Classification and Regression Trees by Example

(Tutorial at 2021 Causal Inference with Big Data Workshop hosted by NUS Institute for Mathematical Sciences)

Professor Wei-Yin Loh

Department of Statistics
University of Wisconsin, Madison

Examples

1. Death from COVID-19 for hospitalized patients (observational study)
2. Soldering circuit boards (factorial experiment; Poisson regression)
3. Weights of new-borns in U.S. (missing values; large sample)
4. Consumer expenditure survey (more than one missing-value code)
5. Breast cancer (subgroup identification; censored response)
6. Diabetes (subgroup identification; longitudinal response)
7. Alzheimer's disease (subgroup identification; clustering sample paths)
8. Right heart catheterization (observational data; causal inference)

COVID-19

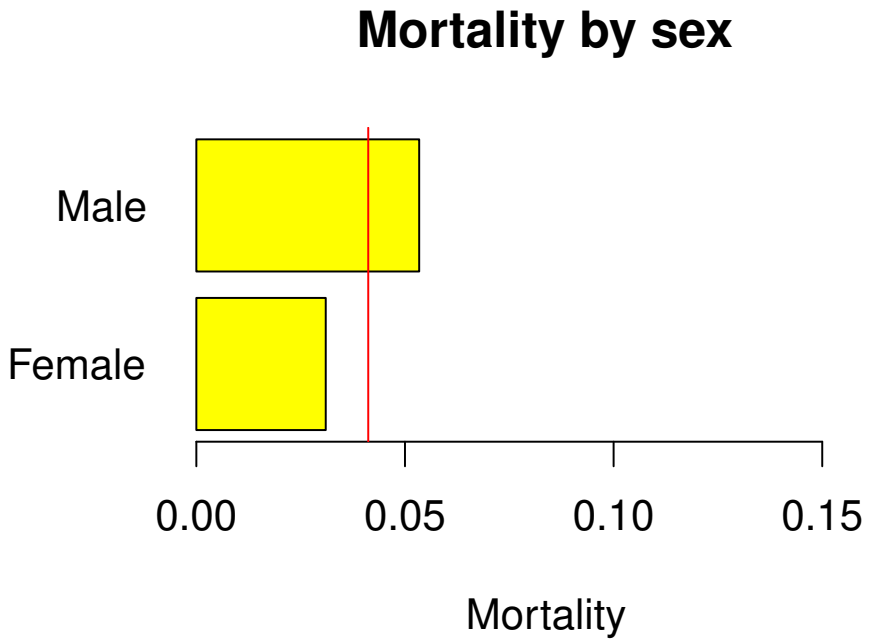
- 31461 patients hospitalized with COVID-19 from Jan 20–May 26, 2020, in USA (Harrison et al., 2020)
- 20 variables:
  - death during hospitalization (4.1% mortality)
  - 5 age groups
  - sex
  - race (6 values)
  - 15 comorbidities (yes/no)
  - Charlson comorbidity index (weighted sum of comorbidities)

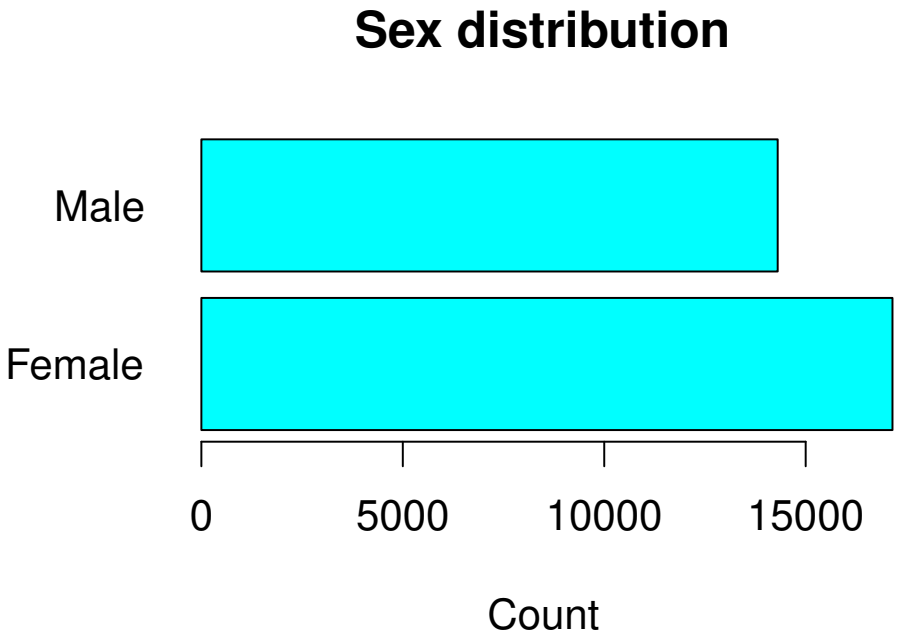
COVID-19 variables

`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)
`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)

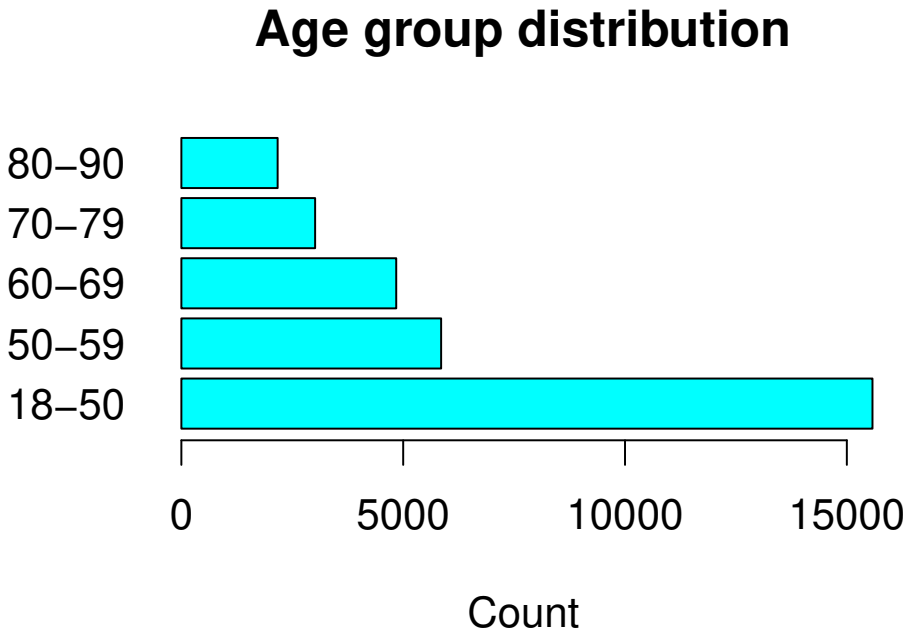
COVID-19 variables (cont'd.)

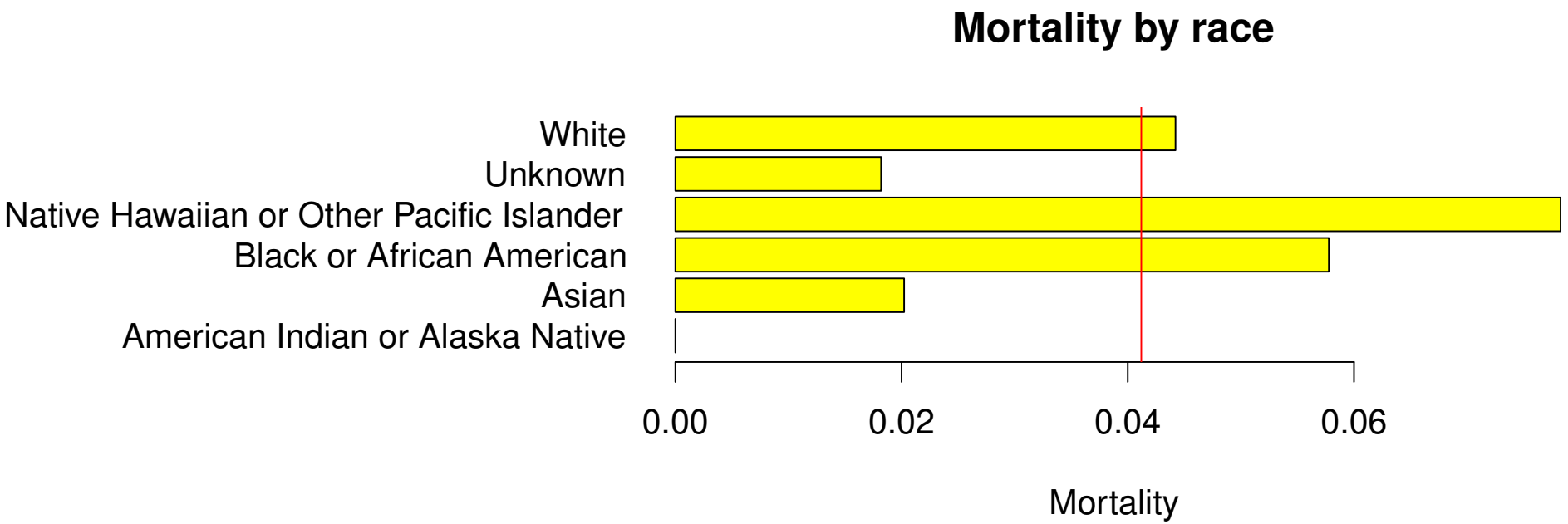
`hemipara`	Hemiplegia or paraplegia (0=no, 1=yes)
`metastatic`	Metastatic solid tumor (0=no, 1=yes)
`MI`	Myocardial infarction (0=no, 1=yes)
`liver`	Liver disease (0=none, 1=mild, 2=severe)
`PUD`	Peptic ulcer disease (0=no, 1=yes)
`PVD`	Peripheral vascular disease (0=no, 1=yes)
`RD`	Rheumatic disease (0=no, 1=yes)
`renal`	Renal disease (0=no, 1=yes)
`charlson`	CHF + CPD + MI + RD + PUD + PVD + cerebro + dementia + diabetes + I(liver=1) + 2$\times$(cancer + hemipara + renal) + 3$\times$I(liver=2) + 6$\times$(metastatic + aids)

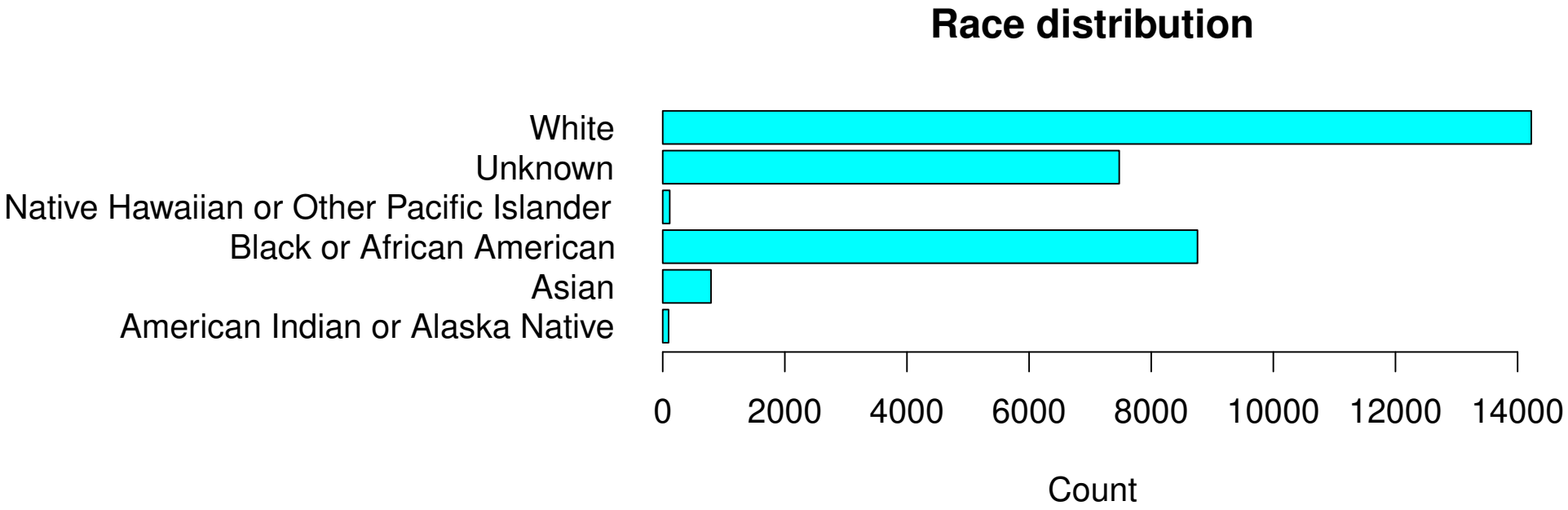

Mortality by sex
Male
Female
0.00
0.05
0.10
0.15
Mortality

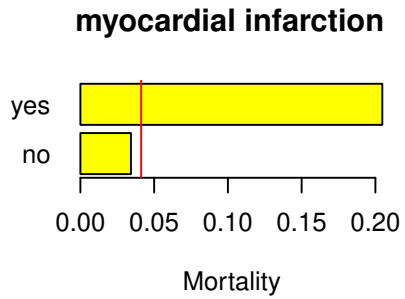

Sex distribution
Male
Female
0
5000
10000
15000
Count

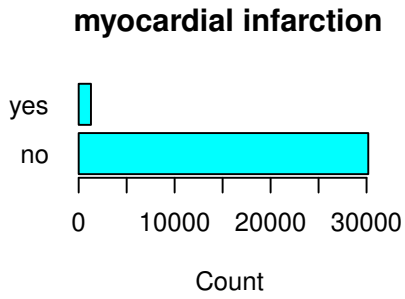

Mortality by age group
80–90
70–79
60–69
50–59
18–50
0.00
0.05
0.10
0.15
Mortality

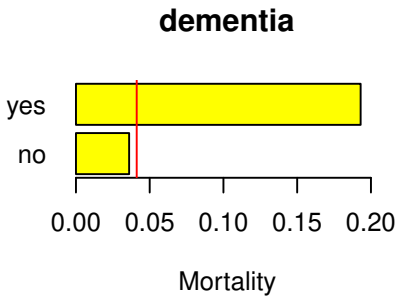

Age group distribution
80–90
70–79
60–69
50–59
18–50
0
5000
10000
15000
Count

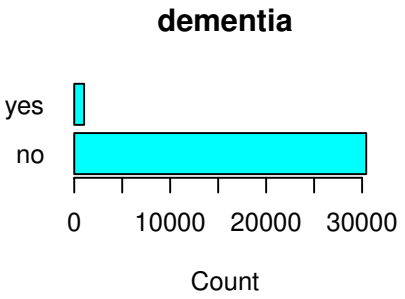

Mortality by race
White
Unknown
Native Hawaiian or Other Pacific Islander
Black or African American
Asian
American Indian or Alaska Native
0.00
0.02
0.04
0.06
Mortality

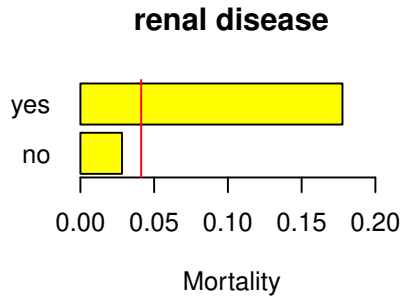

Race distribution
White
Unknown
Native Hawaiian or Other Pacific Islander
Black or African American
Asian
American Indian or Alaska Native
0
2000
4000
6000
8000
10000
12000
14000
Count

