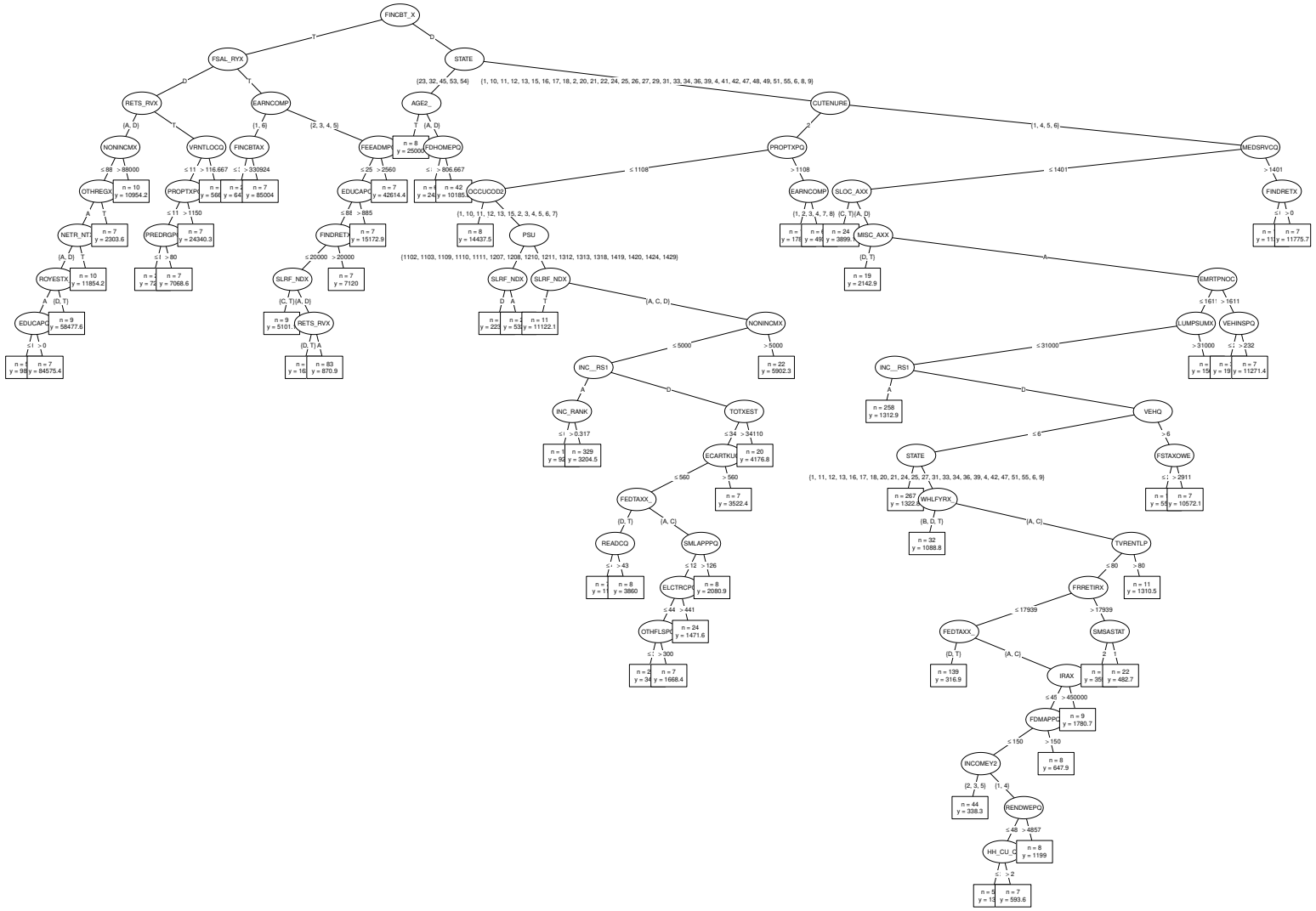
	AGE_REF			
	$\leq$ 43	(43, 58]	(58, 68]	$>$ 68
$>$ \$4778	33	78	152	147
$\leq$ \$4778	718	684	571	539
χ^2_3 = 127.3, p $<$ 2.2E-16, χ^2_1 = 108.4				

	STOCKX			
	$\leq$ 18000	(18000, 133333]	$>$ 133333	NA
$>$ \$4778	3	10	53	344
$\leq$ \$4778	79	67	27	2339
χ^2_3 = 191.5, p $<$ 2.2E-16, χ^2_1 = 168.1				

RPART regression tree



Ctree (party) regression tree without FINLWT21



`partykit` does not work here; neither allows sampling weights

References

Aitkin, M. and Clayton, D. (1980). The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. *Applied Statistics*, 29:156–163.

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

Bollinger, C. R. and Hirsch, B. T. (2006). Match bias from earnings imputation in the Current Population Survey: the case of imperfect matching. *Journal of Labor Economics*, 24:483–519.

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.

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

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

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

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.

Connors, Jr., A. F., , Speroff, T., Dawson, N. V., et al. (1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. *JAMA*, 276(11):889–897.

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

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

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

Henderson, N. C., Louis, T. A., Rosner, G. L., and Varadhan, R. (2017). Individualized treatment effects with censored data via fully nonparametric Bayesian accelerated failure time models. *ArXiv e-print 1706.06611*.

Honaker, J., King, G., and Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of Statistical Software*, 45(7):1–47.

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

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

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

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

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

Laird, N. and Olivier, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, 76:231–240.

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

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

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

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

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

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

Loh, W.-Y. (2014). Fifty years of classification and regression trees (with discussion). *International Statistical Review*, 34:329–370.

Loh, W.-Y. (2019). Logistic regression tree analysis. In Pham, H., editor, *Handbook of Engineering Statistics*. Springer, 2nd edition. To appear.

Loh, W.-Y., Eltinge, J., Cho, M. J., and Li, Y. (2019a). Classification and regression trees and forests for incomplete data from sample surveys. *Statistica Sinica*, 29:431–453.

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

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

Loh, W.-Y., Man, M., and Wang, S. (2019b). Subgroups from regression trees with adjustment for prognostic effects and post-selection inference. *Statistics in Medicine*, 38:545–557.

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

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

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

Negassa, A., Ciampi, A., Abrahamowicz, M., Shapiro, S., and Boivin, J. R. (2005). Tree-structured subgroup analysis for censored survival data:

validation of computationally inexpensive model selection criteria. *Statistics and Computing*, 15:231–239.

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

Schumacher, M., Baster, G., Bojar, H., Hübner, K., Olschewski, M., Sauerbrei, W., Schmoor, C., Beyerle, C., Newmann, R. L. A., and Rauschecker, H. F. (1994). Randomized 2 $\times$ 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. *Journal of Clinical Oncology*, 12:2086–2093.

SOLVD Investigators (1991). Effect of Enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure. *New England Journal of Medicine*, 325(5):293–302.

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

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

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3):1–67.

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