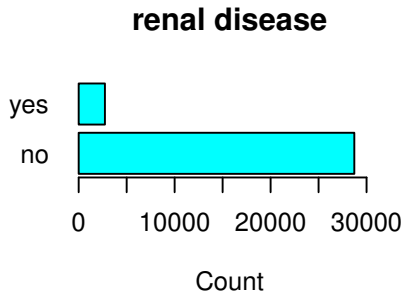

myocardial infarction
yes
no
0.00 0.05 0.10 0.15 0.20
Mortality

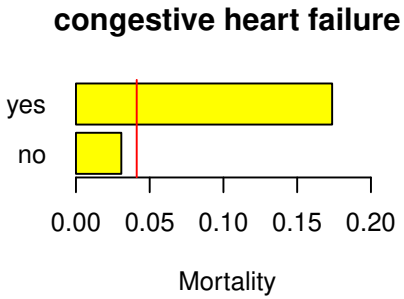

myocardial infarction
yes
no
0 10000 20000 30000
Count

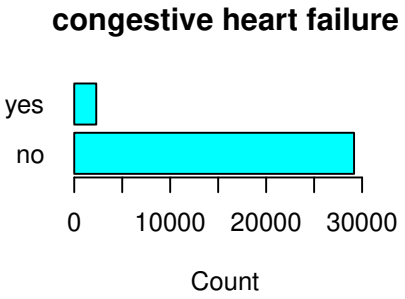

dementia
yes
no
0.00 0.05 0.10 0.15 0.20
Mortality

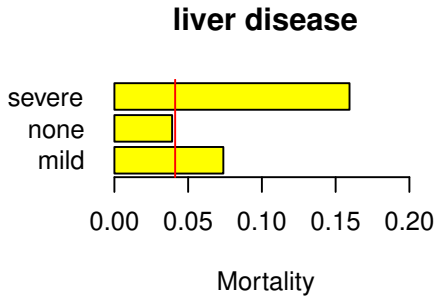

dementia
yes
no
0 10000 20000 30000
Count

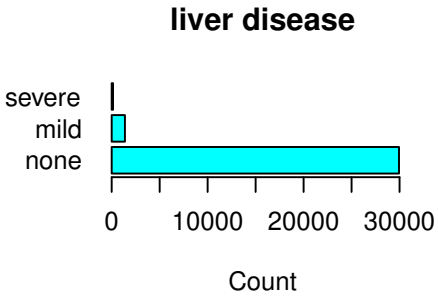

renal disease
yes
no
0.00 0.05 0.10 0.15 0.20
Mortality

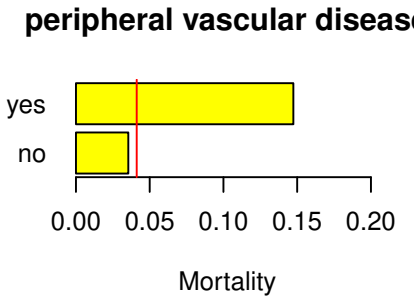

renal disease
yes
no
0 10000 20000 30000
Count

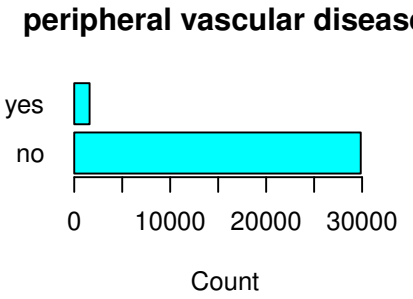

congestive heart failure
yes
no
0.00 0.05 0.10 0.15 0.20
Mortality

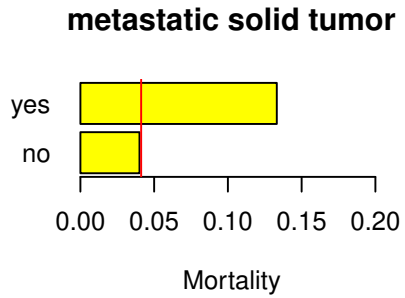

congestive heart failure
yes
no
0 10000 20000 30000
Count

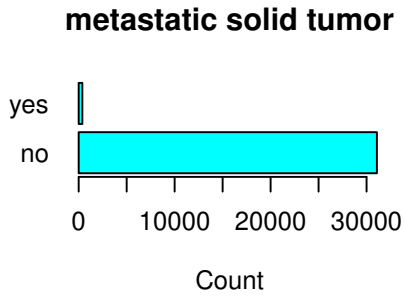

liver disease
severe
none
mild
0.00 0.05 0.10 0.15 0.20
Mortality

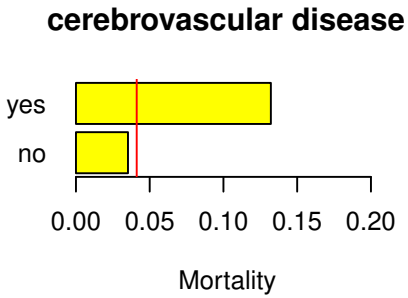

liver disease
severe
mild
none
0 10000 20000 30000
Count

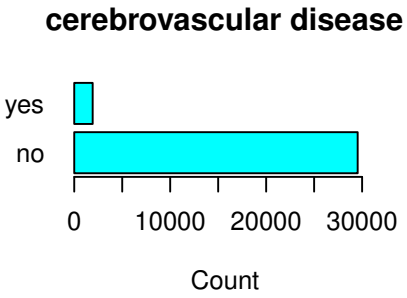

peripheral vascular disease
yes
no
0.00 0.05 0.10 0.15 0.20
Mortality

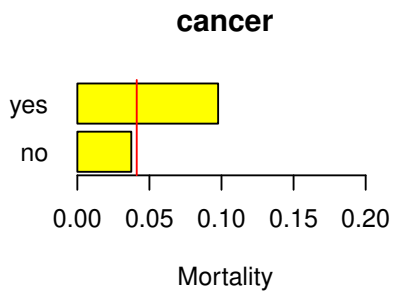

peripheral vascular disease
yes
no
0 10000 20000 30000
Count

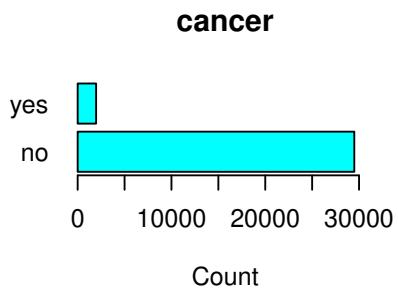

metastatic solid tumor
yes
no
0.00 0.05 0.10 0.15 0.20
Mortality

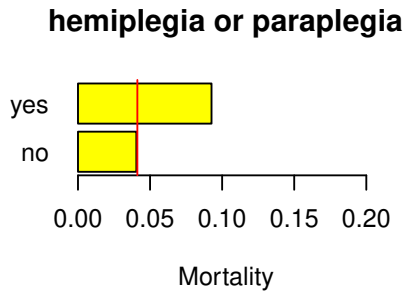

metastatic solid tumor
yes
no
0 10000 20000 30000
Count

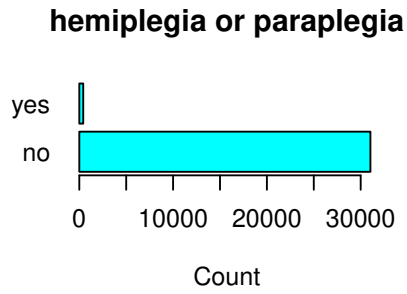

cerebrovascular disease
yes
no
0.00 0.05 0.10 0.15 0.20
Mortality

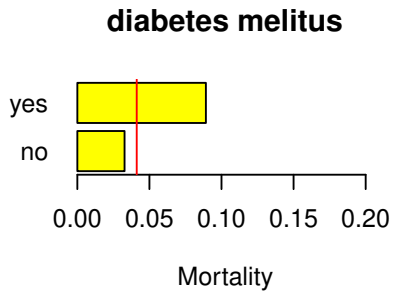

cerebrovascular disease
yes
no
0 10000 20000 30000
Count

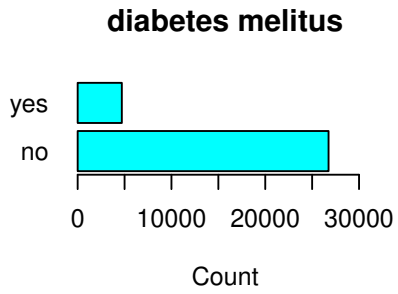

cancer
yes
no
0.00 0.05 0.10 0.15 0.20
Mortality

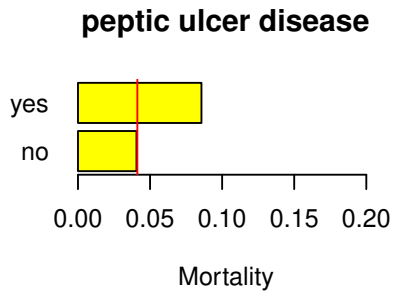

cancer
yes
no
0 10000 20000 30000
Count

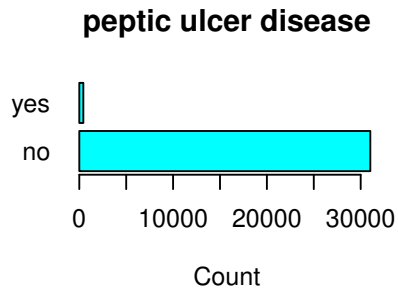

hemiplegia or paraplegia
yes
no
0.00 0.05 0.10 0.15 0.20
Mortality

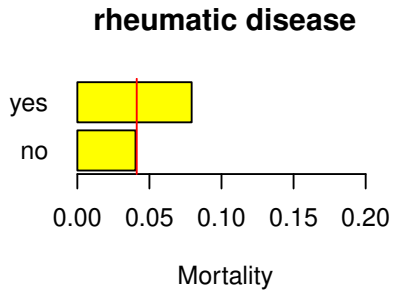

hemiplegia or paraplegia
yes
no
0 10000 20000 30000
Count

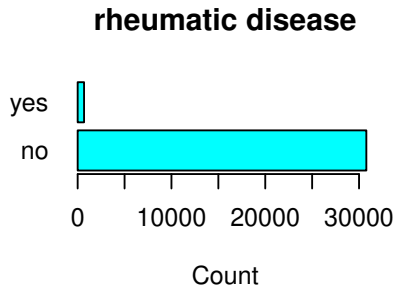

diabetes melitus
yes
no
0.00 0.05 0.10 0.15 0.20
Mortality

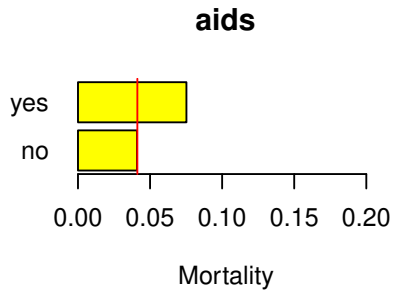

diabetes melitus
yes
no
0 10000 20000 30000
Count

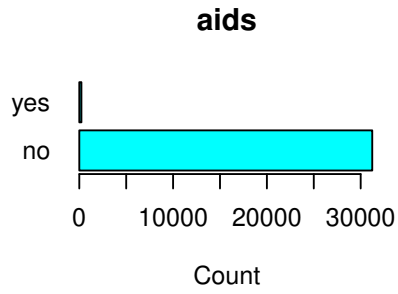

peptic ulcer disease
yes
no
0.00 0.05 0.10 0.15 0.20
Mortality

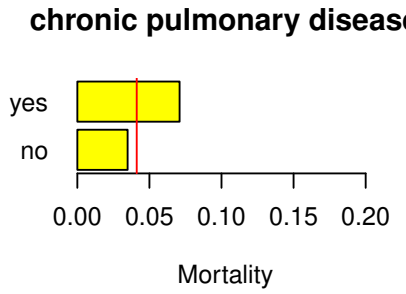

peptic ulcer disease
yes
no
0 10000 20000 30000
Count

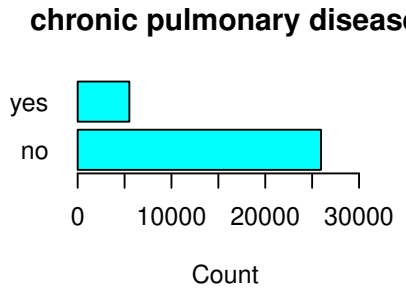

rheumatic disease
yes
no
0.00 0.05 0.10 0.15 0.20
Mortality

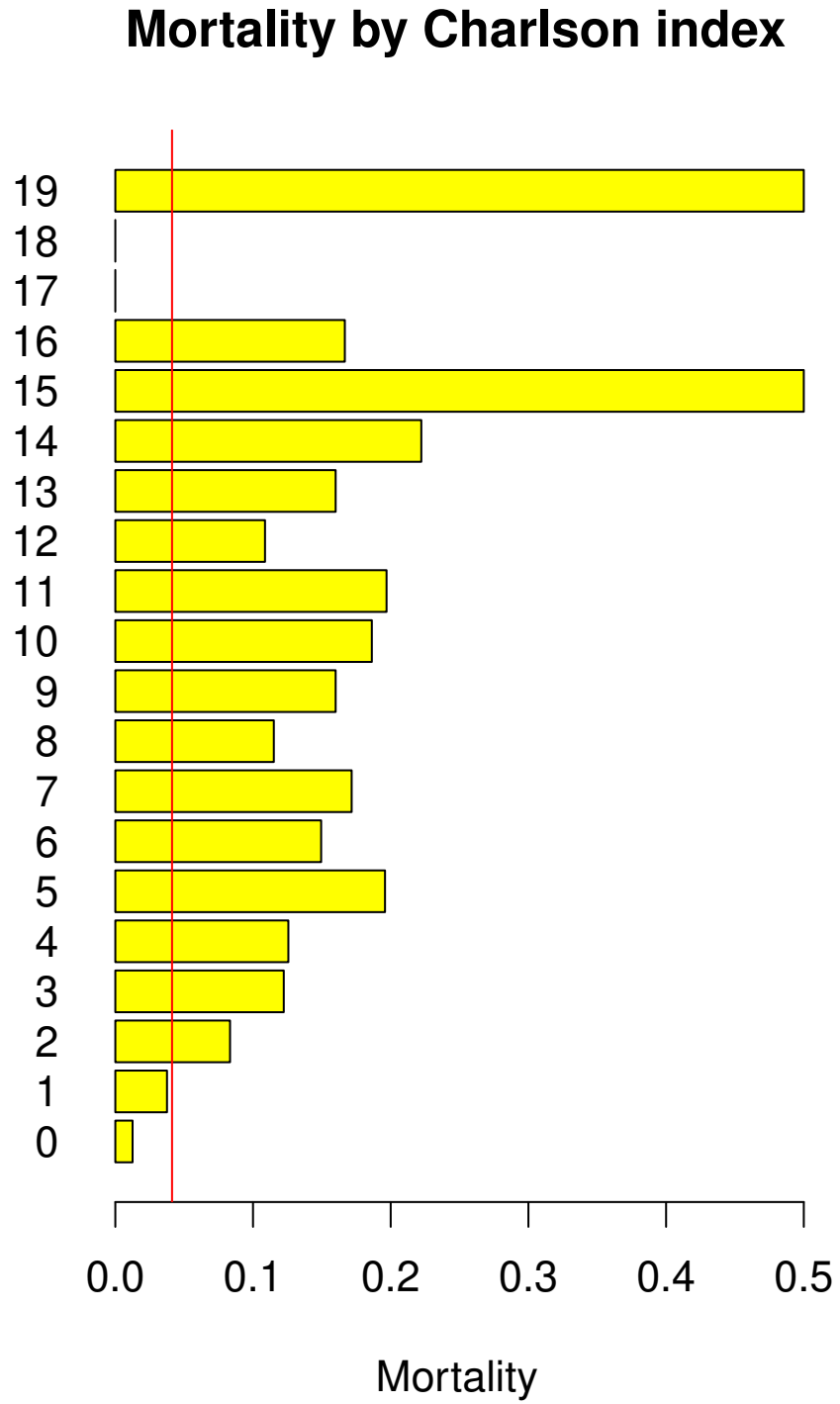

rheumatic disease
yes
no
0 10000 20000 30000
Count

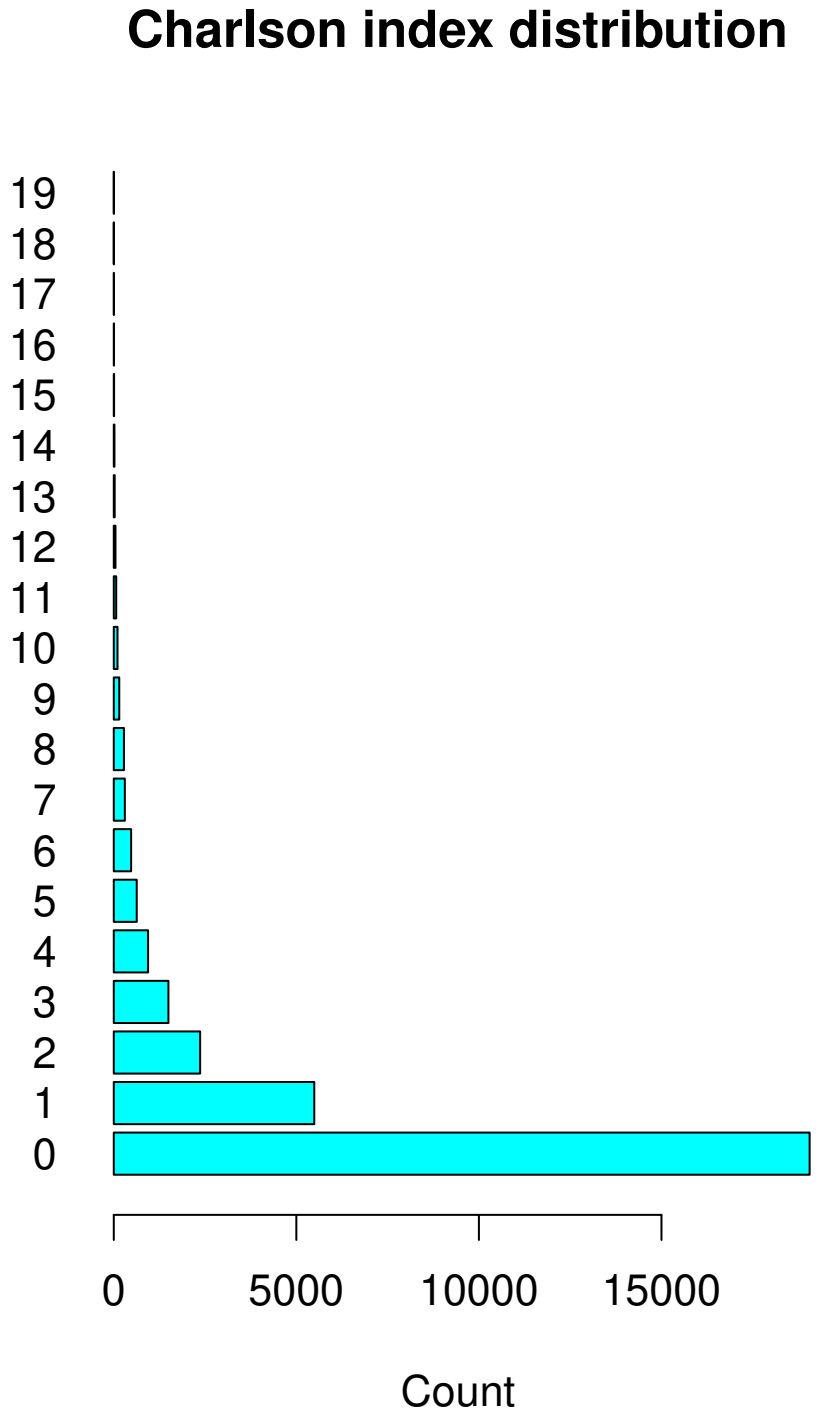

aids
yes
no
0.00 0.05 0.10 0.15 0.20
Mortality


aids
yes
no
0 10000 20000 30000
Count


chronic pulmonary disease
yes
no
0.00 0.05 0.10 0.15 0.20
Mortality


chronic pulmonary disease
yes
no
0 10000 20000 30000
Count


Mortality by Charlson index
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
0.0
0.1
0.2
0.3
0.4
0.5
Mortality


Charlson index distribution
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
0
5000
10000
15000
Count

Research questions

1. Which variables are most highly predictive of mortality?
2. Can a model be built to predict probability of death from COVID-19?
3. Can the model identify the groups at highest mortality risk?

Chi-squared tests (in decreasing significance)

age (X^2_4 = 2056, p-value $<$ 2.2E-16, X^2_1 = 1952)

died	18–50	50–59	60–69	70–79	80–90
0	15474	5710	4558	2616	1807
1	104	145	285	399	363

charlson (X^2_{20} = 1831, p-value $<$ 2.2e-16, X^2_1 = 1527)

died	0	1	2	3	4	5	6	7	8	9	10
0	18510	5250	2286	1365	851	547	411	291	244	146	82
1	224	208	187	187	115	126	72	62	34	23	21
died	11	12	13	14	15	16	17	18	19	20	
0	76	46	25	15	5	6	5	3	0	1	
1	12	9	6	2	4	2	1	0	1	0	

renal (X^2_1 = 1409, p-value $<$ 2.2e-16)

died	no	yes
0	27916	2249
1	810	486

Chi-squared tests (cont'd.)

CHF (X_1^2 = 1098, p-value $<$ 2.2e-16)

died	no	yes
0	28267	1898
1	897	399

MI (X_1^2 = 899, p-value $<$ 2.2e-16)

died	no	yes
0	29147	1018
1	1034	262

dementia (X_1^2 = 618, p-value $<$ 2.2e-16)

died	no	yes
0	29333	832
1	1097	199

PVD (X_1^2 = 479, p-value $<$ 2.2e-16)

died	no	yes
0	28800	1365
1	1060	236

Chi-squared tests (cont'd.)

cerebro (X_1^2 = 426, p-value $<$ 2.2e-16)

died	no	yes
0	28497	1668
1	1042	254

diabetes (X_1^2 = 321, p-value $<$ 2.2e-16)

died	no	yes
0	25875	4290
1	876	420

cancer (X_1^2 = 168, p-value $<$ 2.2e-16)

died	no	yes
0	28391	1774
1	1104	192

CPD (X_1^2 = 149, p-value $<$ 2.2e-16)

died	no	yes
0	25043	5122
1	905	391

Chi-squared tests (cont'd.)

race (X^2_5 = 181, p-value $<$ 2.2E-16, X^2_1 = 145)

died	AIAN	Asian	Black	NHPI	Unknown	White
0	96	775	8252	106	7340	13596
1	0	16	506	9	136	629

sex (X^2_1 = 99, p-value $<$ 2.2E-16)

died	F	M
0	16623	13542
1	532	764

metastatic (X^2_1 = 81, p-value $<$ 2.2e-16)

died	no	yes
0	29833	332
1	1245	51

liver (X^2_2 = 89, p-value $<$ 2.2e-16, X^2_1 = 80)

died	no	mild	severe
0	28770	1279	116
1	1172	102	22

Chi-squared tests (cont'd.)

hemipara (X_1^2 = 27, p-value = 1.755e-07)

died	no	yes
0	29783	382
1	1257	39

RD (X_1^2 = 25, p-value = 7.028e-07)

died	no	yes
0	29538	627
1	1242	54

PUD (X_1^2 = 21, p-value = 5.127e-06)

died	no	yes
0	29770	395
1	1259	37

aids (X_1^2 = 6, p-value = 0.016)

died	no	yes
0	29956	209
1	1279	17

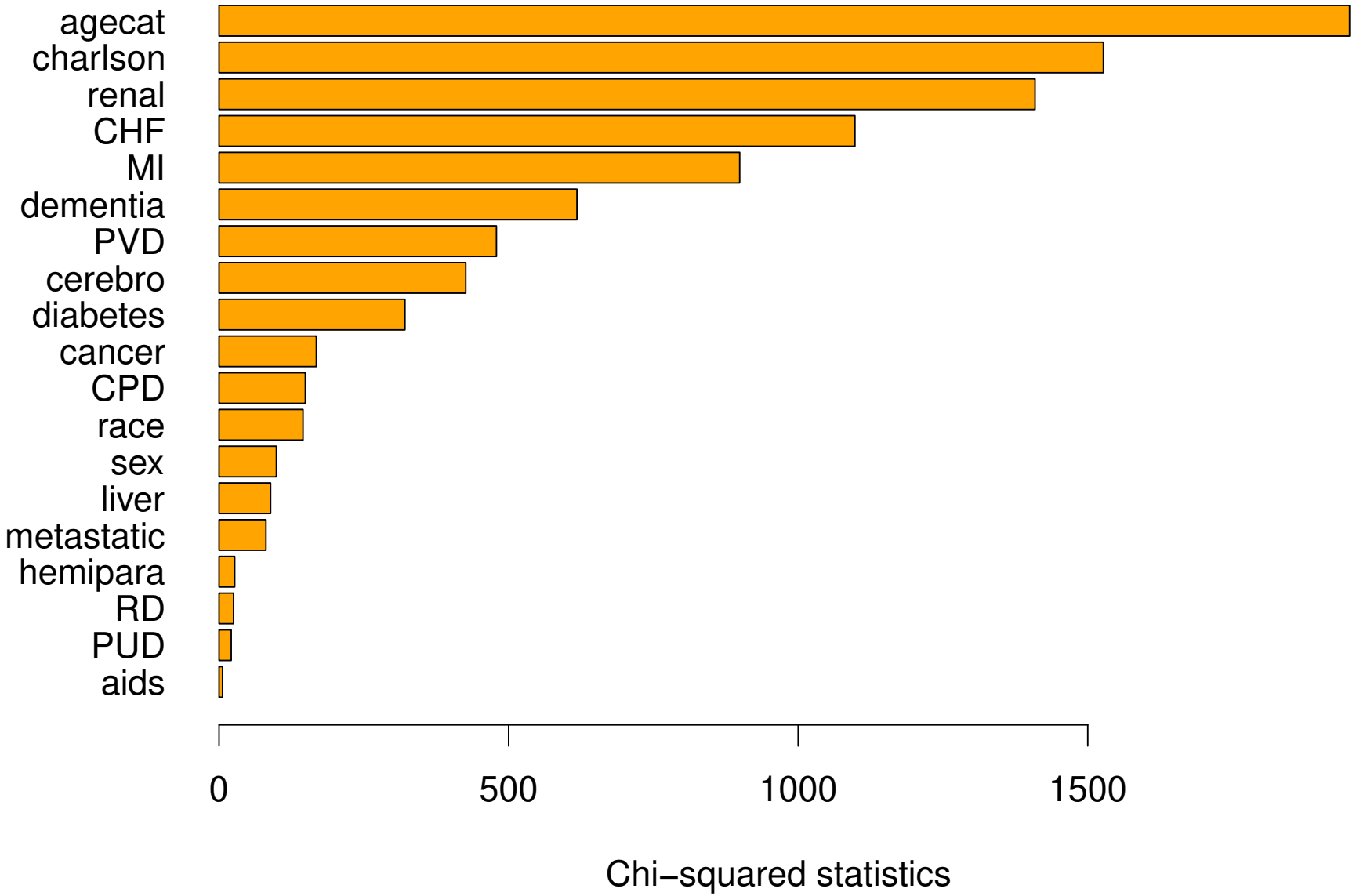
Modified Wilson-Hilferty (1931) approximation

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

Chi-squared statistics for COVID-19 data



Ordinary logistic regression

- Assume that

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

  where $p = P(Y = 1 \,|\, X_1 = x_1, X_2 = x_2, \ldots)$

- Solving for p gives

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

- Estimated coefficients $\beta_0, \beta_1, \ldots$ minimize the deviance

$$\texttt{dev} = -2 \sum_{i=1}^{n} \{y_i \log(\hat{p}_i) + (1 - y_i) \log(1 - \hat{p}_i)\}$$

Two problems with ordinary logistic regression

1. Charlson index is linearly dependent on comorbidities
2. Estimation difficulties even without Charlson index

Logistic regression model without charlson

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-17.29986	135.19306	-0.128	0.898177	
agecat1	1.09227	0.13101	8.337	< 2e-16	***
agecat2	1.82303	0.11929	15.282	< 2e-16	***
agecat3	2.57686	0.11837	21.769	< 2e-16	***
agecat4	2.84696	0.12397	22.965	< 2e-16	***
renal	0.76888	0.07486	10.271	< 2e-16	***
sexM	0.53602	0.06246	8.581	< 2e-16	***
MI	0.66626	0.09161	7.273	3.53e-13	***
CHF	0.37414	0.08352	4.479	7.48e-06	***
liver1	0.20791	0.11933	1.742	0.081454	.
liver2	1.17656	0.26282	4.477	7.58e-06	***
dementia	0.35173	0.09767	3.601	0.000317	***
metastatic	0.51248	0.18310	2.799	0.005127	**
CPD	0.19659	0.07239	2.716	0.006616	**

Logistic regression model w/o charlson (cont'd.)

	Estimate	Std. Error	z value	Pr(>|z|)	
aids	0.51484	0.27847	1.849	0.064491	.
hemipara	-0.32056	0.18858	-1.700	0.089156	.
PUD	-0.30429	0.19328	-1.574	0.115413	
cancer	-0.13542	0.09983	-1.356	0.174942	
PVD	-0.11120	0.09241	-1.203	0.228875	
diabetes	0.06361	0.07226	0.880	0.378715	
RD	0.14071	0.16091	0.875	0.381845	
cerebro	0.06029	0.08964	0.673	0.501253	
raceAsian	11.38412	135.19327	0.084	0.932892	
raceBlack.African American	12.27601	135.19303	0.091	0.927649	
raceNative Hawaiian/Other Pac.	13.20392	135.19352	0.098	0.922197	
raceUnknown	11.58188	135.19305	0.086	0.931729	
raceWhite	11.88757	135.19303	0.088	0.929932	

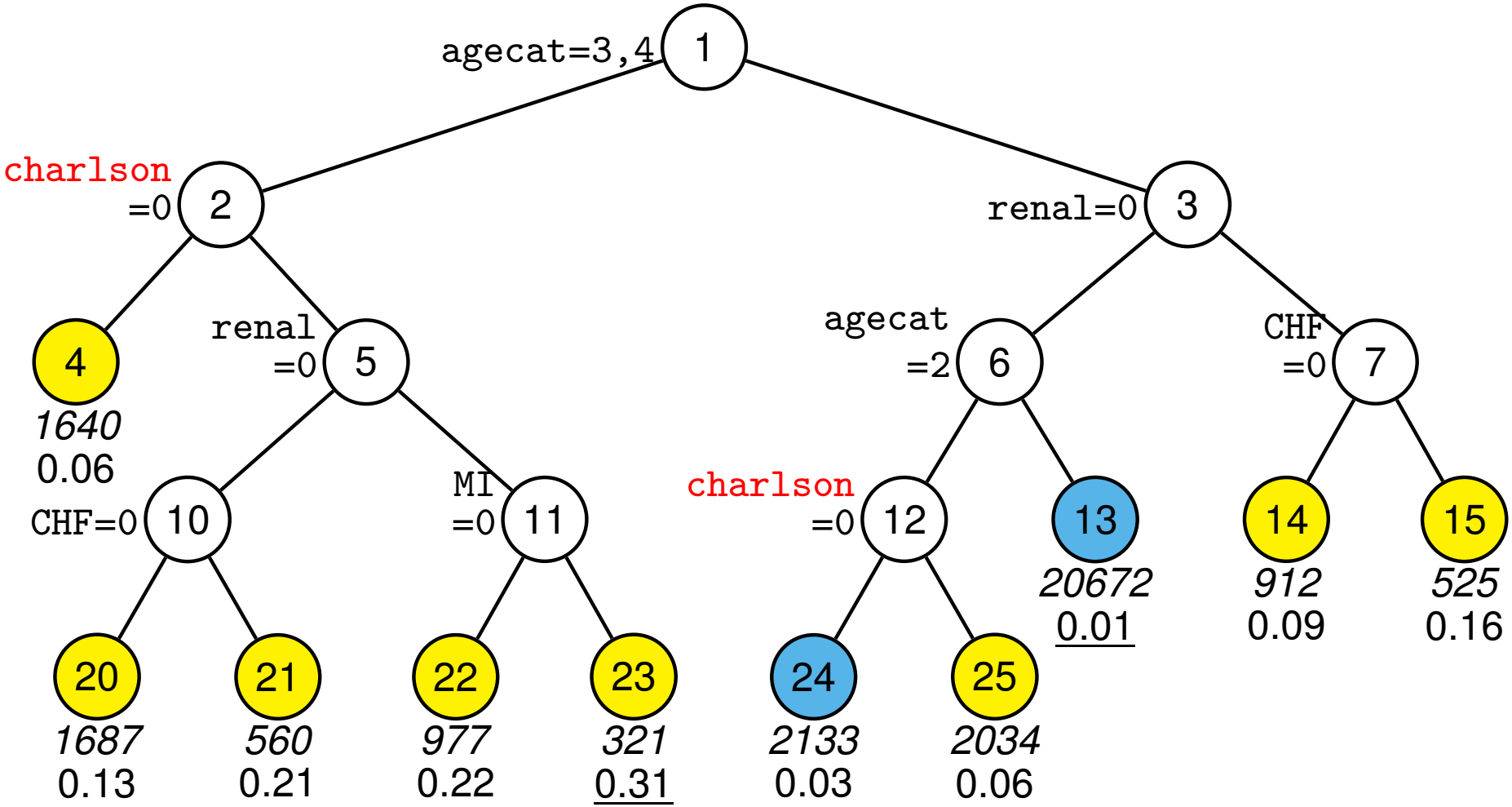
Model without American Indian and Alaska Native

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-5.91574	0.27957	-21.160	< 2e-16	***
agecat1	1.09227	0.13099	8.338	< 2e-16	***
agecat2	1.82303	0.11928	15.284	< 2e-16	***
agecat3	2.57686	0.11835	21.772	< 2e-16	***
agecat4	2.84696	0.12395	22.968	< 2e-16	***
renal	0.76888	0.07486	10.271	< 2e-16	***
sexM	0.53602	0.06246	8.582	< 2e-16	***
MI	0.66626	0.09161	7.273	3.53e-13	***
CHF	0.37414	0.08352	4.479	7.48e-06	***
liver1	0.20791	0.11933	1.742	0.081450	.
liver2	1.17656	0.26282	4.477	7.58e-06	***
raceBlack/AfricanAmerican	0.89189	0.26595	3.354	0.000798	***
raceNativeHawaiian/OtherPac.	1.81980	0.45112	4.034	5.49e-05	***
raceUnknown	0.19776	0.27596	0.717	0.473613	
raceWhite	0.50345	0.26480	1.901	0.057271	.

Model without Amer. Indian/Alaska Native (cont'd.)

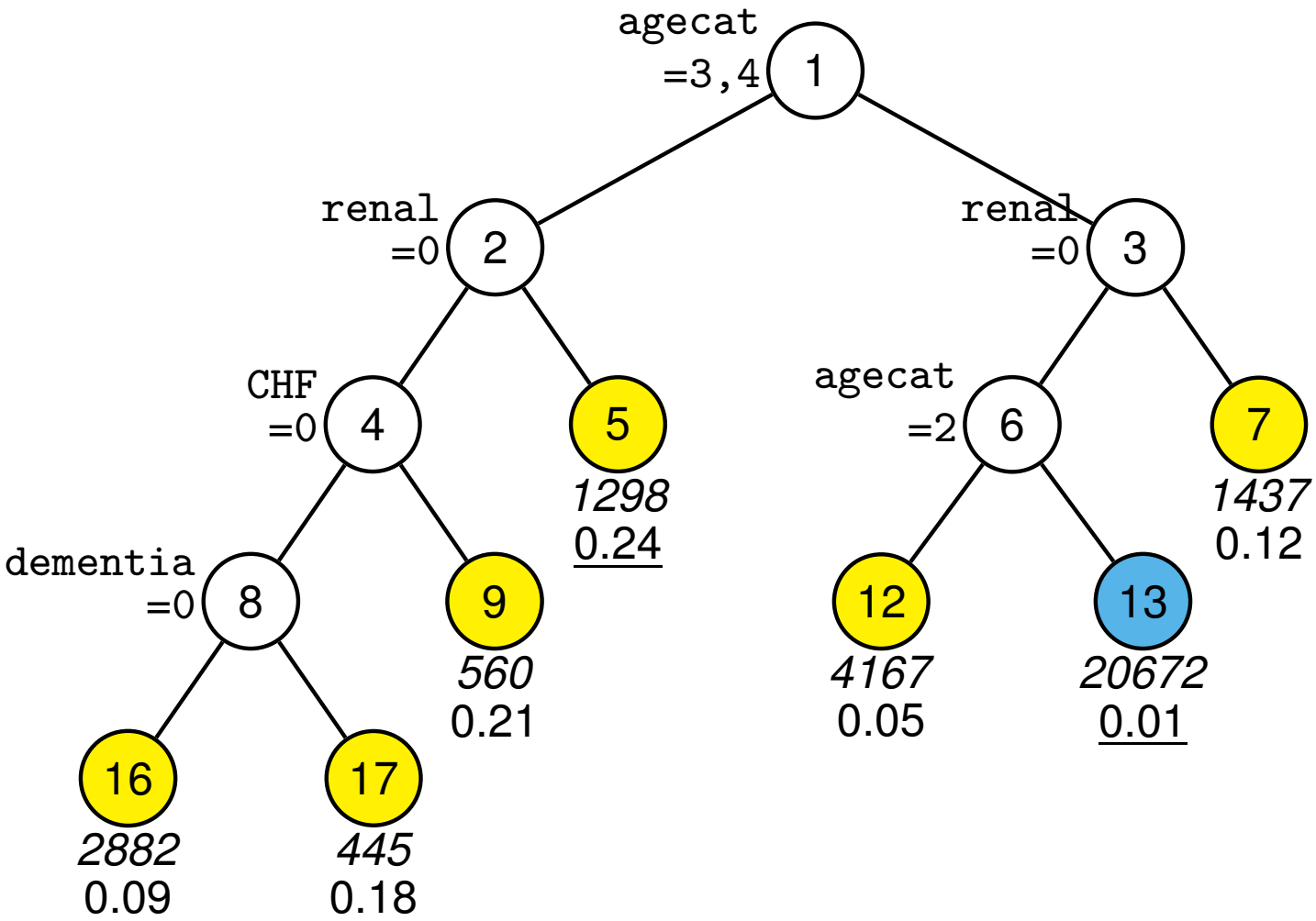
	Estimate	Std. Error	z value	Pr(>|z|)	
dementia	0.35173	0.09767	3.601	0.000317	***
metastatic	0.51248	0.18310	2.799	0.005126	**
CPD	0.19659	0.07239	2.716	0.006615	**
aids	0.51484	0.27847	1.849	0.064484	.
hemipara	-0.32056	0.18858	-1.700	0.089155	.
PUD	-0.30429	0.19328	-1.574	0.115411	
cancer	-0.13542	0.09983	-1.356	0.174940	
PVD	-0.11120	0.09241	-1.203	0.228874	
diabetes	0.06361	0.07226	0.880	0.378711	
RD	0.14071	0.16090	0.875	0.381841	
cerebro	0.06029	0.08964	0.673	0.501252	

GUIDE regression tree using all variables and obs



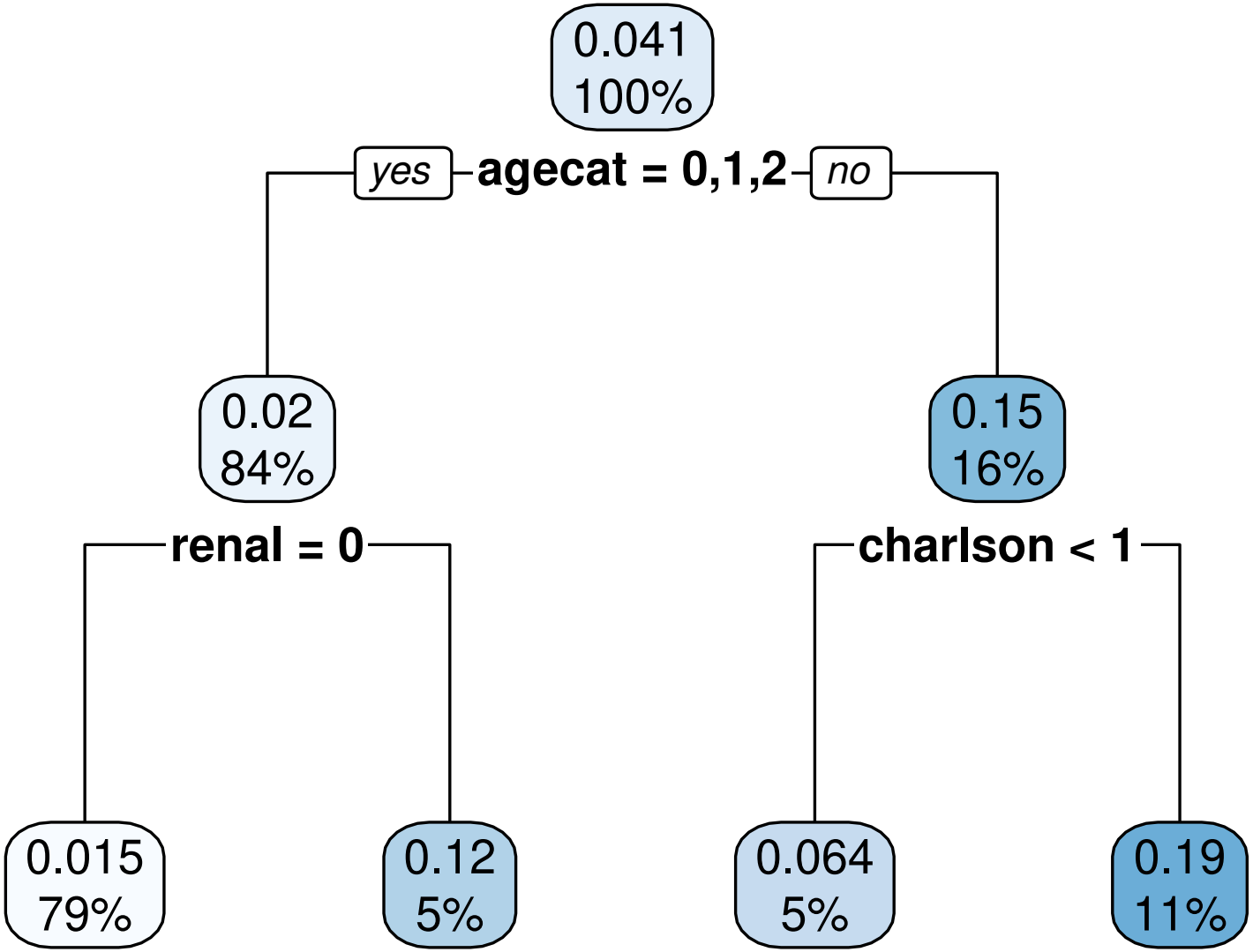
- Sample size (in *italics*) and mortality rate printed below nodes
- Terminal nodes with mortality rates above and below value of 0.04 at root node are colored yellow and skyblue, respectively

Regression tree without charlson

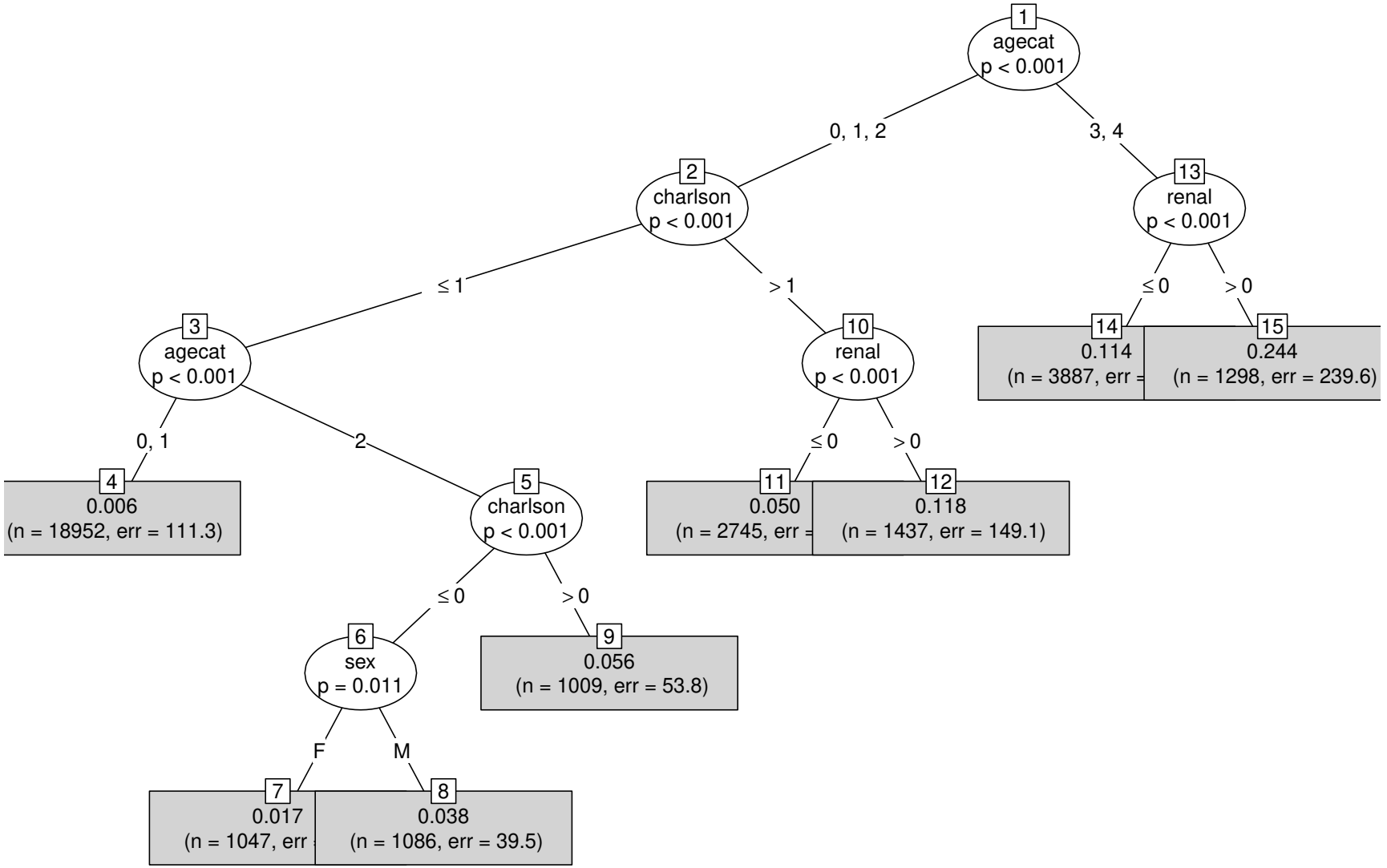


- Sample size (*in italics*) and mortality rate printed below nodes
- Terminal nodes with mortality rates above and below value of 0.04 at root node are colored yellow and skyblue, respectively

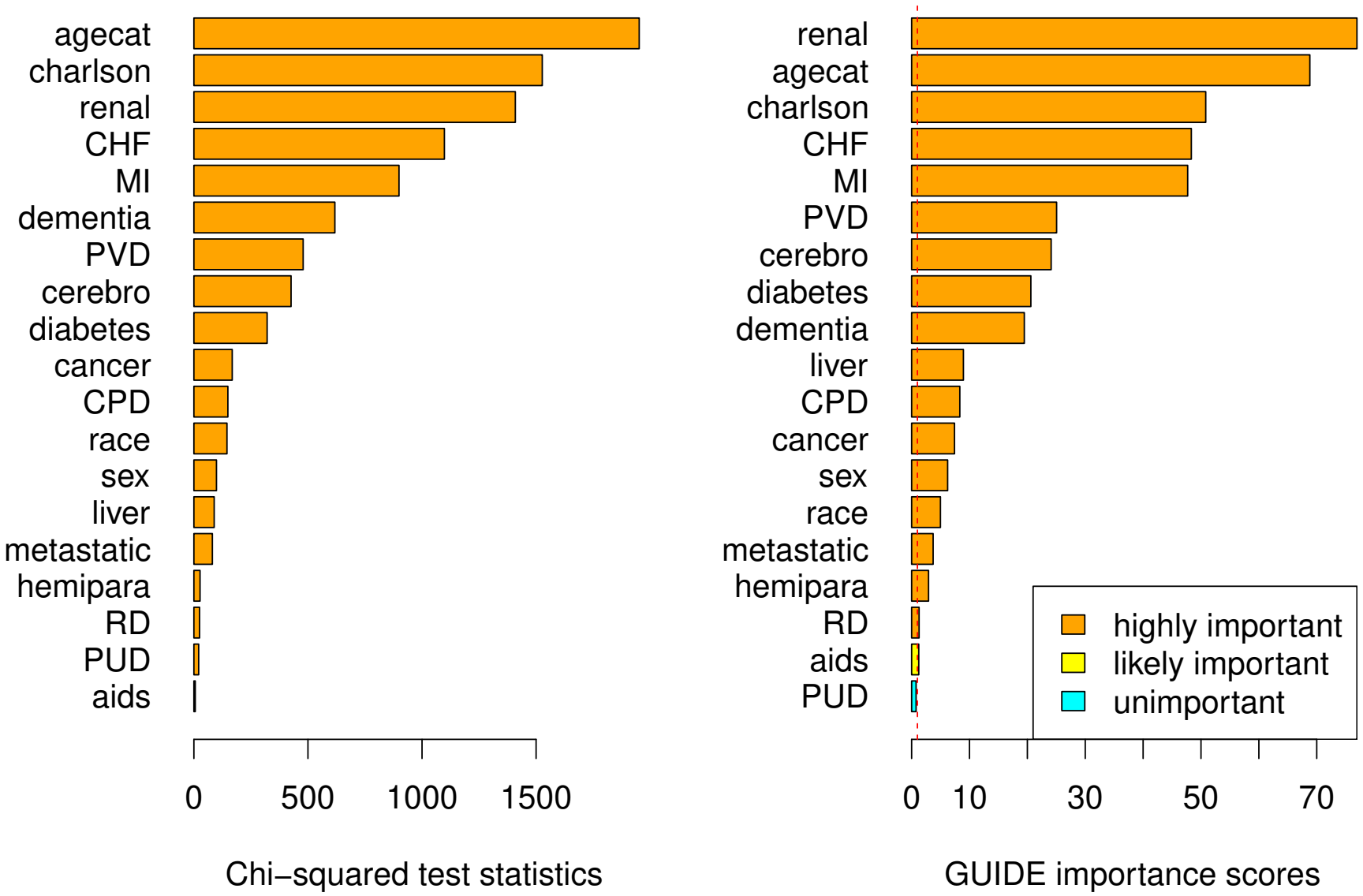
RPART (Therneau et al., 2019) tree



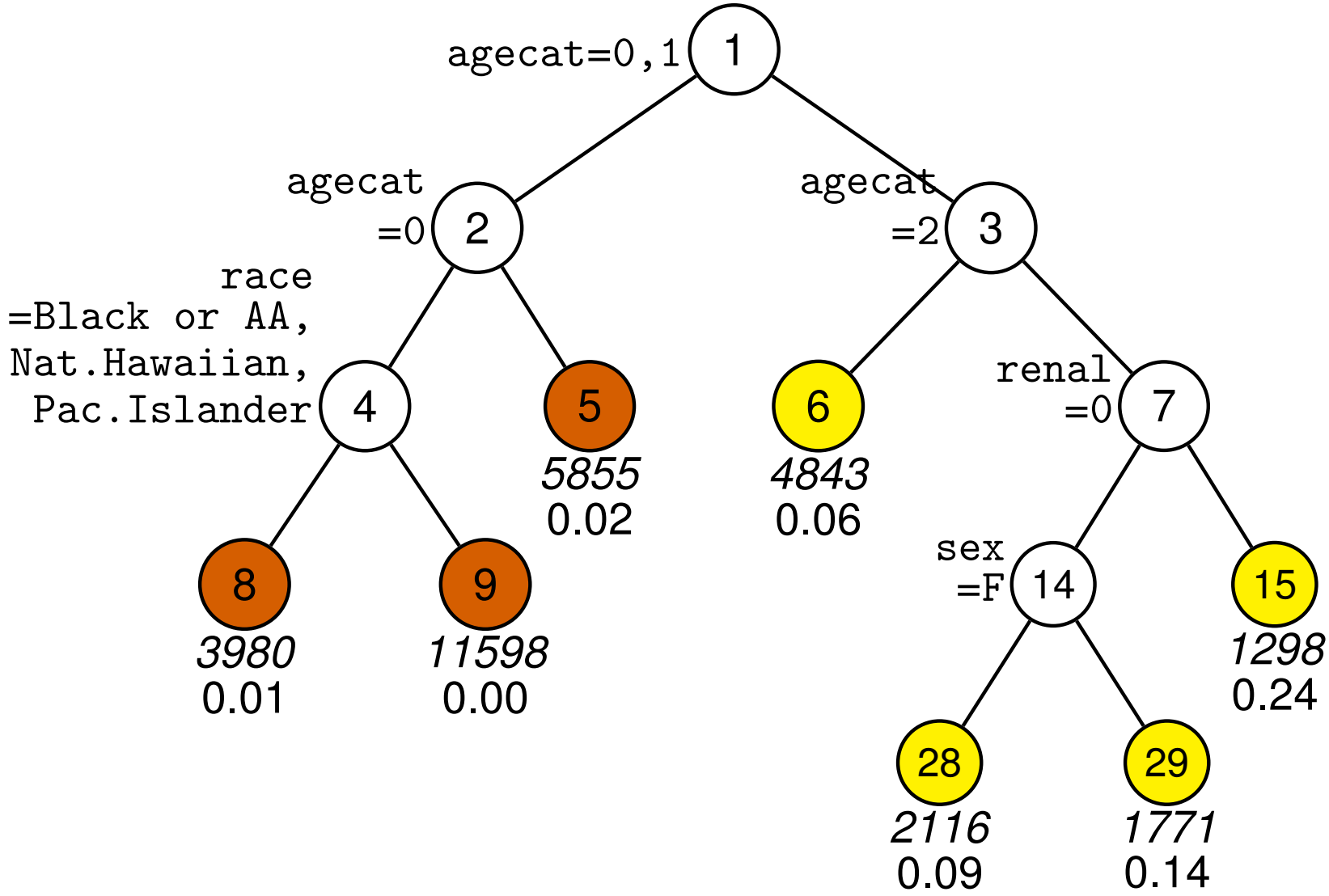
CTREE (Hothorn et al., 2006) tree



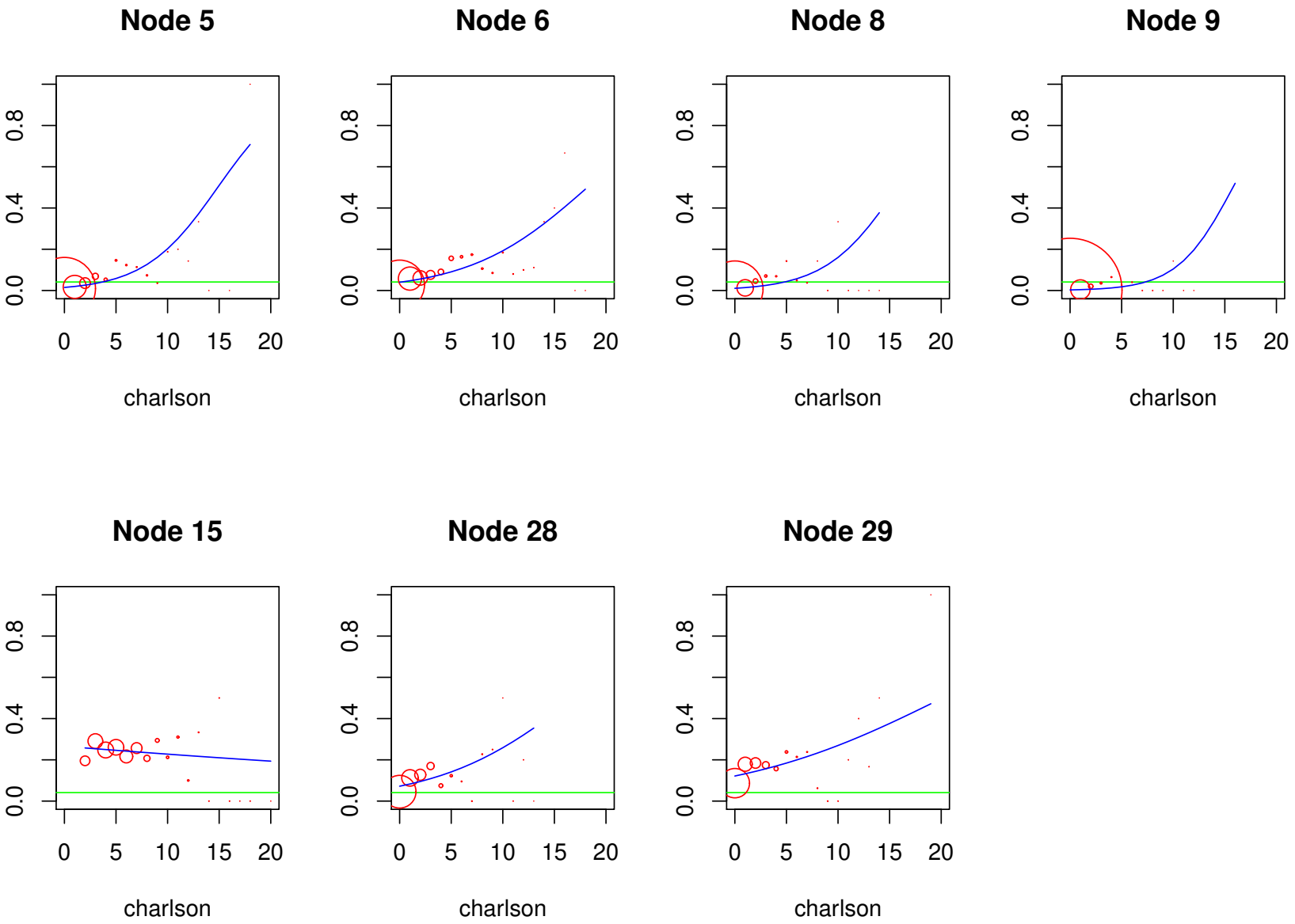
Chi-squared and GUIDE importance scores



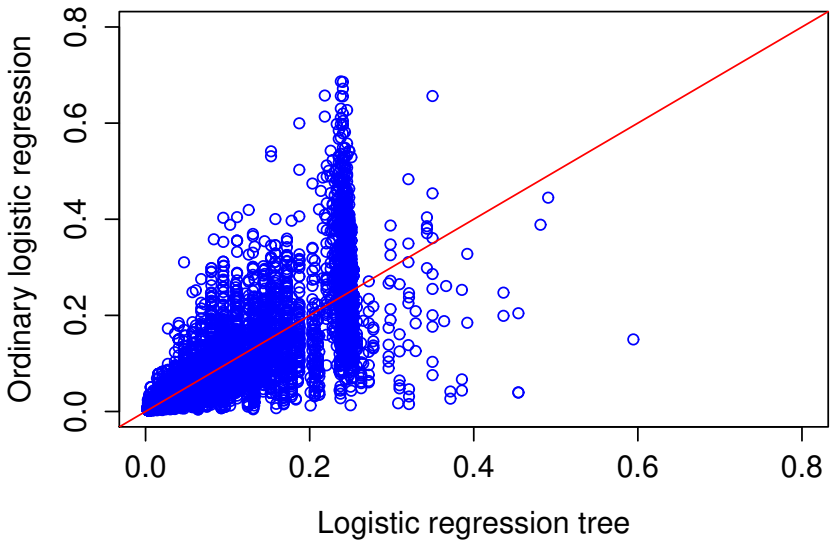
GUIDE logistic regression tree with charlson as sole linear predictor

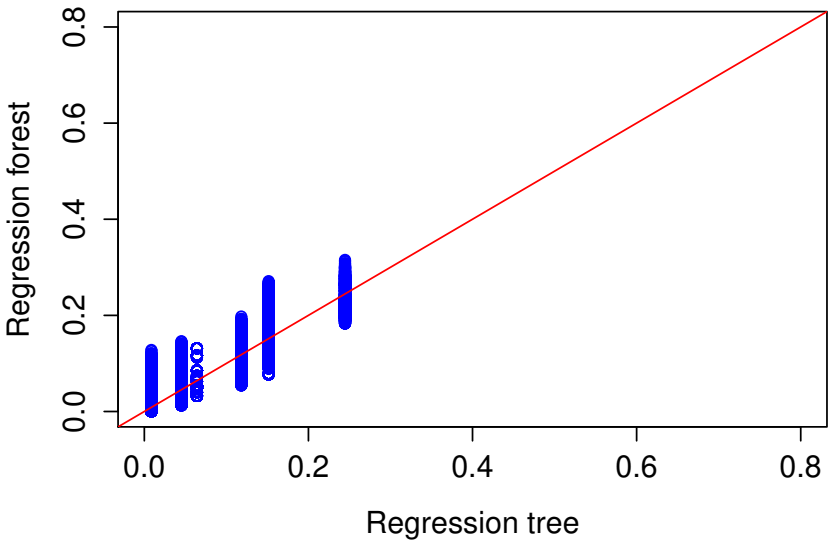


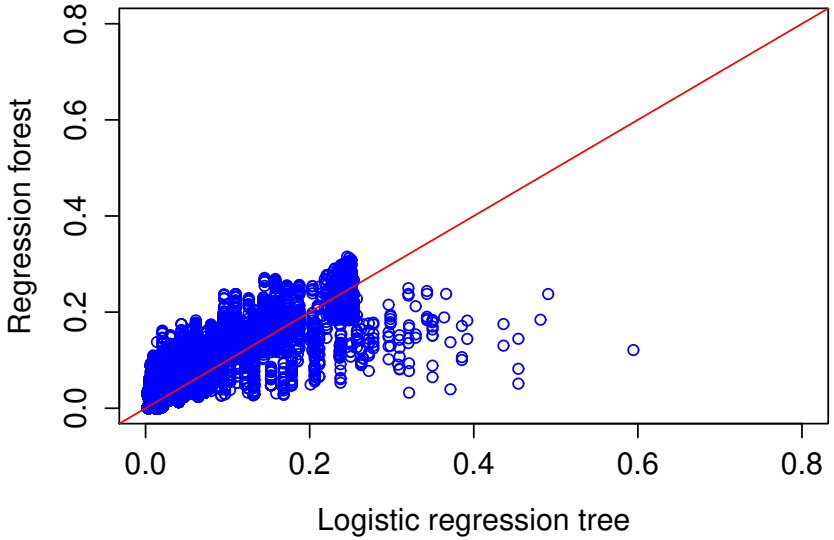
Logistic curves (area of circles $\propto$ sample size)

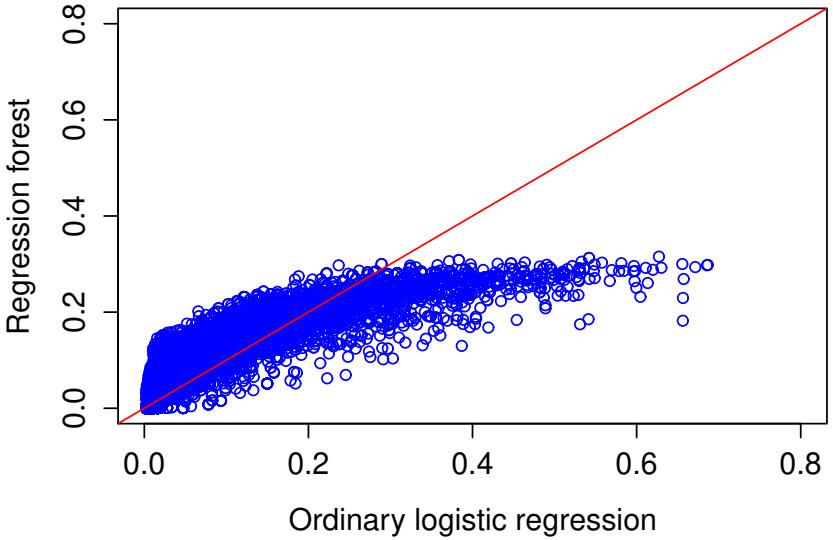


Fitted probabilities w/o Amer. Indian/Alaska Native









About GUIDE

- GUIDE algorithm and software have been in development for 35+ 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, 2019b,c, 2020); Loh and Zhou (2021)

Poisson regression: Unreplicated 3x2x4x10x3 soldering experiment (Comizzoli et al., 1990; Chambers and Hastie, 1992)

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

Solder: Amount of solder (thin, thick)

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

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

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

Response: Number of solder skips (0–48)

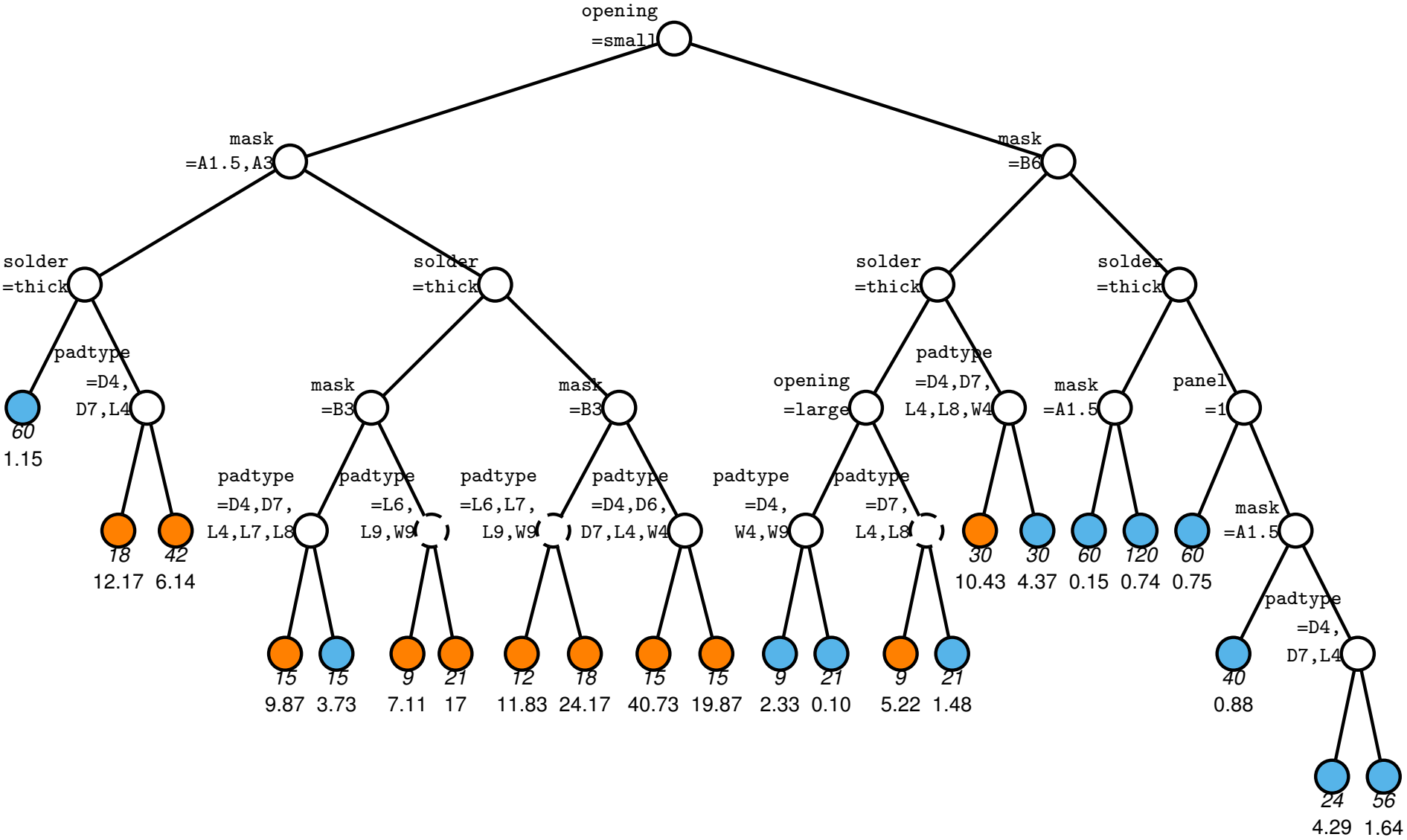
Full 2nd-degree Poisson loglinear model

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

Chambers-Hastie model with 2-factor interactions among 3 predictors with largest main effects

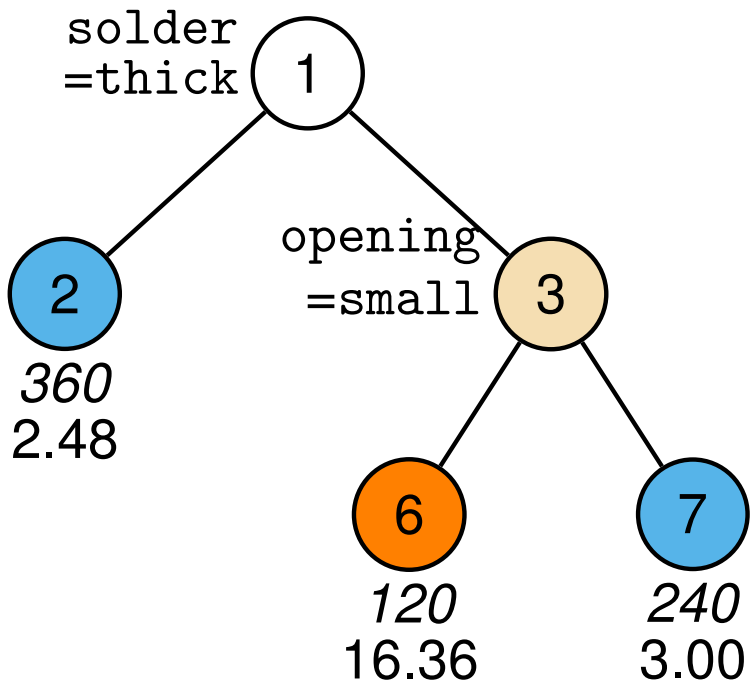
Regressor	Coef	t-stat	Regressor	Coef	t-stat
Constant	-2.668	-9.25			
maskA3	0.396	1.21	openmedium	0.921	2.95
maskB3	2.101	7.54	opensmall	2.919	11.63
maskB6	3.010	11.36	soldthin	2.495	11.44
padD6	-0.369	-5.17	maskA3:openmedium	0.816	2.44
padD7	-0.098	-1.49	maskB3:openmedium	-0.447	-1.44
padL4	0.262	4.32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

GUIDE piecewise-constant Poisson model



Sample size (in *italics*) and mean skip printed below each terminal node

GUIDE piecewise main effects Poisson model



Sample size (in *italics*) and mean skip below each terminal node

Nodes with means above and below value of 4.97 at root node are colored orange and skyblue respectively

Node 3 (in wheat color) has interaction between opening and mask

Regression coefficients for GUIDE model

	solder = thick		solder = thin			
			opening = small		medium or large	
Regressor	Coef	t-stat	Coef	t-stat	Coef	t-stat
Constant	-2.43	-10.68	2.08	21.5	-0.37	-1.9
maskA3	0.47	2.37	0.31	3.3	0.81	4.5
maskB3	1.83	11.01	1.05	12.8	1.01	5.8
maskB6	2.52	15.71	1.50	19.3	2.27	14.6
openmedium	0.86	5.57	aliased		0.10	1.4
opensmall	2.46	18.18	aliased		aliased	
panel2	0.22	2.72	0.31	5.5	0.58	5.7
panel3	0.07	0.81	0.19	3.2	0.69	6.9

Regression coefficients (cont'd.)

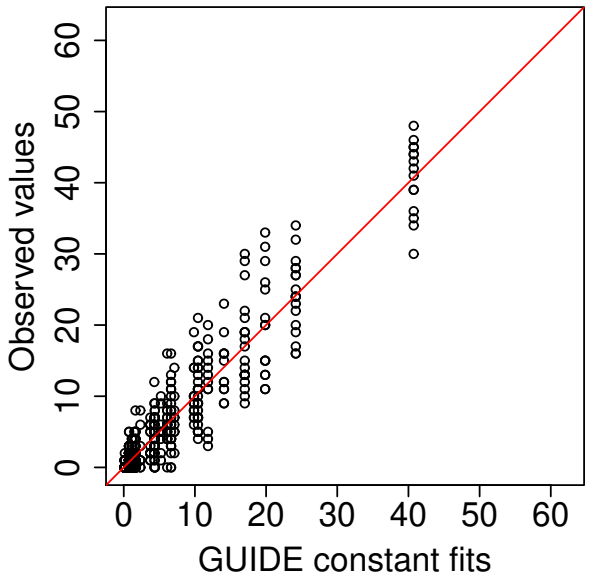
	solder = thick		solder = thin			
			opening = small		medium or large	
Regressor	Coef	t-stat	Coef	t-stat	Coef	t-stat
padD6	-0.32	-2.03	-0.25	-2.8	-0.80	-4.6
padD7	0.12	0.85	-0.15	-1.7	-0.19	-1.3
padL4	0.70	5.53	0.08	1.0	0.21	1.6
padL6	-0.40	-2.46	-0.72	-6.8	-0.82	-4.7
padL7	0.04	0.29	-0.65	-6.3	-0.76	-4.5
padL8	0.15	1.05	-0.43	-4.5	-0.36	-2.4
padL9	-0.59	-3.43	-0.64	-6.3	-0.67	-4.1
padW4	-0.05	-0.37	-0.09	-1.0	-0.23	-1.6
padW9	-1.32	-5.89	-1.38	-10.3	-1.75	-7.0

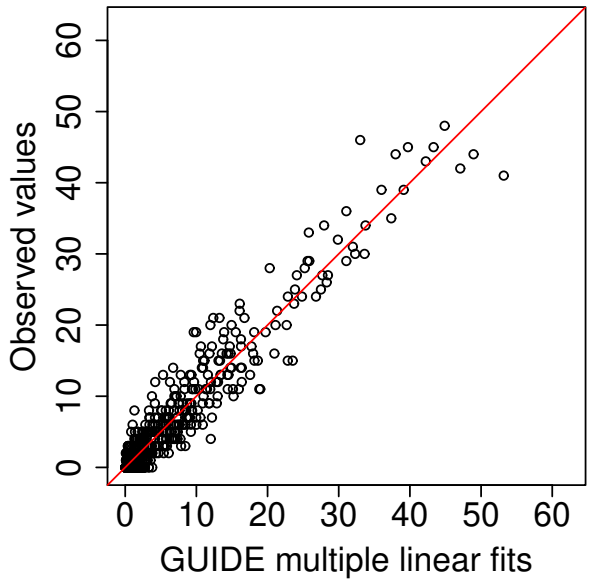
GUIDE model in equation form

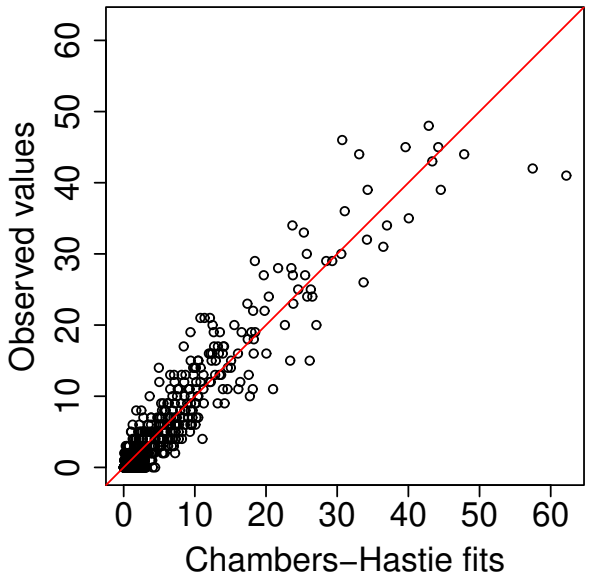
$$
\begin{aligned}
\log(EY) \;=\;& I(\texttt{solder} = \texttt{thick}) \left(\beta_{20} + \sum \beta_{2i} x_i\right) \\
&+ I(\texttt{solder} = \texttt{thin}, \texttt{opening} = \texttt{small}) \left(\beta_{60} + \sum \beta_{6i} x_i\right) \\
&+ I(\texttt{solder} = \texttt{thin}, \texttt{opening} = \texttt{large}, \texttt{medium}) \left(\beta_{70} + \sum \beta_{7i} x_i\right) \\
=\;& (1 - \texttt{solderthin}) \left(-2.43 + 0.47\,\texttt{maskA3} + 1.83\,\texttt{maskB3} + \ldots - 1.32\,\texttt{padW9}\right) \\
&+ \texttt{solderthin} \times \texttt{openingsmall} \left(2.08 + 0.31\,\texttt{maskA3}\right. \\
&\left. + \; 1.05\,\texttt{maskB3} + \ldots - 1.38\,\texttt{padW9}\right) \\
&+ \texttt{solderthin} \times (1 - \texttt{openingsmall}) \left(-0.37 + 0.81\,\texttt{maskA3}\right. \\
&\left. + \; 1.01\,\texttt{maskB3} + \ldots - 1.75\,\texttt{padW9}\right)
\end{aligned}
$$

- Model has some three-factor interactions

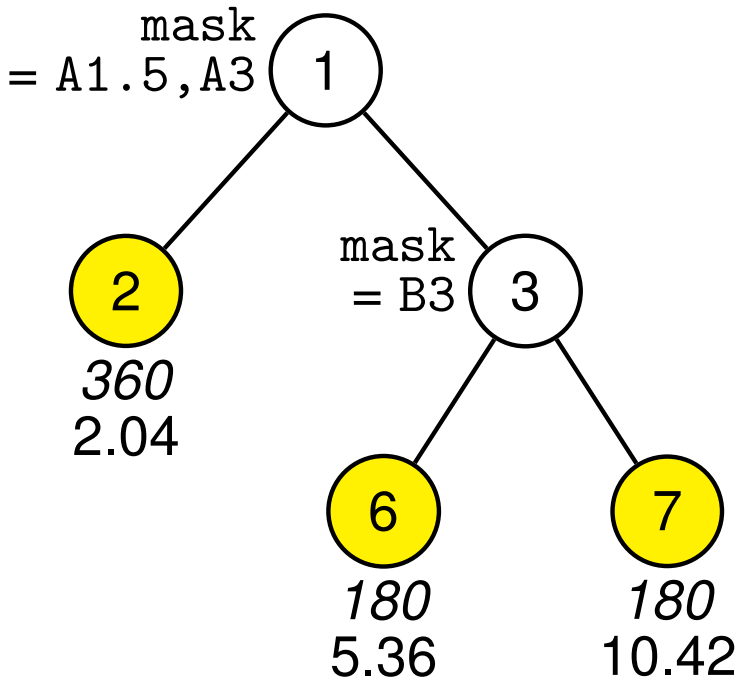
Observed vs. fitted values







MOB (Hothorn et al., 2006)
piecewise main effects Poisson model



Sample size (in *italics*) and mean number of solder skips each leaf node

Regression coefficients for MOB model

	mask = A1.5, A3		B3		B6	
Regressor	Coef	t-stat	Coef	t-stat	Coef	t-stat
Constant	-2.15	-10.8	-0.02	-0.1	0.93	9.2
openingmedium	0.74	4.8	-0.17	-1.2	0.25	2.9
openingsmall	2.19	16.2	2.06	19.4	1.72	24.4
solderthin	1.74	16.8	0.99	13.6	0.95	18.5
maskA3	0.43	5.7				
padtypeD6	-0.52	-3.1	-0.32	-2.3	-0.34	-3.5
padtypeD7	-0.02	-0.2	-0.15	-1.2	-0.10	-1.1
padtypeL4	0.45	3.4	0.28	2.4	0.17	2.0
padtypeL6	-0.52	-3.1	-0.58	-3.9	-0.78	-7.0
padtypeL7	-0.59	-3.5	-0.28	-2.1	-0.58	-5.5
padtypeL8	-0.27	-1.8	-0.12	-0.9	-0.36	-3.7
padtypeL9	-0.36	-2.3	-0.69	-4.5	-0.73	-6.7
padtypeW4	-0.30	-1.9	-0.27	-2.0	0.02	0.2
padtypeW9	-1.73	-6.6	-1.91	-7.8	-1.19	-9.2
panel2	0.35	3.6	0.30	3.7	0.35	6.1
panel3	0.46	4.9	0.28	3.4	0.16	2.6

Leave-one-out cross-validation (CV) estimates of mean deviance for solder data

Method	Mean	Time[a]
GUIDE multiple linear	1.43	3.65
MOB multiple linear	1.65	0.08
GLM	1.67	0.01
GUIDE constant	1.89	0.64
MOB constant	1.99	6.05
RPART	2.44	0.01

Mean deviance = $2n^{-1}\sum_i\{y_i\log(y_i/\hat{\mu}_i)-(y_i-\hat{\mu}_i)\}$, n = 720

[a]average time to fit one data set in seconds

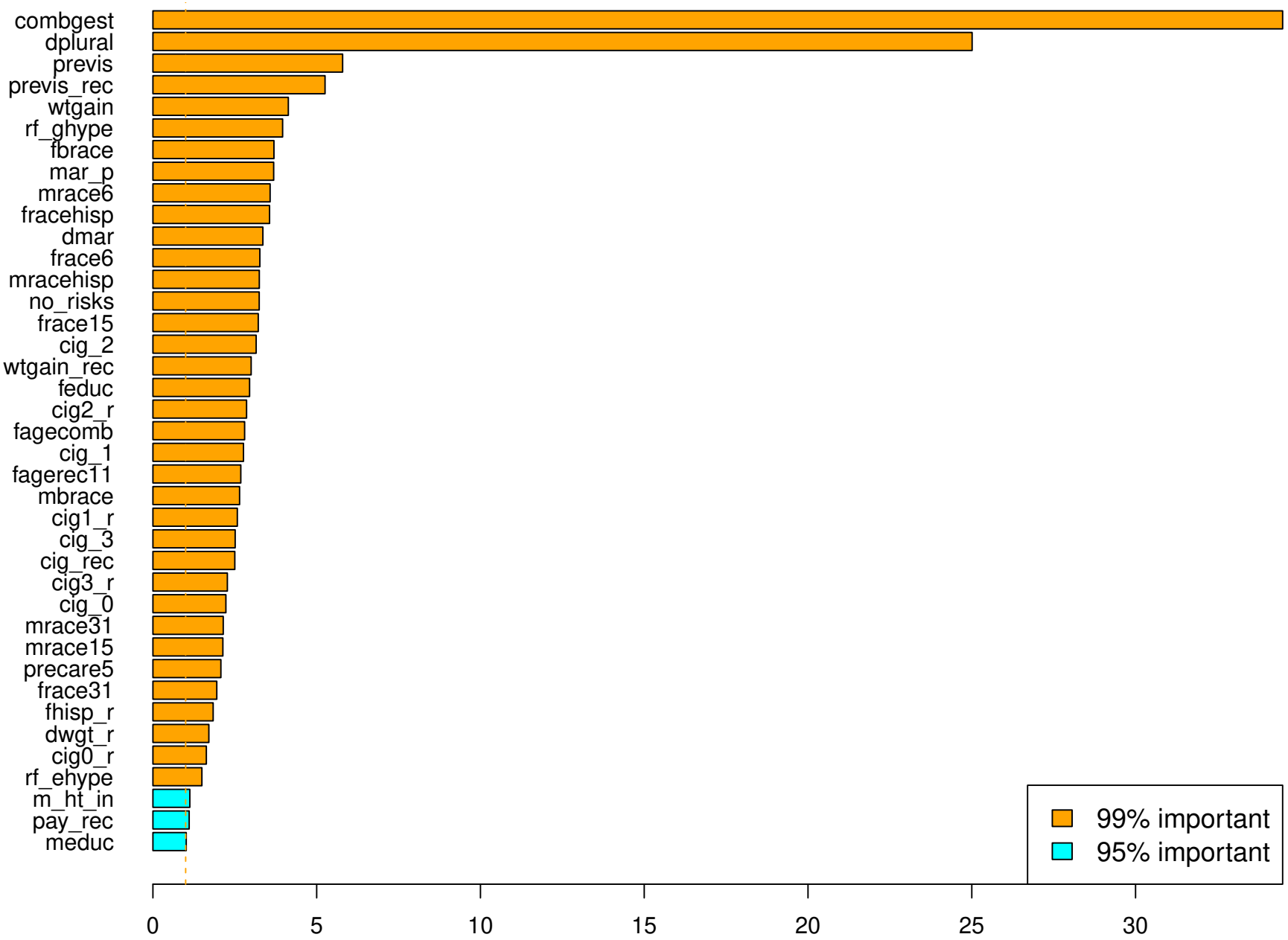
Missing values: birth weight data

- Data from 2016 CDC Natality Public Use File
- Birth weight and more than 200 predictor variables for 3,956,112 births in U.S. in 2016
- 8.15% have low birth weight (defined as less than 2500 gm $\approx$ 5.5 lbs)
- 99.6% of subjects have missing values
- Question: what factors and how are they predictive of low birth weight?

Approach using logistic regression

- Logistic regression is inapplicable to missing data
- Options:
  1. Use only observations with complete data—0.4% of data
  2. Delete observations (row deletion) and/or variables (column deletion) with missing values
  3. Impute missing values (more than 2.4 million)
     - imputation is a much harder task than logistic regression
  4. Reduce number of variables with GUIDE:
     (a) Use GUIDE to find the important variables
     (b) Apply logistic regression to set of complete observations (3,169,938) in the selected variables

GUIDE importance scores for predicting lowbwt



39 variables found by GUIDE [# missing values]

combgest	Combined gestation in weeks [**3,516**]
dplural	Plurality: 1=single, 2=twin, 3=triplet, 4=quadruplet, 5=quintuplet or higher [**0**]
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**]
dwgt_r	Mother's delivery weight in pounds [**70,304**]
wtgain	Weight gain in lbs [**143,049**]
wtgain_rec	Weight gain recode: 1= $<$11, 2=11–20, 3=21–30, 4=31–40, 5=41–98 lbs [**143,049**]
m_ht_in	Mother's height in inches [**28,356**]

dmar	Marital status: 1=married, 2=unmarried [**0**]
mar_p	Paternity acknowledged: Y=yes, N=no, U=unknown, X=not app. [**0**]
fagecomb	Father's combined age in years [**469,589**]
fagerec11	Father's age recode: 1=under 15, 2=15–19, 3=20–24, 4=25–29, 5=30–34, 6=35–39, 7=40–44, 8=45–49, 9=50–54, 10=55-98 [**469,589**]
feduc/ meduc	Father's/mother's education: 1=8th grade or less, 2=9–12th grade with no diploma, . . . , 8=doctorate or professional degree [**555,897**/**51,721**]
pay_rec	Pay recode: 1=medicaid, 2=private insurance, 3=self pay, 4=other [**0**]
fbrace/ mbrace	Father's/mother's bridged race (individuals reporting more than one race bridged into one race): 1=White, 2=Black, 3=American Indian or Alaskan Native (AIAN), 4=Asian or Pacific Islander [**0**]
frace6/ mrace6	Father's/mother's race recode to 6 values: 1=White, 2=Black, 3=AIAN, 4=Asian, 5=Native Hawaiian or other Pacific Islander (NHOPI), 6=more than 1 race, 9=unknown or not stated [**0**]

frace15/ mrace15	Father's/mother's race recode to 15 values: 1=White, 2=Black, 3=AIAN, 4=Asian Indian, 5=Chinese, 6=Filipino, 7=Japanese, 8=Korean, 9=Vietnamese, 10=Other Asian, 11=Hawaiian, 12=Guamanian, 13=Samoan, 14=Other Pacific Islander, 15=More than one race, 99=Unknown [**0**]
frace31/ mrace31	Father's/mother's race recode to 31 values [**0**]
fhisp_r	Father's Hispanic origin recode: 0=non-Hispanic, 1=Mexican, 2=Puerto Rican, 3=Cuban, 4=Central and South American, 5=Other and unknown Hispanic origin, 9=not stated [**0**]
fracehisp/ mracehisp	Father's/mother's race/Hispanic origin: 1=non-Hispanic White, 2=non-Hispanic Black, 3=non-Hispanic AIAN, 4=non-Hispanic Asian, 5=non-Hispanic NHOPI, 6=non-Hispanic more than 1 race, 7=Hispanic, 8=origin unknown or not stated, 9=non-Hispanic race, unknown or not stated [**0**]
no_risks	No risk factors reported: 1=true, 0=false, 9=not reported [**0**]

rf_ehype	Hypertension eclampsia (Y=yes, N=no, U=unknown or not stated) [**0**]
rf_ghype	Risk factor for gestational hypertension: Y=yes, N=no, U=unknown or not stated [**0**]
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**]
cig0_r	cigarettes before pregnancy recode: 0=nonsmoker, 1=1–5, 2=6–10, 3=11–20, 4=21–40, 5=41 or more [**19,350**]
cig1_r	cigarettes 1st trimester recode: same codes as cig0_r [**19,719**]
cig2_r	cigarettes 2nd trimester recode: same codes as cig0_r [**19,985**]
cig3_r	cigarettes 3rd trimester recode: same codes as cig0_r [**20,035**]
cig_rec	cigarette recode: Y=yes, N=no, U=unknown or not stated [**0**]

Logistic regression on 3,169,938 complete cases

```
Coefficients: (28 not defined because of singularities)
Estimate Std. Error  z value Pr(>|z|)
(Intercept)  2.279e+01  1.027e-01  221.982  < 2e-16 ***
combgest    -5.678e-01  1.235e-03 -459.731  < 2e-16 ***
dplural      2.458e+00  8.835e-03  278.205  < 2e-16 ***
meduc        2.098e-02  2.424e-03    8.655  < 2e-16 ***
feduc       -2.577e-02  2.394e-03  -10.764  < 2e-16 ***
precare5    -3.168e-01  5.531e-03  -57.278  < 2e-16 ***
previs       5.254e-02  1.945e-03   27.012  < 2e-16 ***
previs_rec  -2.351e-01  4.587e-03  -51.249  < 2e-16 ***
cig_0        8.230e-04  1.980e-03    0.416 0.677644
cig_1        2.651e-03  3.668e-03    0.723 0.469899
cig_2       -1.333e-02  4.934e-03   -2.702 0.006892 **
cig_3        2.844e-03  4.419e-03    0.644 0.519843
cig0_r       8.543e-02  1.622e-02    5.266 1.39e-07 ***
cig1_r      -2.057e-02  3.057e-02   -0.673 0.500915
cig2_r       1.871e-01  3.493e-02    5.355 8.53e-08 ***
cig3_r       1.089e-02  3.189e-02    0.342 0.732691
cig_recY     3.036e-01  2.915e-02   10.415  < 2e-16 ***
```

```
rf_ghypeU     7.519e-02  1.217e-01    0.618 0.536820
rf_ghypeY     9.571e-01  9.675e-03   98.923  < 2e-16 ***
rf_ehypeU            NA         NA       NA       NA
rf_ehypeY     1.192e+00  3.352e-02   35.558  < 2e-16 ***
no_risks1    -4.994e-02  6.675e-03   -7.482 7.34e-14 ***
no_risks9            NA         NA       NA       NA
pay_rec2     -9.855e-03  7.390e-03   -1.334 0.182364
pay_rec3     -1.717e-01  1.572e-02  -10.918  < 2e-16 ***
pay_rec4     -4.212e-02  1.499e-02   -2.810 0.004950 **
pay_rec9     -3.257e-02  4.318e-02   -0.754 0.450693
fagecomb     -1.246e-04  1.950e-03   -0.064 0.949051
fagerec11    -5.120e-03  9.587e-03   -0.534 0.593294
dmar2                NA         NA       NA       NA
```

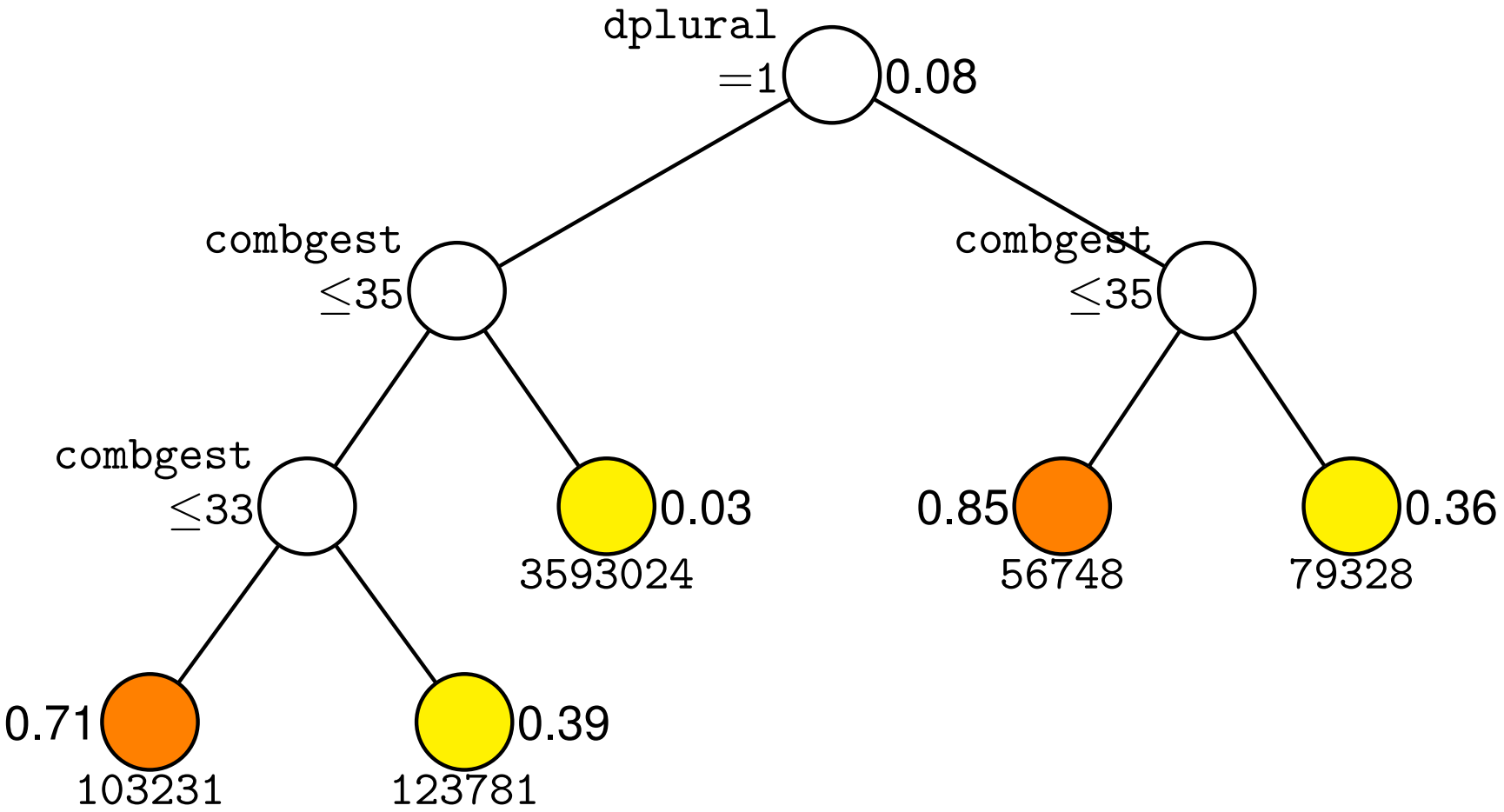
```
frace62             NA          NA       NA       NA
frace63             NA          NA       NA       NA
frace64             NA          NA       NA       NA
frace65             NA          NA       NA       NA
frace66             NA          NA       NA       NA
frace69             NA          NA       NA       NA
frace1510    1.175e-01  6.344e-02    1.853 0.063923 .
frace1511    3.142e-01  1.862e-01    1.687 0.091578 .
frace1512    2.537e-01  1.906e-01    1.331 0.183311
frace1513    2.419e-02  1.718e-01    0.141 0.888027
frace1514           NA          NA       NA       NA
frace1515           NA          NA       NA       NA
frace152            NA          NA       NA       NA
frace153            NA          NA       NA       NA
frace154     1.626e-01  6.659e-02    2.441 0.014633 *
frace155     4.841e-02  6.478e-02    0.747 0.454883
frace156     2.675e-01  6.498e-02    4.116 3.85e-05 ***
frace157     1.033e-01  9.800e-02    1.054 0.291793
frace158    -7.140e-02  8.191e-02   -0.872 0.383374
frace159            NA          NA       NA       NA
frace1599           NA          NA       NA       NA
```

```
frace312     3.932e-02  6.722e-02    0.585 0.558599
frace313    -4.883e-02  1.148e-01   -0.425 0.670592
frace314     1.165e-01  1.220e-01    0.955 0.339572
frace315    -3.482e-01  2.066e-01   -1.685 0.091913 .
frace316    -4.010e-02  6.205e-02   -0.646 0.518064
frace317     2.400e-01  1.042e-01    2.303 0.021266 *
frace318     3.484e-01  1.153e-01    3.022 0.002508 **
frace319     5.784e-02  2.005e-01    0.288 0.773027
frace3110   -8.834e-02  6.694e-02   -1.320 0.186950
frace3111   -6.669e-01  4.080e-01   -1.635 0.102076
frace3112   -3.416e-01  5.443e-01   -0.628 0.530298
frace3113    6.976e-02  6.749e-02    1.034 0.301270
frace3114   -6.176e-02  1.444e-01   -0.428 0.668825
frace3115   -1.869e-01  1.226e-01   -1.524 0.127398
frace3116   -1.325e-01  1.241e-01   -1.068 0.285363
frace3117    2.804e-01  4.415e-01    0.635 0.525389
frace3118    1.063e+00  5.748e-01    1.849 0.064491 .
frace3119   -4.717e-02  2.250e-01   -0.210 0.833944
frace3120    1.075e-01  4.988e-01    0.216 0.829327
```

```
frace3121    1.088e-01  3.814e-01    0.285 0.775499
frace3122    1.859e-01  2.230e-01    0.833 0.404616
frace3123    4.232e-01  4.329e-01    0.978 0.328267
frace3124   -1.026e+00  8.956e-01   -1.146 0.251852
frace3125   -1.021e-01  1.174e-01   -0.870 0.384340
frace3126   -1.390e+00  9.688e-01   -1.435 0.151398
frace3127   -7.145e+00  3.272e+01   -0.218 0.827126
frace3128   -6.950e+00  3.033e+01   -0.229 0.818766
frace3129    3.251e-01  6.335e-01    0.513 0.607823
frace3130   -3.531e-01  3.977e-01   -0.888 0.374636
frace3131   -6.736e-01  1.412e+00   -0.477 0.633308
frace3199   -6.735e-02  1.470e-02   -4.583 4.59e-06 ***
```

```
fbrace2      1.978e-02  5.540e-02    0.357 0.721107
fbrace3      4.694e-02  7.239e-02    0.648 0.516732
fbrace4      5.851e-02  6.567e-02    0.891 0.372944
fbrace9             NA         NA       NA       NA
fhisp_r1     2.718e-02  1.280e-02    2.124 0.033647 *
fhisp_r2     2.344e-01  2.227e-02   10.527  < 2e-16 ***
fhisp_r3     1.089e-01  3.737e-02    2.914 0.003563 **
fhisp_r4    -6.465e-02  1.956e-02   -3.305 0.000950 ***
fhisp_r5     1.384e-01  1.770e-02    7.819 5.31e-15 ***
fhisp_r9     1.632e-01  4.668e-02    3.496 0.000473 ***
fracehisp2   1.710e-01  3.910e-02    4.373 1.23e-05 ***
fracehisp3  -5.324e-02  9.669e-02   -0.551 0.581866
fracehisp4  -2.159e-02  8.916e-02   -0.242 0.808673
fracehisp5   2.098e-01  1.926e-01    1.090 0.275895
fracehisp6   7.900e-02  5.178e-02    1.526 0.127067
fracehisp7          NA         NA       NA       NA
fracehisp8          NA         NA       NA       NA
fracehisp9   2.635e-01  4.866e-02    5.416 6.09e-08 ***
```

GUIDE classification tree using all variables & obs



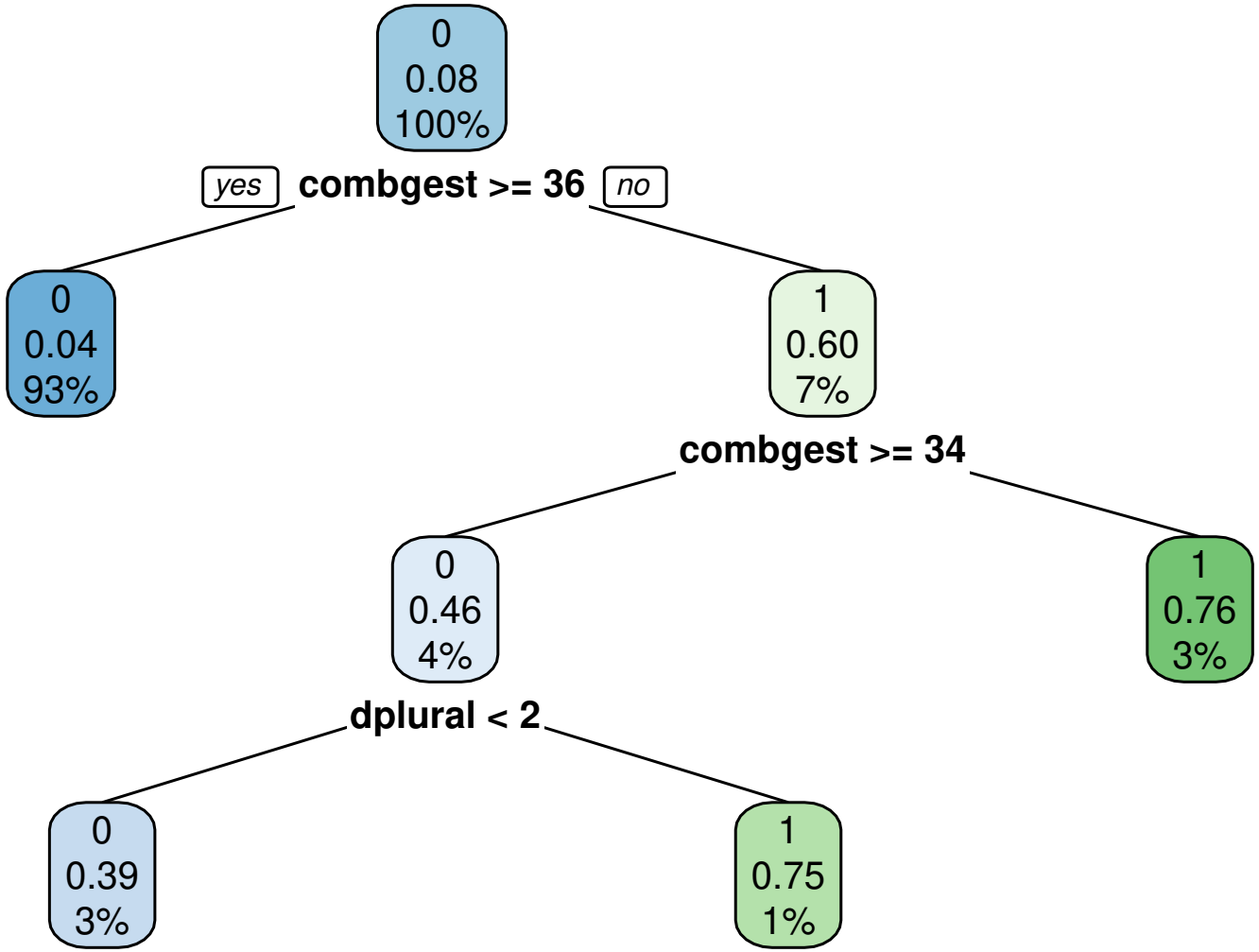
Estimated posterior P(lowbwt=1) beside nodes

Chi-squared tests

	dplural (X_4^2 = 447868, X_1^2 = 446312)				
lowbwt	1	2	3	4	5
0	3574458	59017	188	11	5
1	245578	73044	3579	206	26

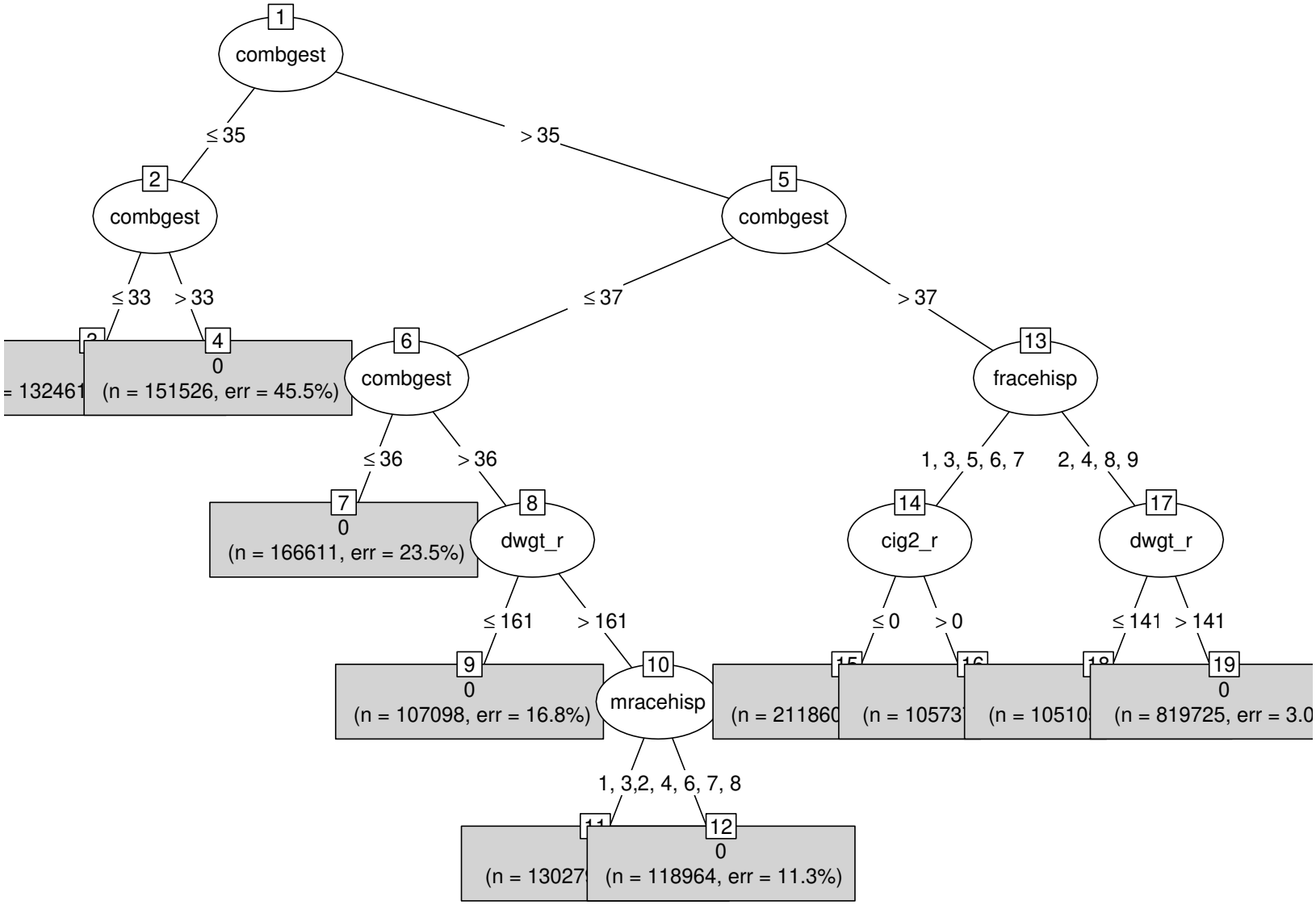
	combgest (X_2^2 = 115935, X_1^2 = 115583)		
lowbwt	(16,39]	(39,47]	NA
0	2332002	1298559	3118
1	301908	20127	398

RPART classification tree for low birthweight



CTREE constrained to have at least 10^5 obs/node

Tree with default parameters has 1206 terminal nodes



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.

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)

Multiple missing-value codes: Consumer Expenditure (CE) Data

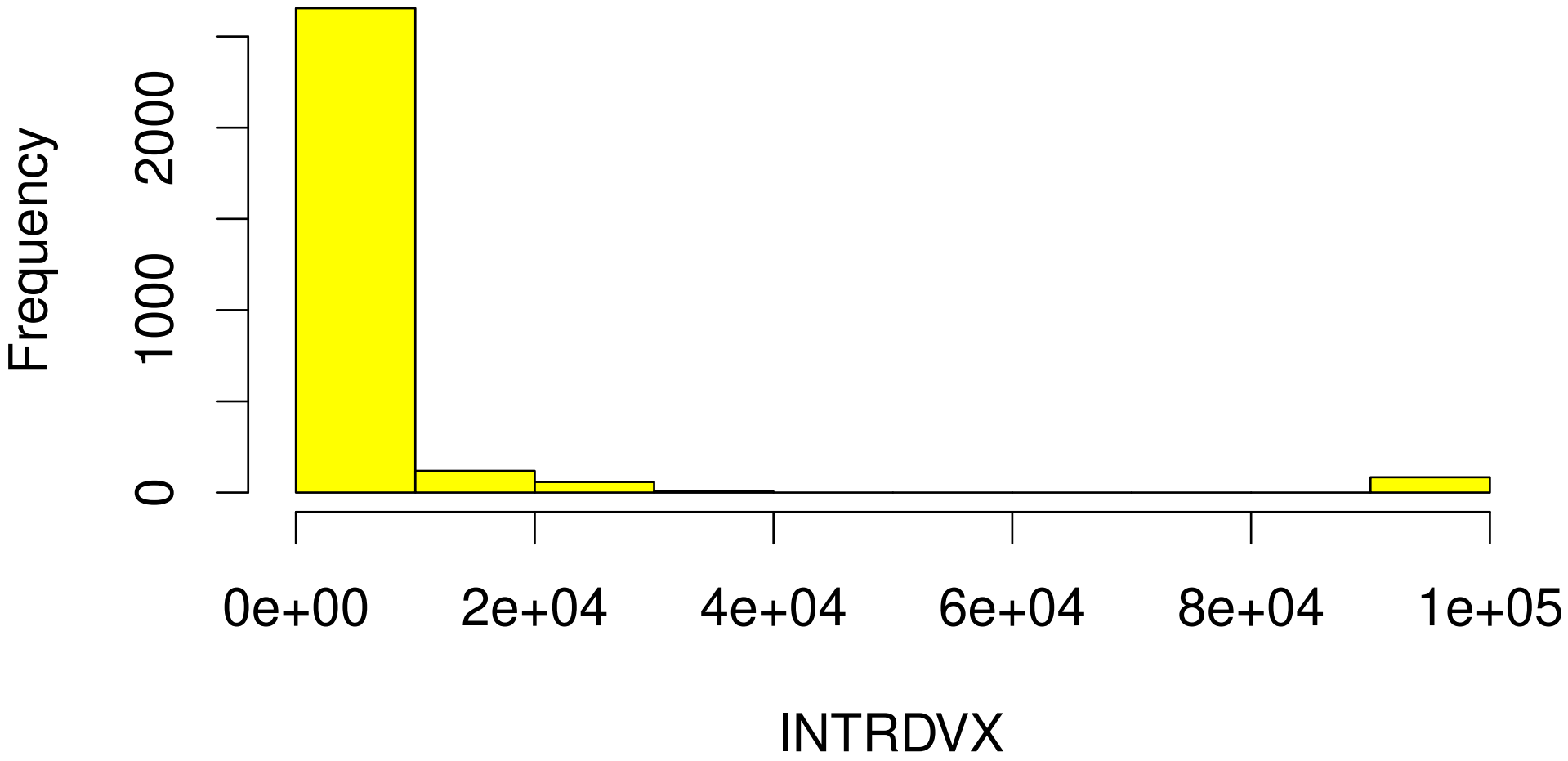
- 2013 Consumer Expenditure Survey, Bureau of Labor Statistics
- 25,822 consumer units (CUs) interviewed quarterly on hundreds of items
- Goal: estimate population mean interest and dividend (INTRDVX)
- Top 3% of INTRDVX are "topcoded" (above $32,000 changed to $98,338)
- 4693 CUs remain after deleting those with valid nonresponse in INTRDVX (INTRDVX_ = A): 1771 missing and 2922 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

INTRDVX_ is missing-value flag variable for INTRDVX

Histogram of 2922 nonmissing INTRDVX values



Some variables and their proportions missing

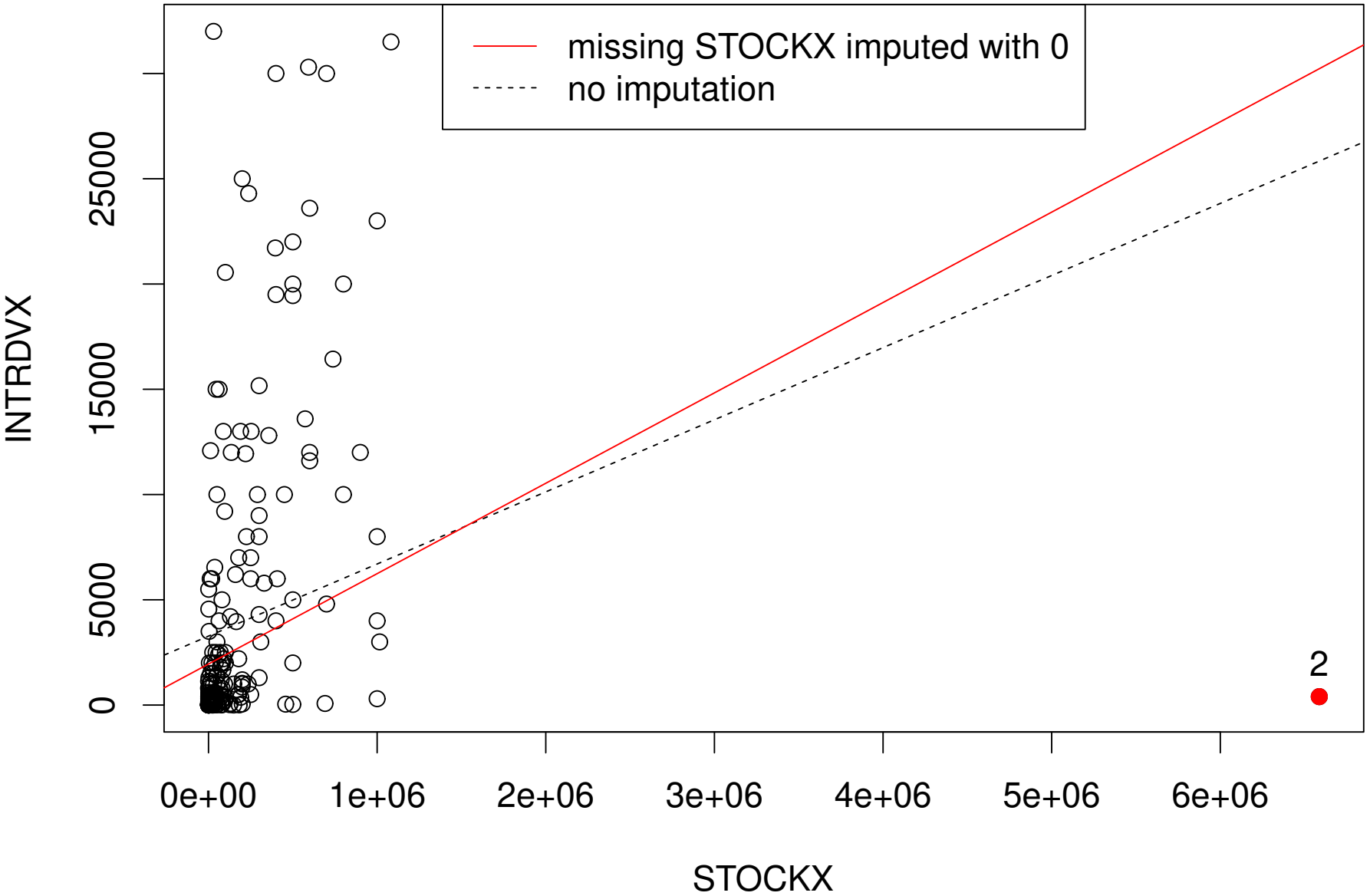
Name	Definition	Prop
AGE_REF	Age of reference person	
AGE2	Age of spouse	0.41
AS_COMP3	Number of males age 2 through 15 in CU	
BUILT	Year range property was built	0.13
CUTENURE	Housing tenure	
EDUCA2	Education of spouse	0.41
EMRTPNOP	Mortgage principal outlays last quarter	
ERANKH	Percent expenditure 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

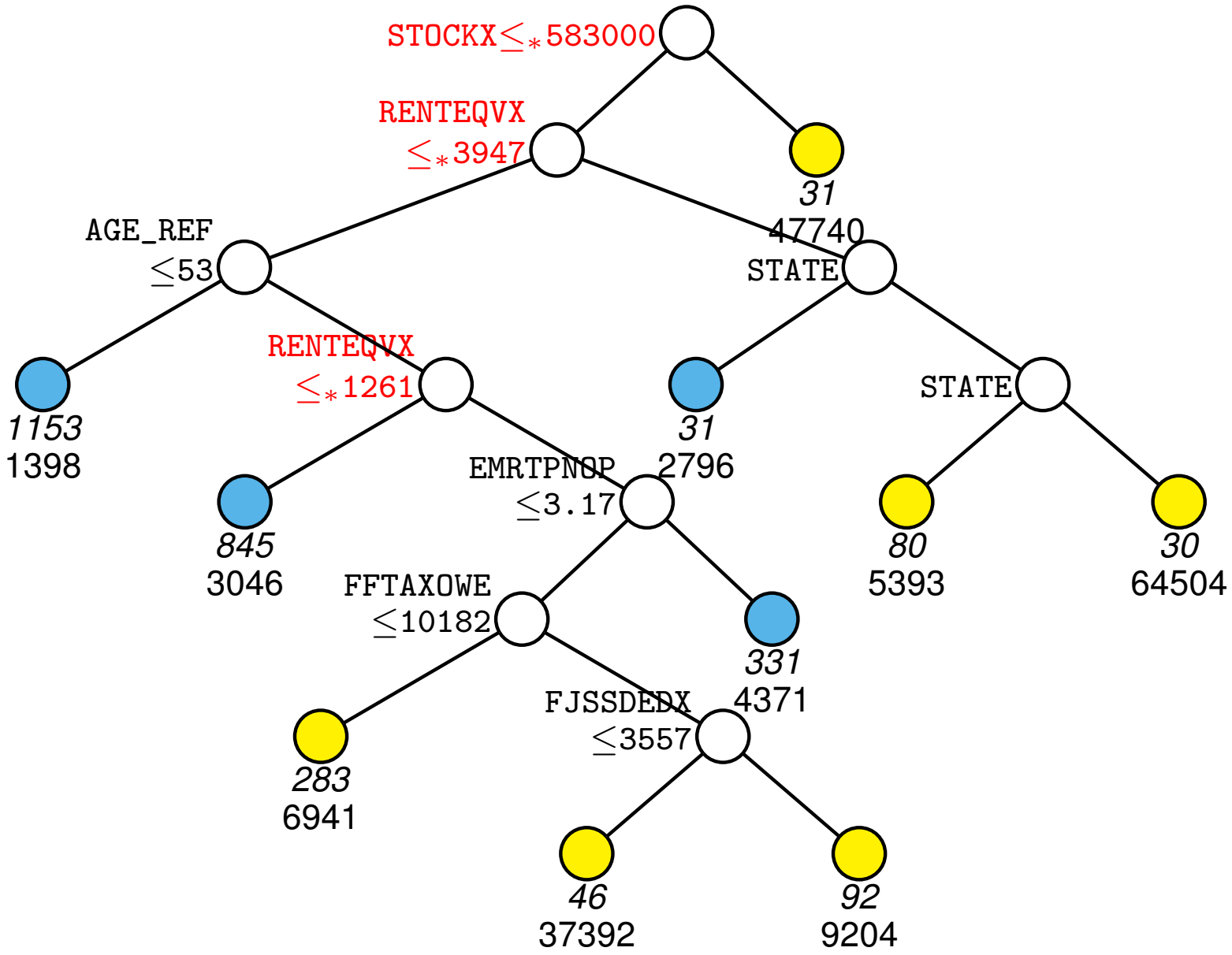
INCOMEY2	Employer from which spouse received most earnings during the past 12 months	0.61
INCWEEK1	Weeks worked full or part time last 12 months	
IRAX	Total value of retirement accounts	0.84
LIQUIDB	Bracket range of bank accounts	0.97
LIQUIDX	Total value of checking, savings, CD, etc., accounts	0.83
LIQUDYRX	Total value of bank accounts one year ago	0.84
NO_EARNR	Number of earners	
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 (1=complete, 2=incomplete)	
RETPENPQ	Retirement, pensions, Social Security last quarter	
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
ROYESTX	Royalties and income from estates and trusts	0.93
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
TOTTXPDX	Personal taxes paid by CU in past 12 months	
TOTXEST	Estimated total taxes paid	
UTILCQ	Utilities, fuels and public services this quarter	

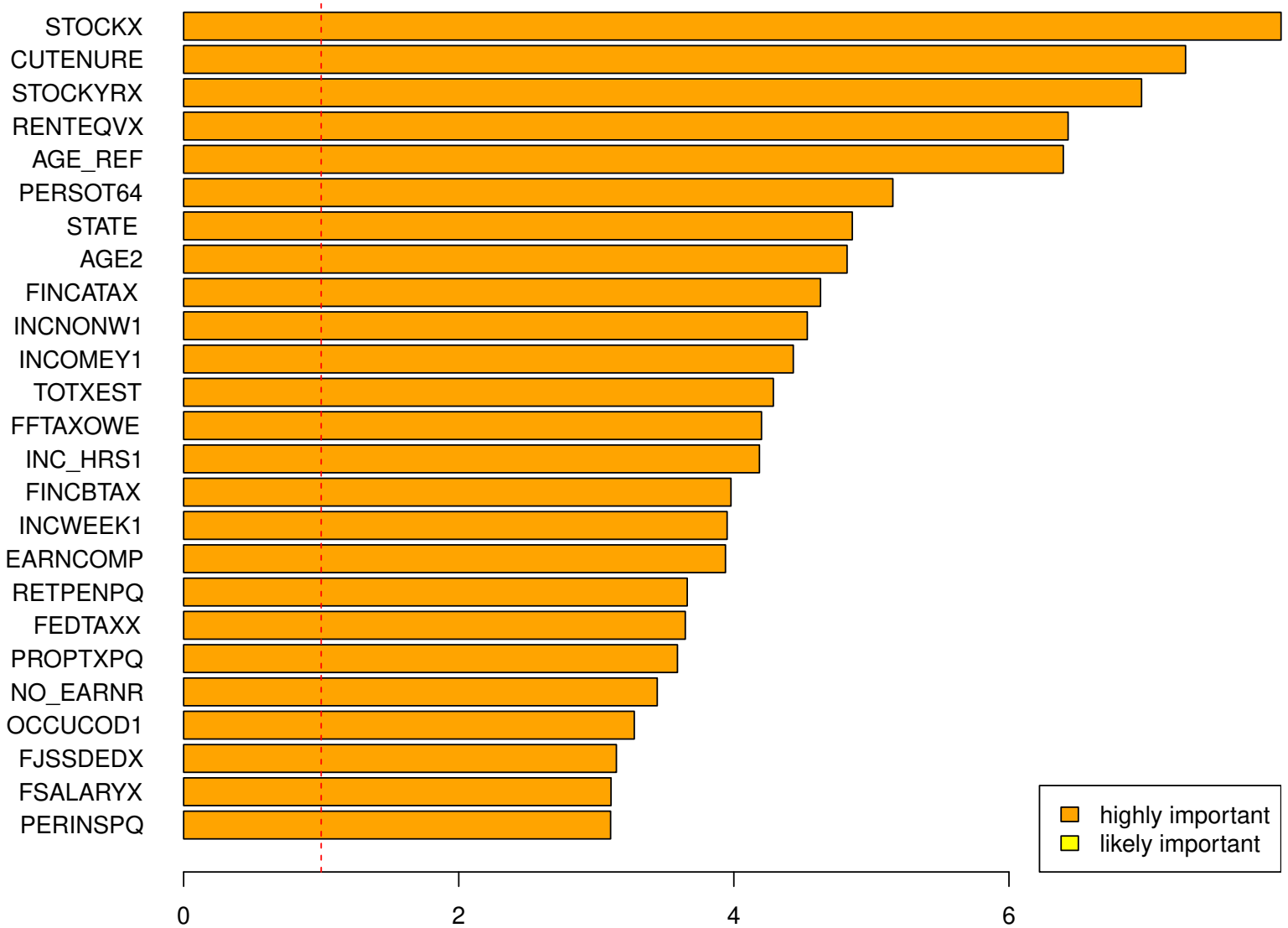
Plot of 6% of data nonmissing STOCKX



GUIDE regression tree for predicting INTRDVX

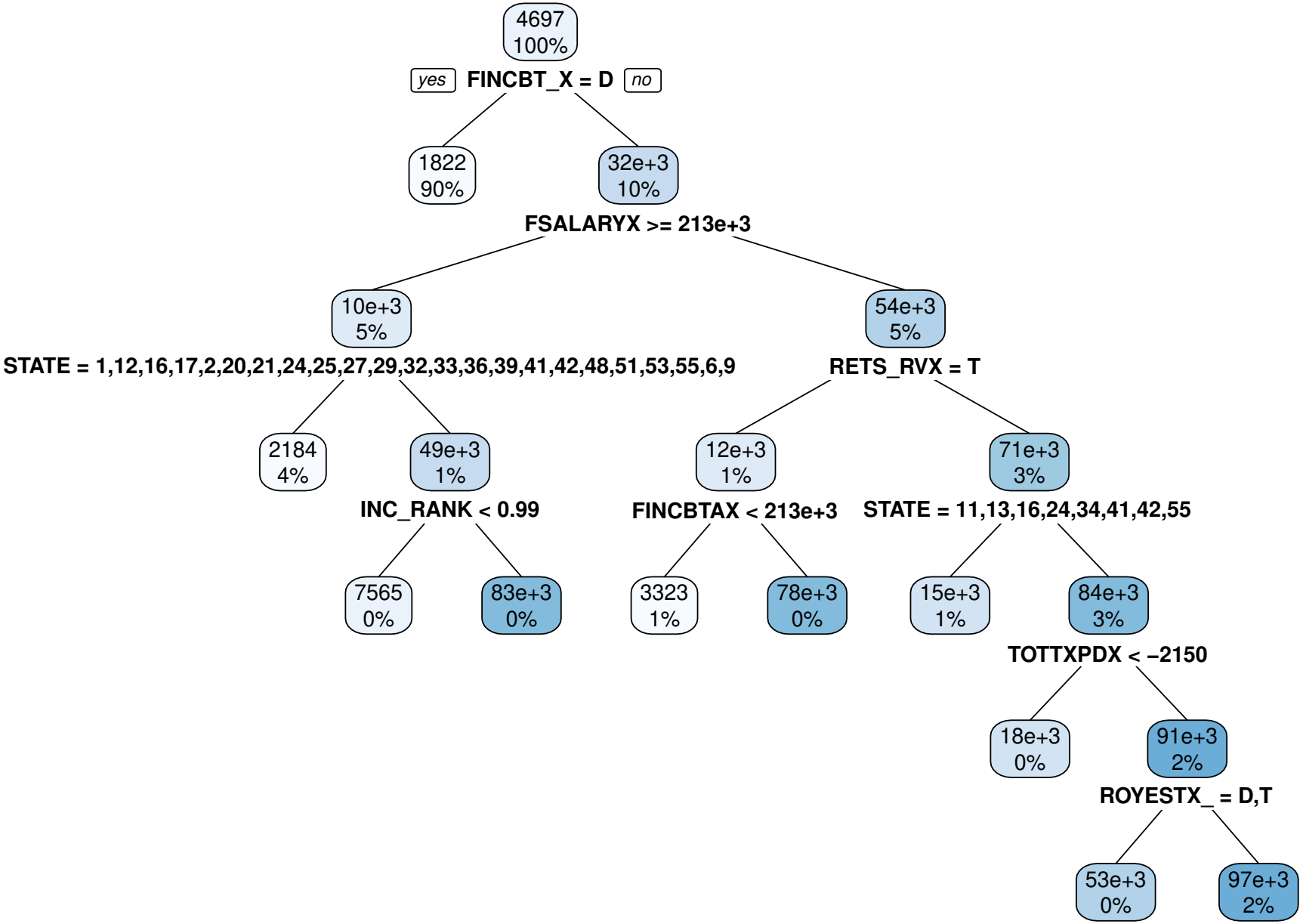


Top 25 predictors of INTRDVX



72 highly important, 21 likely important, 355 unimportant nontrivial predictors in total

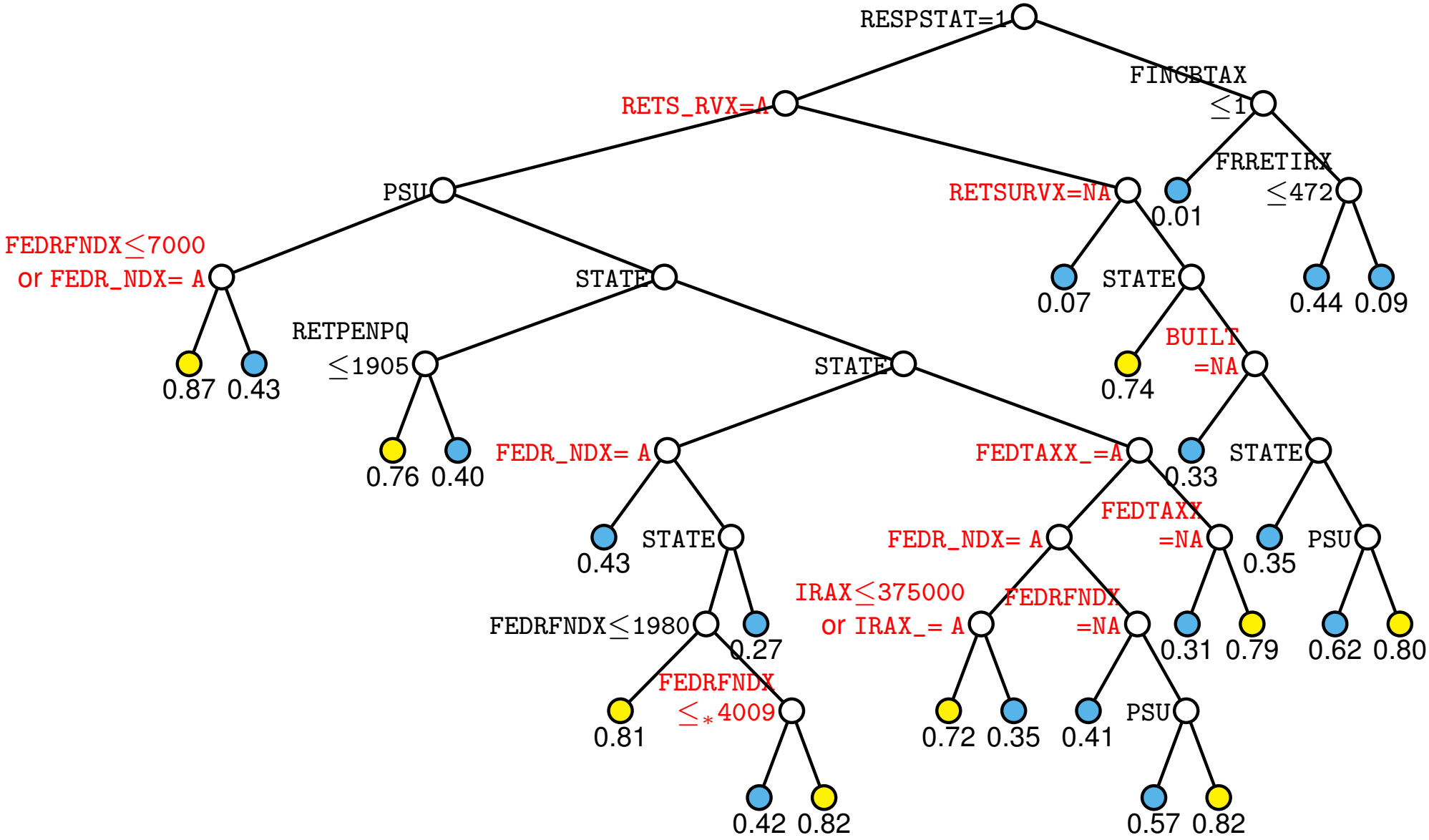
RPART regression tree for predicting INTRDVX



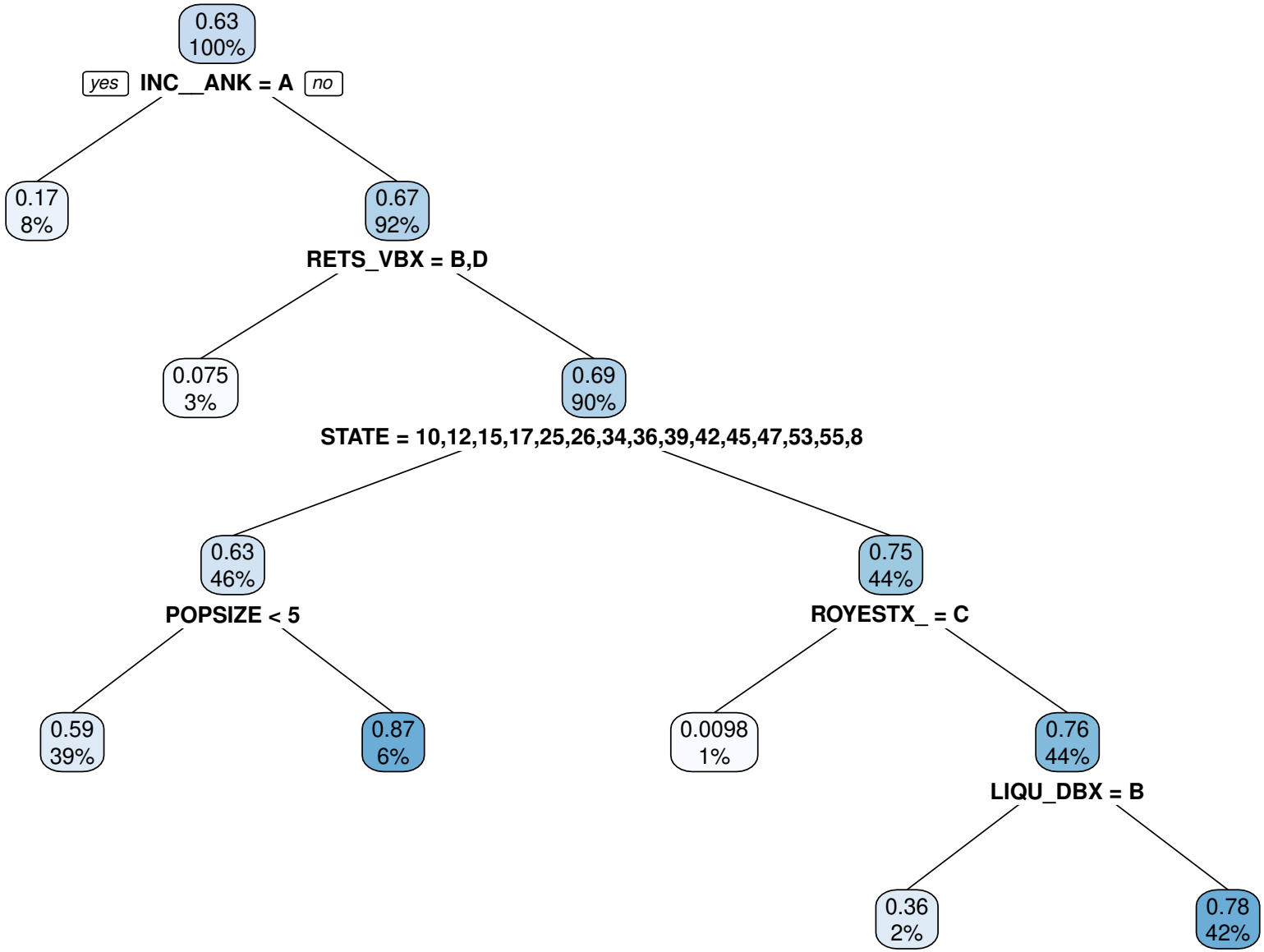
CTREE regression not applicable to CE data

1. `party` and `partykit` allow replicate weights but not sampling weights
2. `partykit` does not allow categorical variables with more than 31 levels

GUIDE regression tree for P(INTRDVX nonmissing)



RPART regression tree for P(INTRDVX nonmissing)



Two approaches to mean estimation

- Let μ be the population mean
- Let S_1 and S_2 be the subsets of nonmissing and missing y_i, respectively
- Let $\hat{\pi}_i$ be the estimated probability that y_i is nonmissing
- Let $\hat{y}_i$ be an estimate of y_i if it is missing
- Let w_i be the sampling weight (if any)

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

$$\Big(\sum_{i\in S_1} w_i/\hat{\pi}_i\Big)^{-1}\sum_{i\in S_1} w_i y_i/\hat{\pi}_i$$

Missing value estimation (MVE). The MVE estimate is

$$\Big(\sum_{i\in S_1\cup S_2} w_i\Big)^{-1}\Big(\sum_{i\in S_1} w_i y_i + \sum_{j\in S_2} w_j\hat{y}_j\Big)$$

Estimates of INTRDVX population mean

Type	Method	Non-tree	Tree	Forest
	Weighted average of nonmissing values	4697		
IPW	Lasso logistic regression[a]	4303		
IPW	GUIDE classification[b]		4736	5005
IPW	GUIDE regression		4557	4786
IPW	CTREE/CFOREST classification[c]		4446	4695
IPW	RPART/RF[d] classification		4425	4735
IPW	RPART/RF[d] regression		4525	4725
MVE	Weighted least-squares regression[e]	4726		
MVE	GUIDE regression		4833	4681
MVE	RPART/RF[d] regression		3997	–

[a] With mean imputation and addition of missing-value indicators; ordinary logistic fails

[b] GUIDE treats all positive sampling weights as 1 in classification

[c] Sampling weights not allowed (only integer-valued replicate weights) in party (and partykit)

[d] RF does not allow sampling weights

[e] With mean imputation, missingness indicators & deleting variables with levels in S_2 but not S_1

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 two practical constraints:

1. Number of splits **increases linearly** in n and **exponentially** in c for ordinal and categorical, resp., X
2. Computationally expensive for other than **piecewise constant** 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

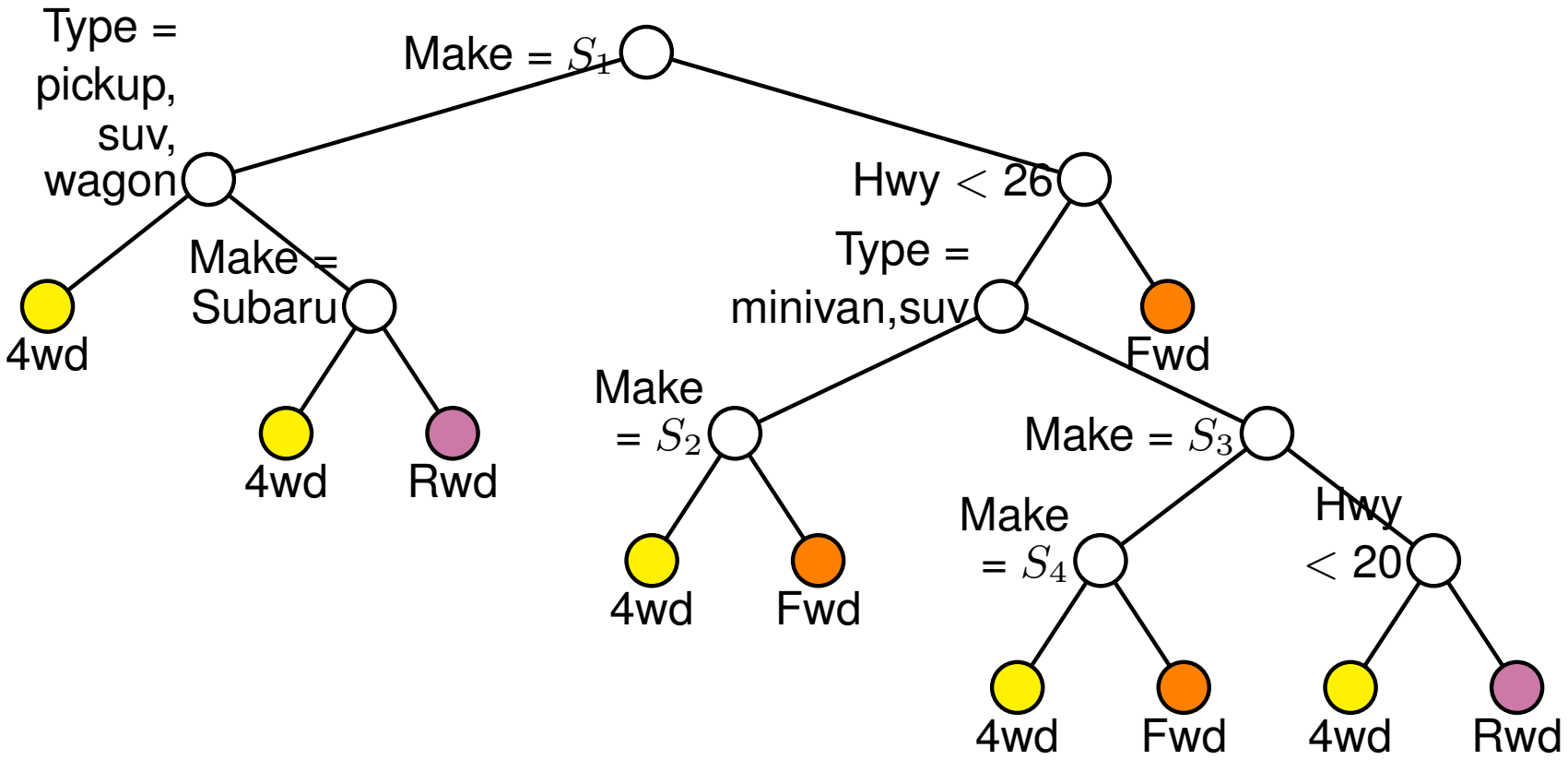
Predictor variables

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!

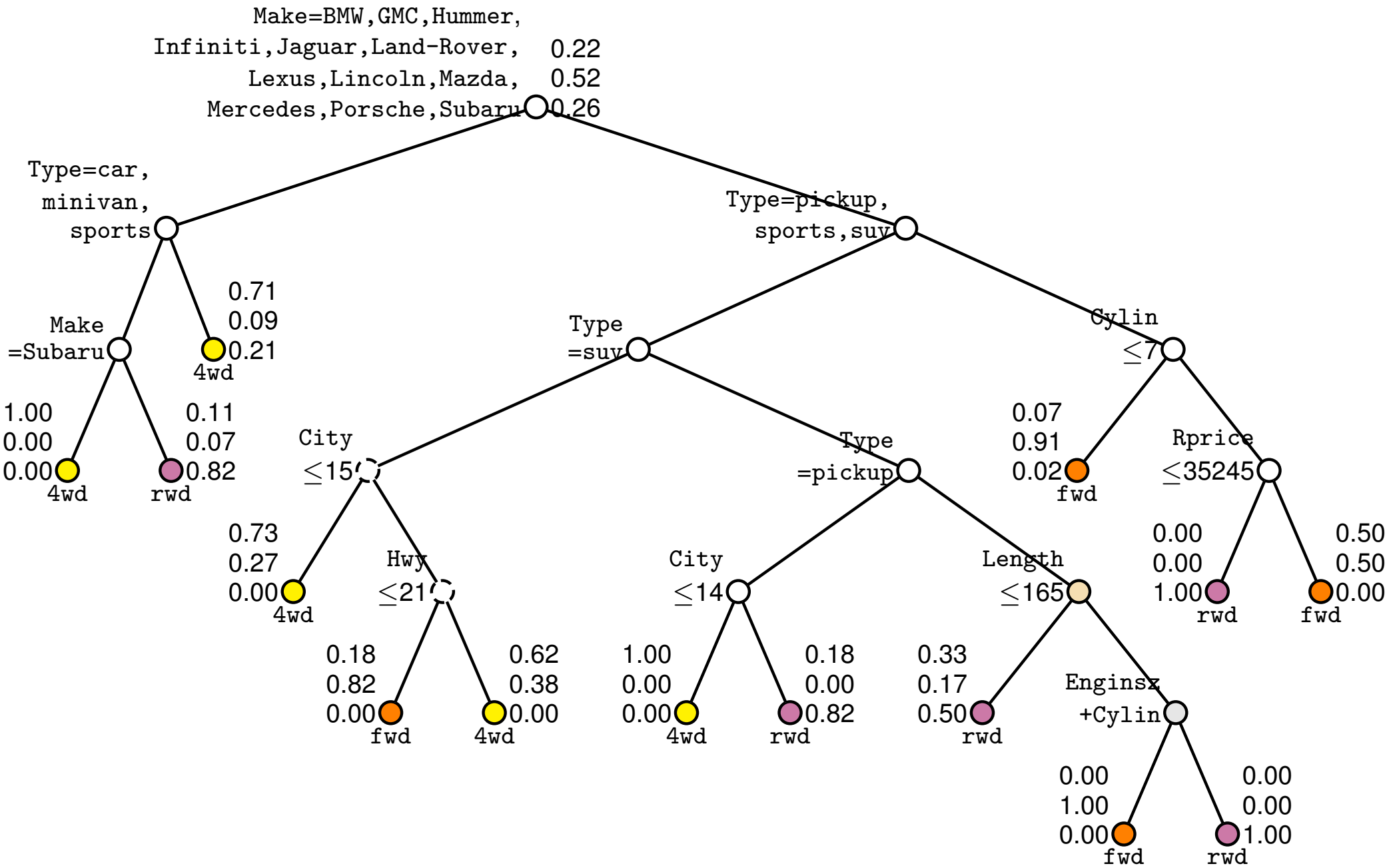
`party` accepts categorical variables with $>$ 31 levels, but not `partykit`

RPART tree for car data (36 cpu hrs!)

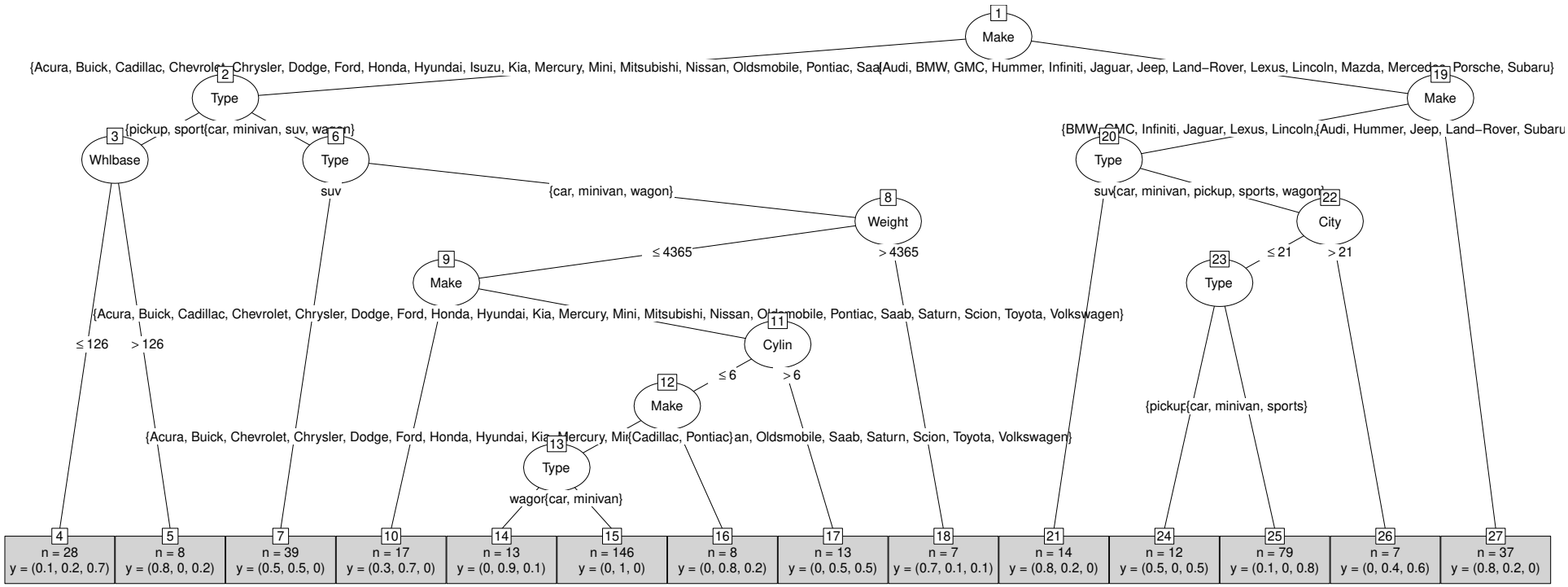


S_1 = {BMW, GMC, Hummer, Infiniti, Jaguar, Land-Rover, Lexus, Lincoln, Mazda, Mercedes, Porsche, Subaru}; S_2 = {Acura, Buick, Chevrolet, Dodge, Ford, Honda, Isuzu, Jeep, Mitsubishi, Pontiac, Suzuki, Toyota, Volkswagen, Volvo}; S_3 = {Audi, Kia, Mitsubishi, Nissan, Pontiac, Volkswagen, Volvo}; S_4 =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

GUIDE tree for car data (0.5 sec.)



CTREE (party) tree for car data



partykit inapplicable because Make has more than 31 categories

Leave-one-out error counts for car data[a]

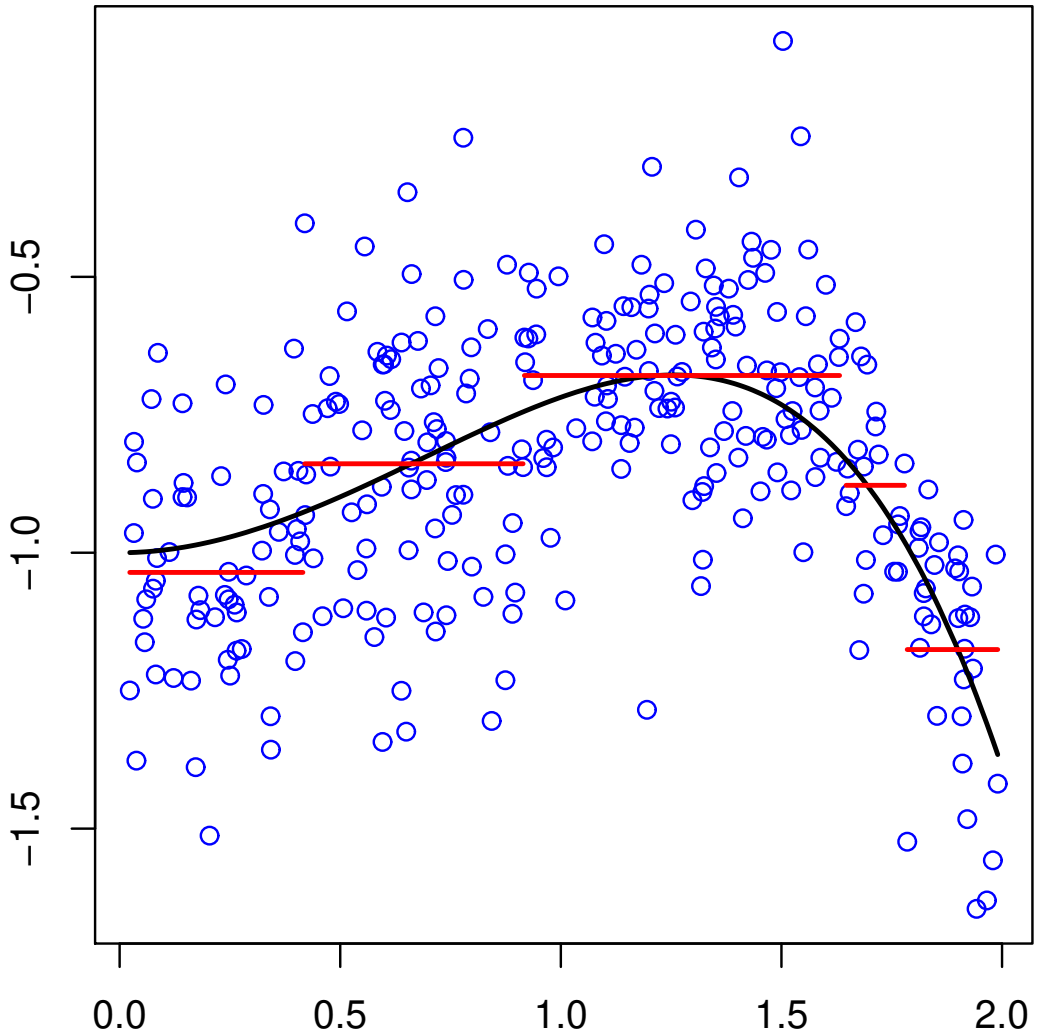
Method	Errors	Time[b]
GUIDE forest	74	21.29
CFOREST (party)	78	8.44
Linear discriminant analysis (LDA)	84	0.02
GUIDE tree	99	0.68
CTREE (party)	111	0.07
Logistic regression[c]	-	-
RPART[d]	-	!
randomForest and CTREE, CFOREST (partykit)[e]	-	-

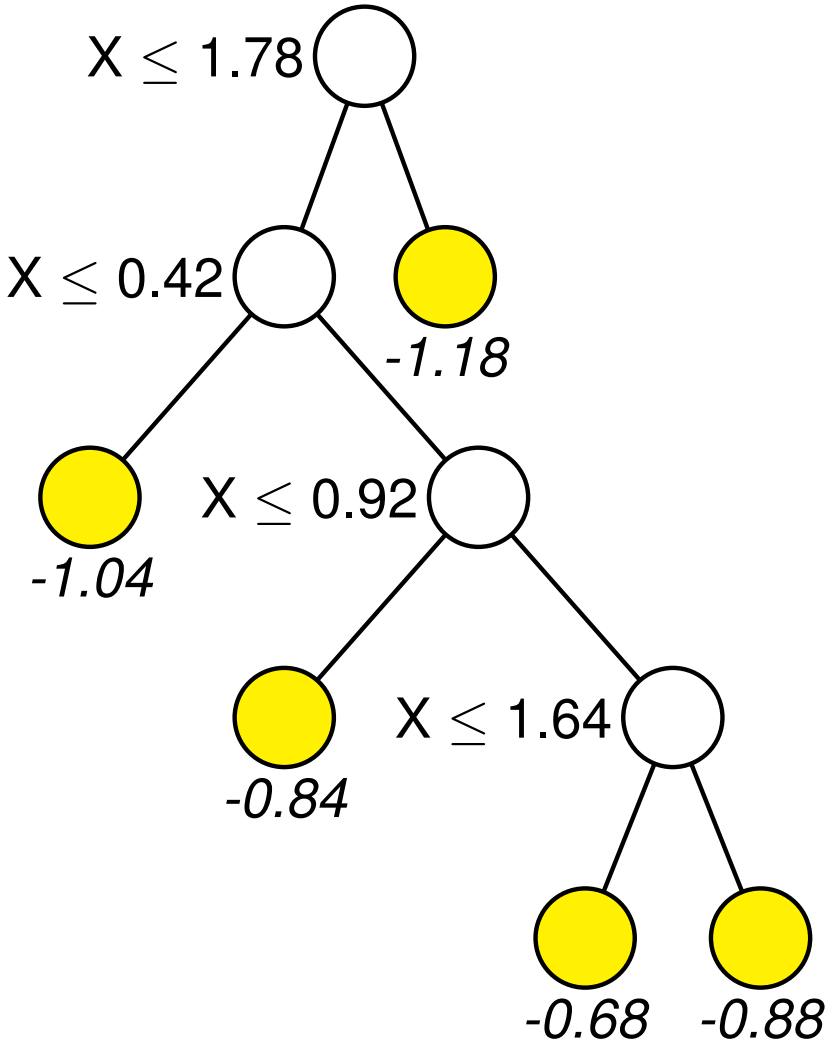
[a] out of 427 obs, excluding single Hummer which crashes LDA and CFOREST
[b] average time (sec.) to fit one data set
[c] logistic regression does not converge
[d] RPART would take 36×427 hrs $\approx$ 1.8 years to complete
[e] inapplicable to categorical variables with more than 31 levels

CART regression

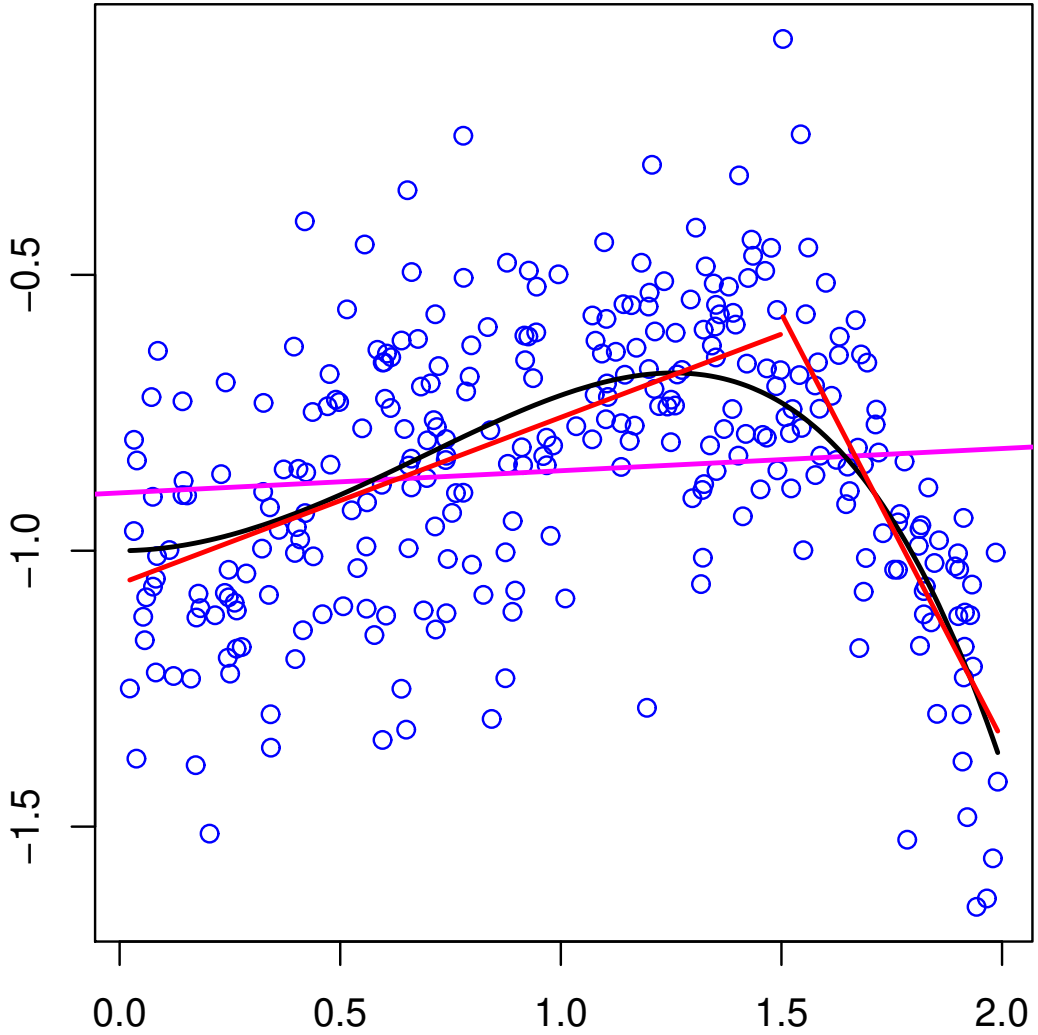
- Fit a constant $\bar{y}$ to each node
- Use residual sum of squares as node impurity and error measure
- Everything else the same as in CART classification

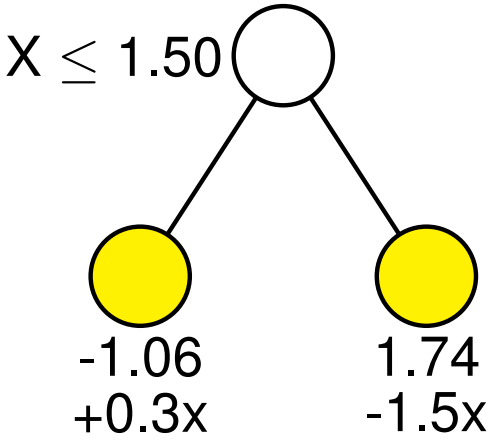
Piecewise-constant regression model





Piecewise-linear regression model





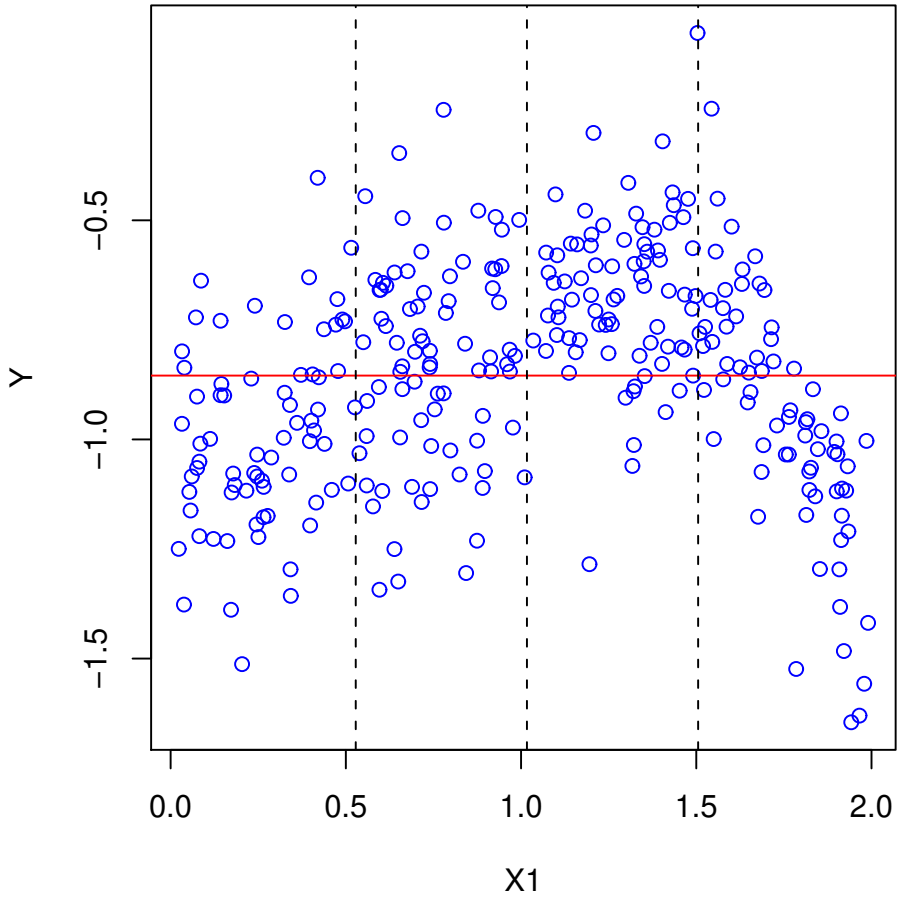
GUIDE regression tree models

- 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

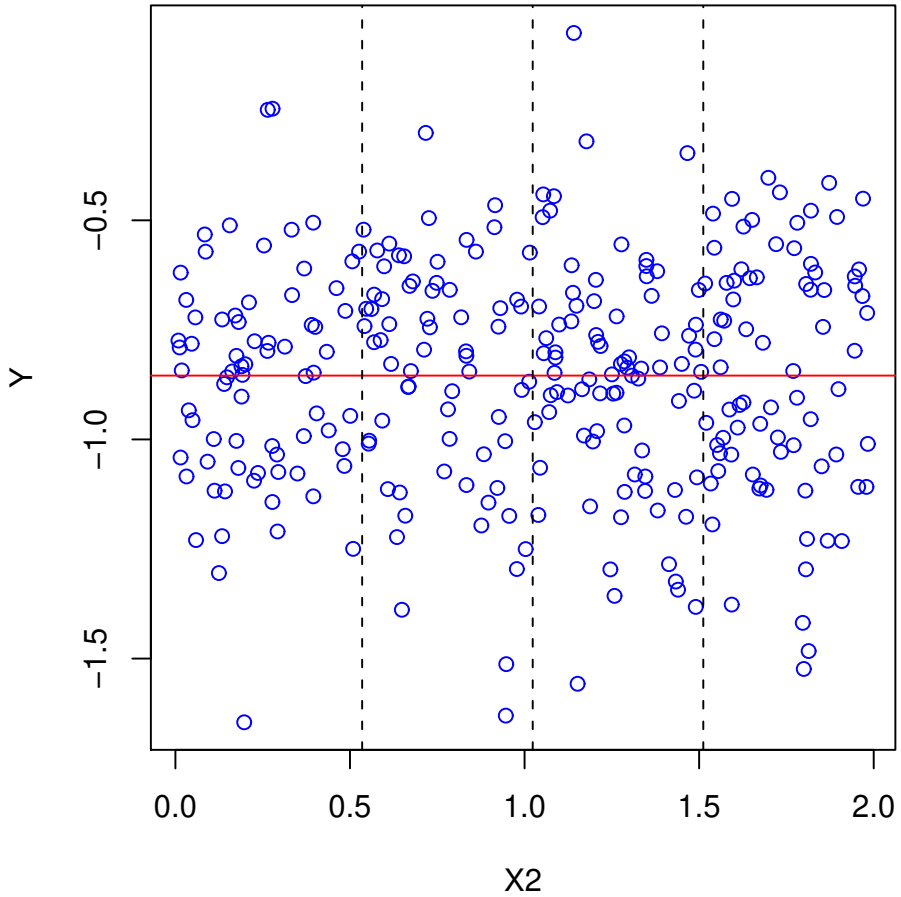
1. Fit a model to the data in the node and obtain the residuals
2. Define a “class” variable that equals +1 if residual is positive, -1 otherwise
3. Follow GUIDE classification procedure to select a variable to split node

Split variable selection based on residual patterns



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

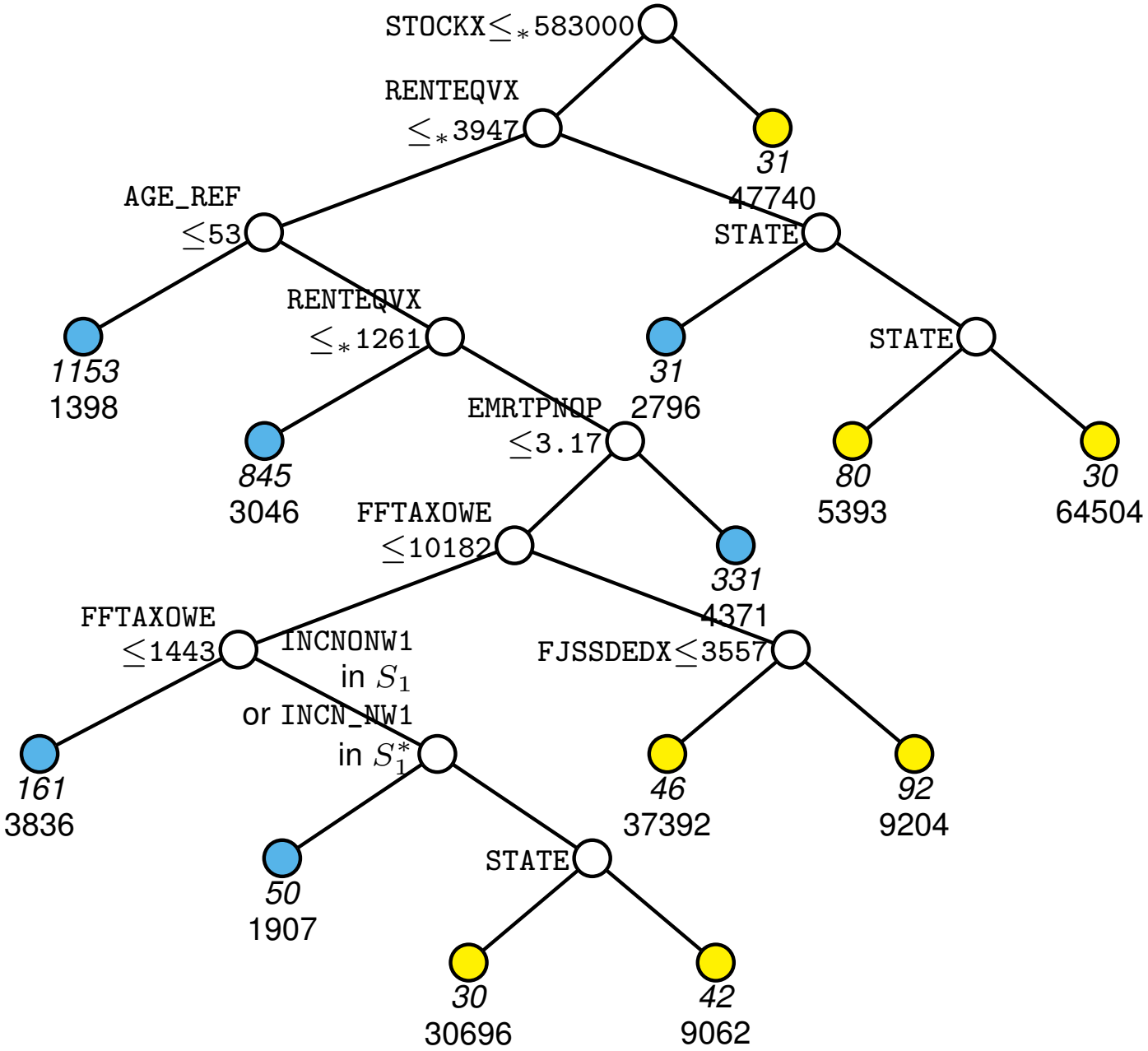
χ^2_3 = 66.7, p = 2×10^{-14}



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

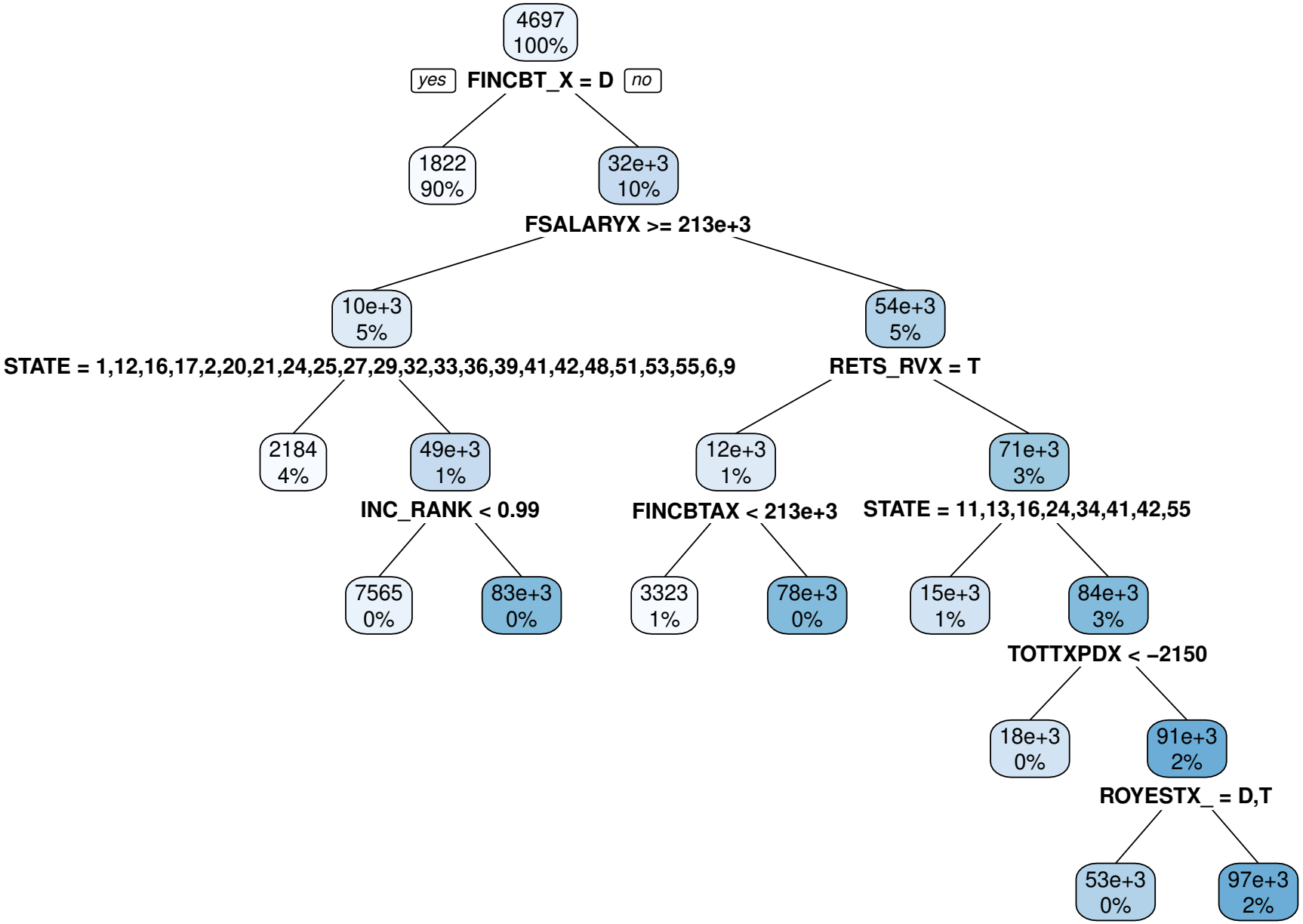


CE data (weighted mean INTRDVX = $4778)

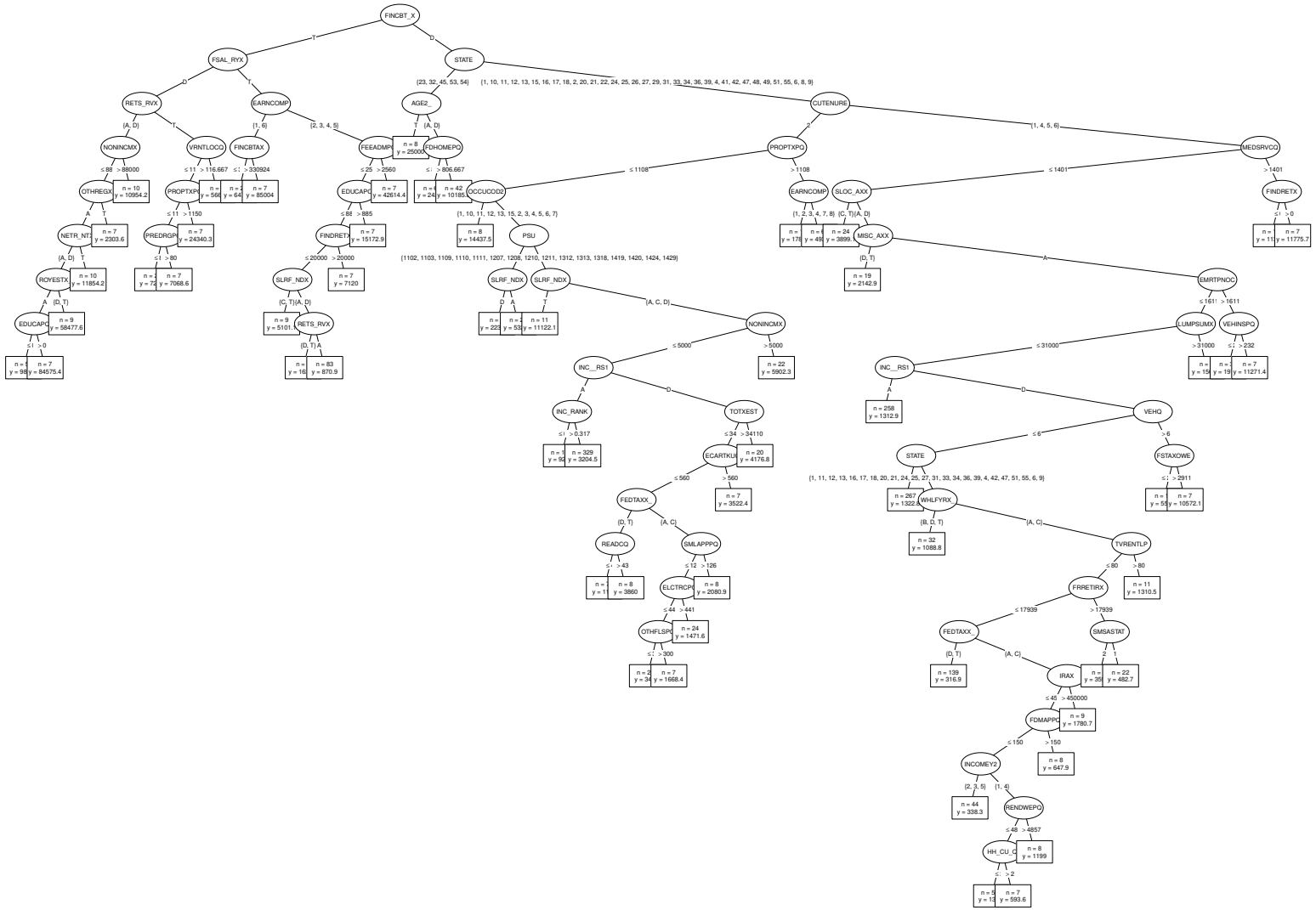
	AGE_REF			
	$\leq$ 43	(43, 58]	(58, 68]	$>$ 68
$>$ $4778	33	78	152	147
$\leq$ $4778	718	684	571	539
χ^2_3 = 127.3, p $<$ 2.2E-16, χ^2_1 = 108.4				

	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

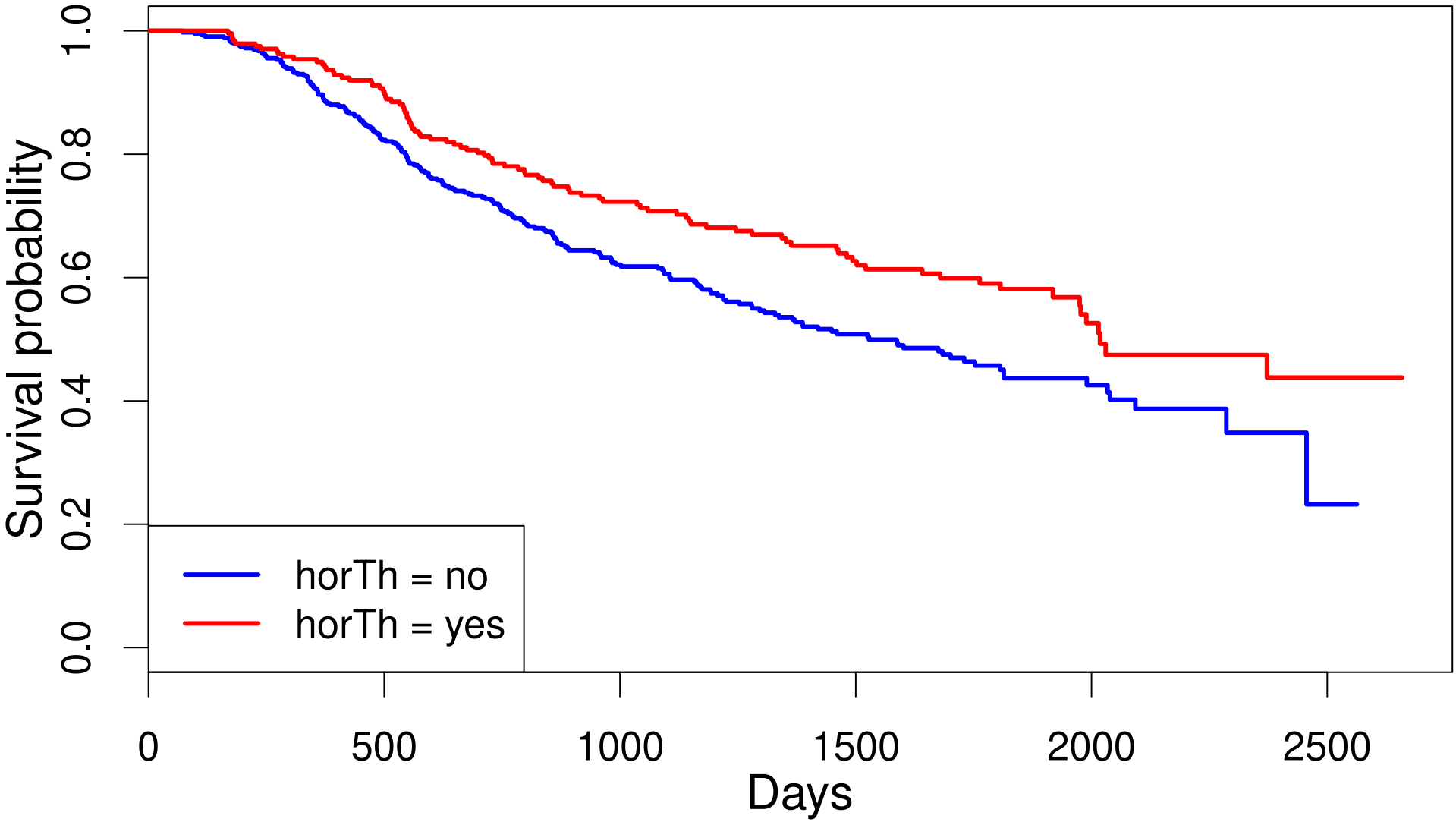


`partykit` does not work here; neither allows sampling weights

Censored data: breast cancer trial

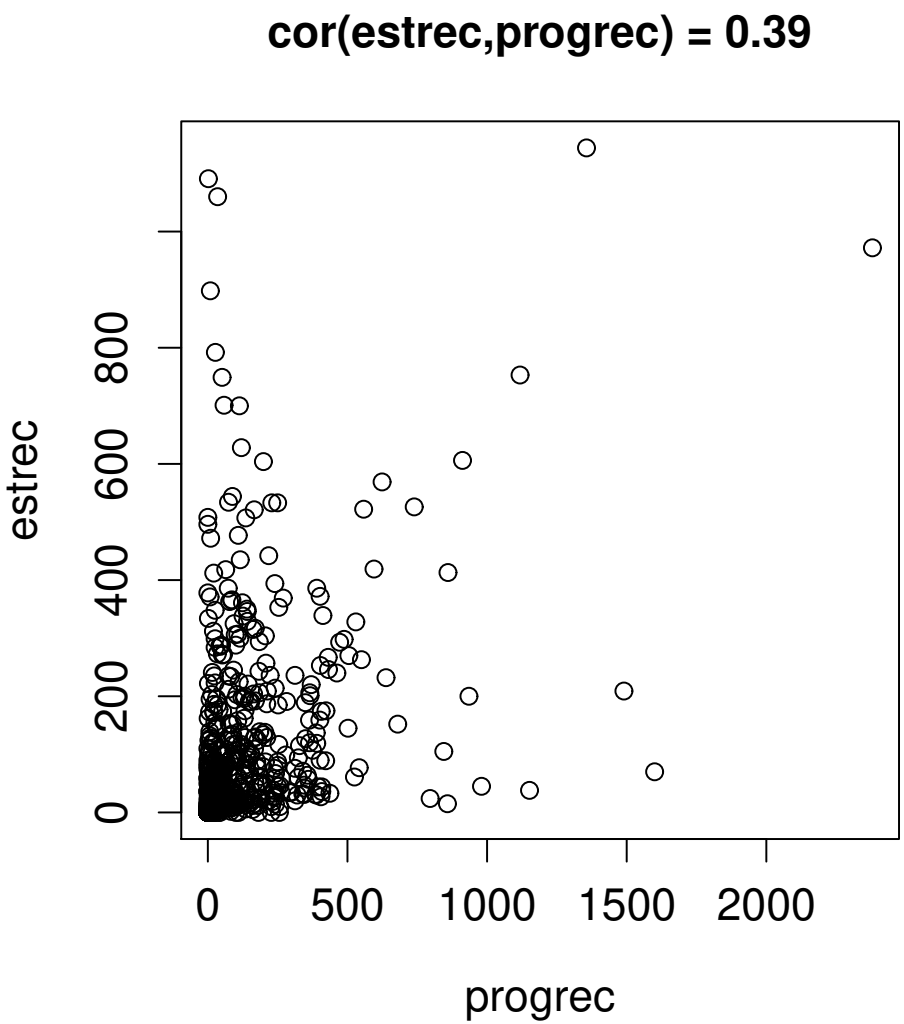
- Randomized clinical trial of 672 subjects with primary node positive breast cancer (Schumacher et al., 1994); 14 subjects with censored times less than smallest uncensored time excluded; data from TH.data R package
- Response is recurrence-free survival time (8–2659 days, 299 uncensored, 387 censored)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

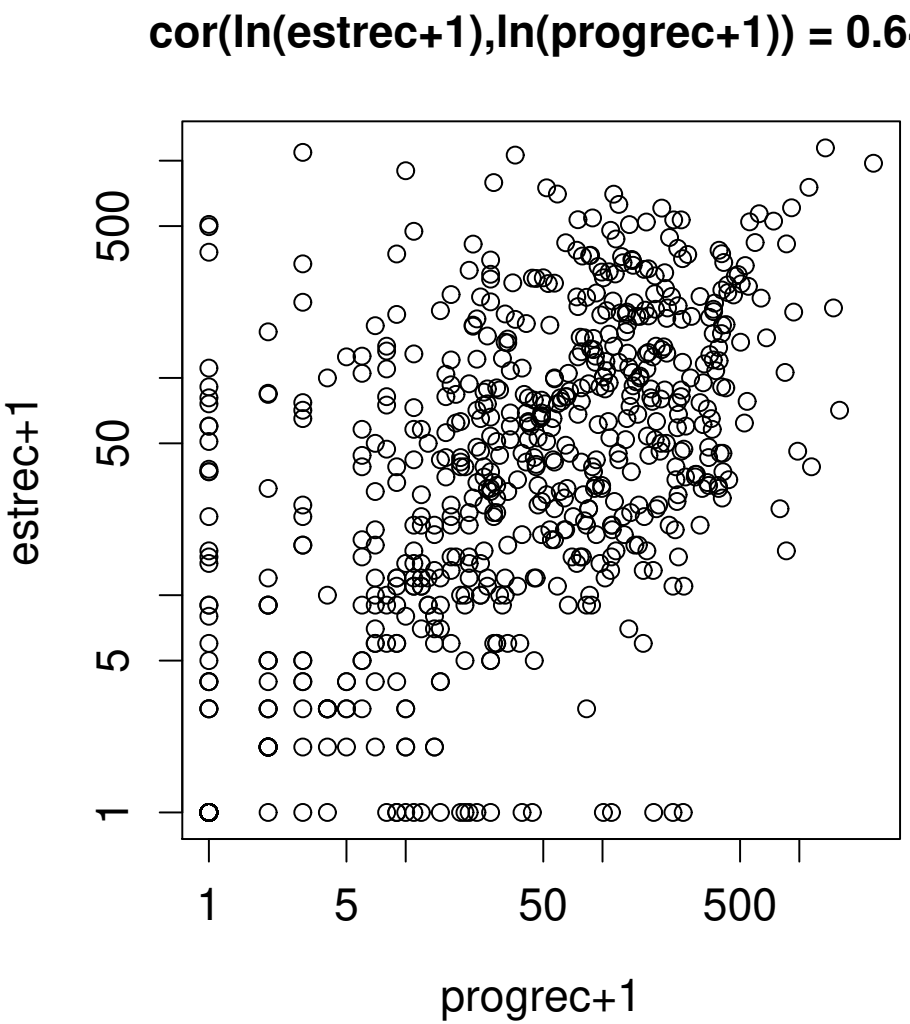
Kaplan-Meier survival curves



Proportional hazards model

Variable	Coef	p-value	Variable	Coef	p-value
horTh=yes	-0.3372	0.0089	tgrade	0.2803	0.0082
age	-0.0094	0.3111	pnodes	0.0499	1.7e-11
meno=Pre	-0.2673	0.1449	progrec	-0.0022	0.0001
tsize	0.0077	0.0507	estrec	0.0002	0.7084


cor(estrec,progrec) = 0.39
estrec
0
200
400
600
800
0
500
1000
1500
2000
progrec


cor(ln(estrec+1),ln(progrec+1)) = 0.64
estrec+1
1
5
50
500
1
5
50
500
progrec+1

Question

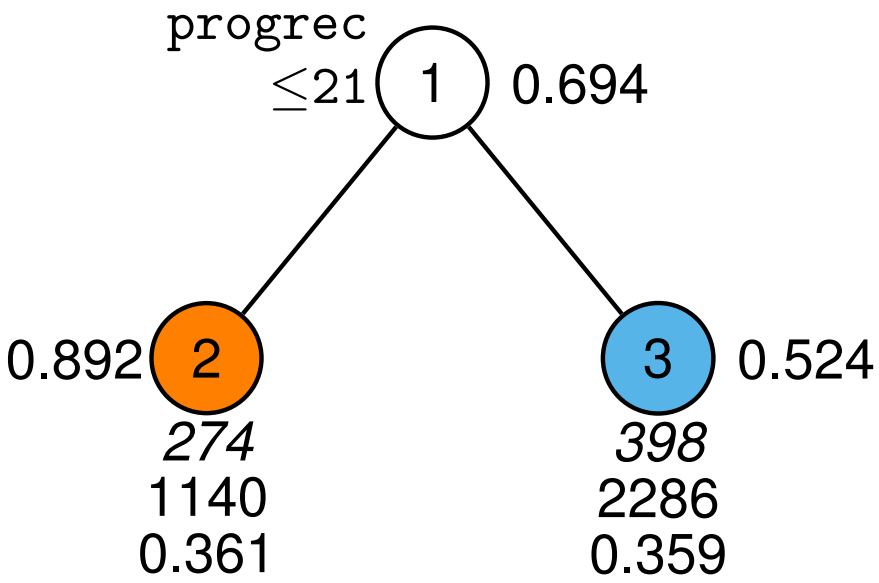
- Is there a subgroup where hormone therapy is ineffective?
- If affirmative, then the therapy in the complementary subgroup must be more effective than average

Cox model with treatment interactions

Variable	Coef	p-value	Variable	Coef	p-value
horThyes	-1.3741	0.322			
horThyes:age	0.0099	0.639	age	-0.0126	0.256
horThyes:menostatPre	0.0834	0.848	menostatPre	-0.3176	0.135
horThyes:tsize	0.0017	0.838	tsize	0.0074	0.148
horThyes:tgrade	0.1879	0.429	tgrade	0.2335	0.065
horThyes:pnodes	0.0255	0.173	pnodes	0.0425	2.7e-06
horThyes:progrec	-0.0028	0.054	progrec	-0.0014	0.025
horThyes:estrec	0.0003	0.736	estrec	-0.0002	0.771

- Main effect of horTh and all its interactions not significant!
- Are there no subgroups with differential treatment effects?

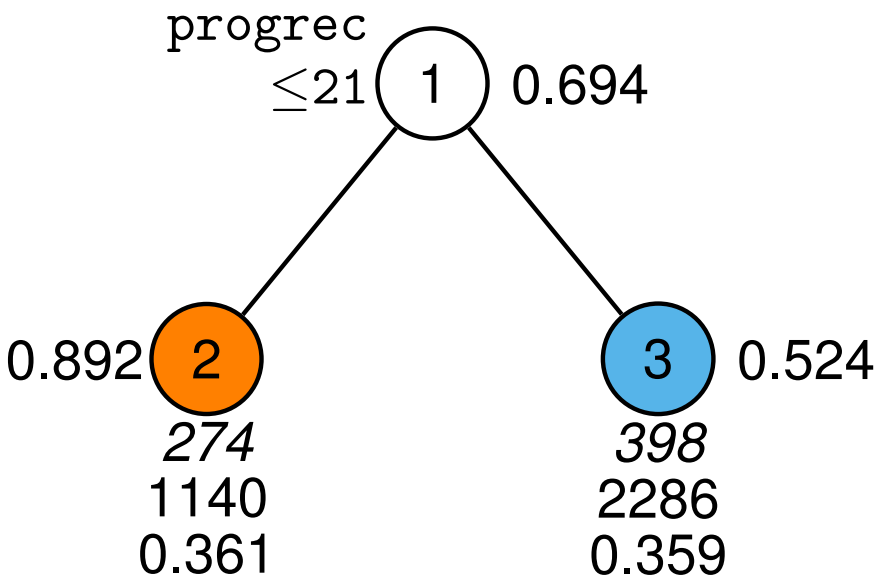
GUIDE Gi model

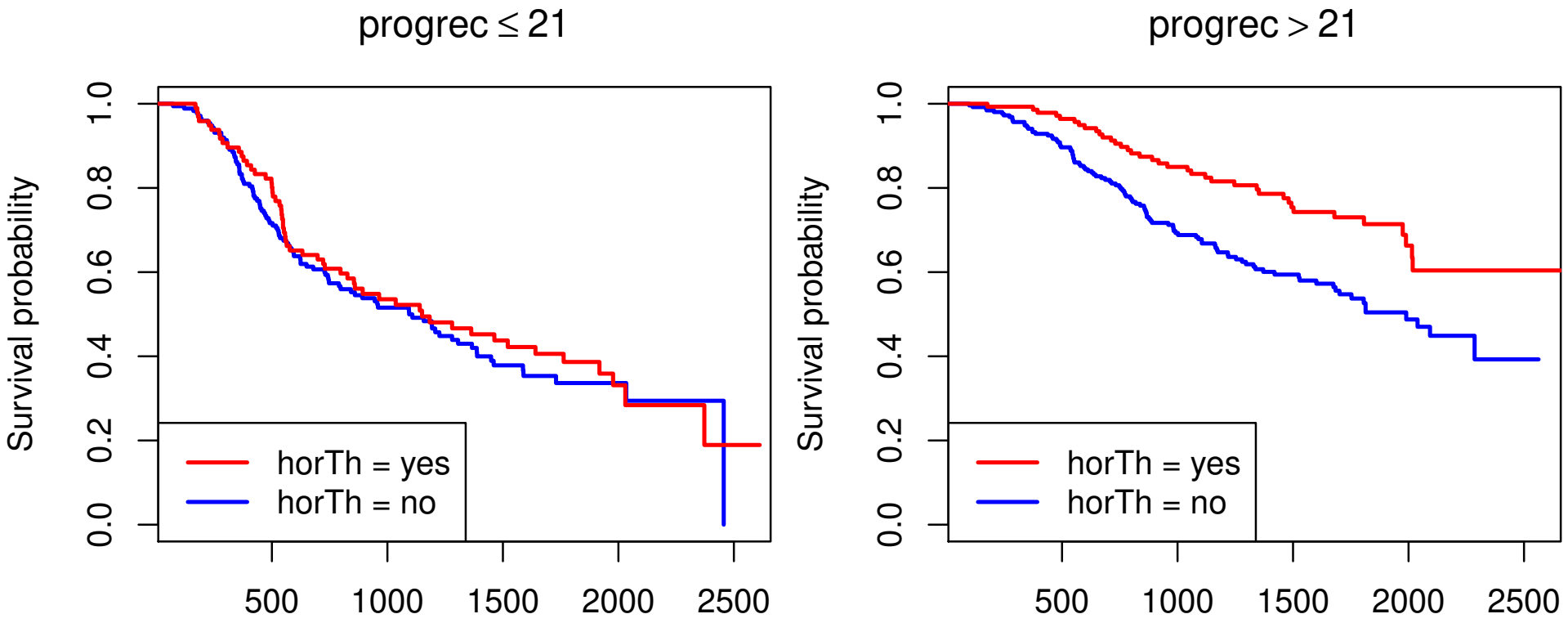


Hazard ratio of `horTh=yes` vs `no` beside nodes

Sample size (in *italics*), median survival time, P(`horTh = yes`) below nodes

Second best split variable is estrec


progrec
≤21
1
0.694
0.892
2
3
0.524
274
1140
0.361
398
2286
0.359


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

IBRANCE (Pfizer)

- IBRANCE is taken with a hormonal therapy and is used to treat hormone receptor positive (HR+), HER2- metastatic breast cancer
- Hormone receptor positive includes both ER+ (estrogen receptor positive) and/or PR+ (progesterone receptor positive) subtypes
- IBRANCE (palbociclib) is in a class of drugs called CDK 4/6 inhibitors that work to put the brakes on cell growth and division in both healthy and cancer cells
- `https://www.ibrance.com/ibrance-overview`

Verzenio (Lilly)

- Verzenio is a prescription medicine used to treat a type of breast cancer
- It is a medicine you can take if you have a type of breast cancer called HR+/HER2- (hormone receptor positive/human epidermal growth factor receptor 2 negative) and the cancer has spread to other parts of the body (metastasized)
- Verzenio is in a class of drugs known as CDK4 & 6 inhibitors. CDK4 & 6 are proteins that control how fast cells grow and divide.
- These proteins are found on both normal and cancer cells. They become overactive in metastatic breast cancer (MBC), causing cells to grow and divide uncontrollably. This leads to the spread of cancer.
- Verzenio interrupts these proteins and cells just as they are deciding to grow and divide. It slows down cancer cell growth and division, causing cancer cells to become inactive or even die.
- `https://www.verzenio.com/about`

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

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

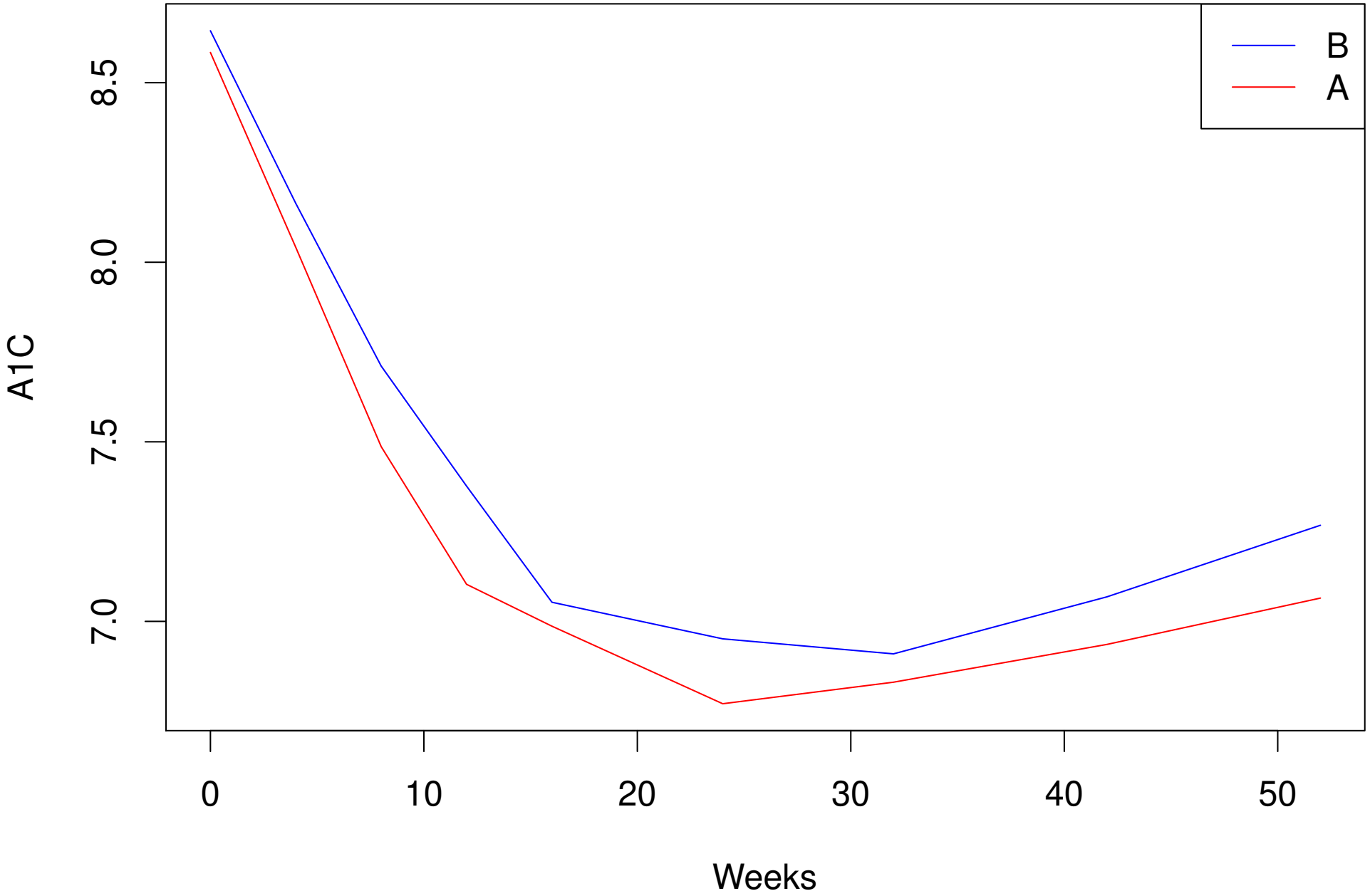
Insulin sensitizers

- Thiazolidinediones (TZDs) work to lower your blood sugar by increasing the muscle, fat and liver's sensitivity to insulin
- TZDs are referred to as "insulin sensitizers" and also are blood sugar normalizing or euglycemics (drugs that help return the blood sugar to the normal range without the risk of low blood sugars)
- TZDs take a while to begin working (several weeks); so don't stop the pill if you don't notice your blood sugar responding right away
- `https://dtc.ucsf.edu/`

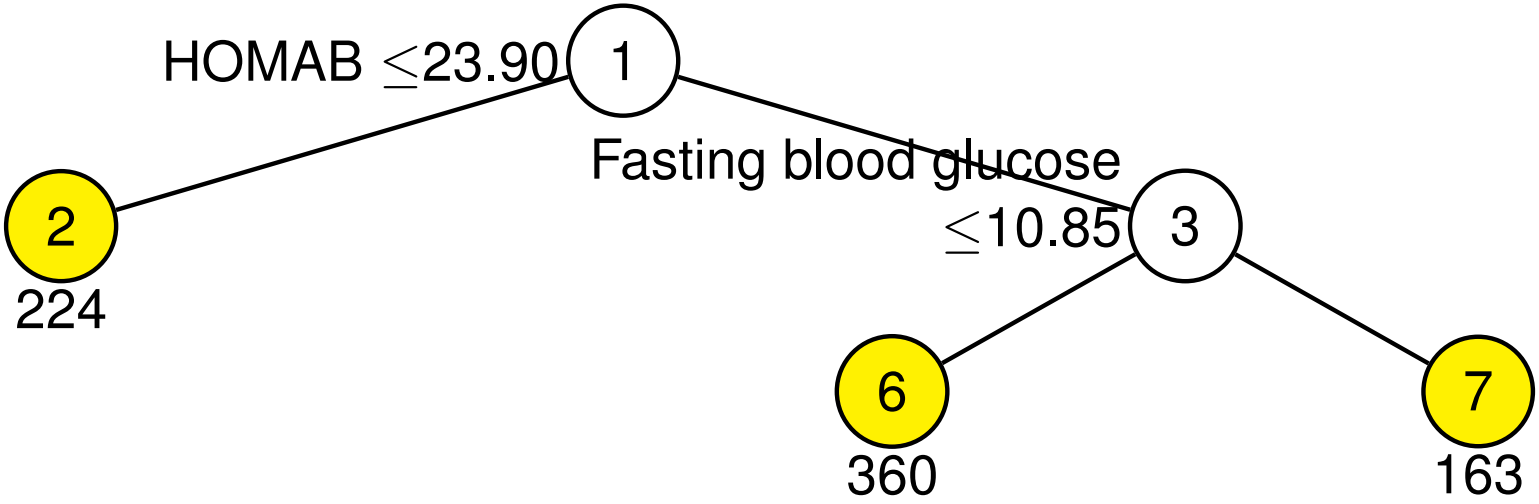
Baseline variables and their missing values

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

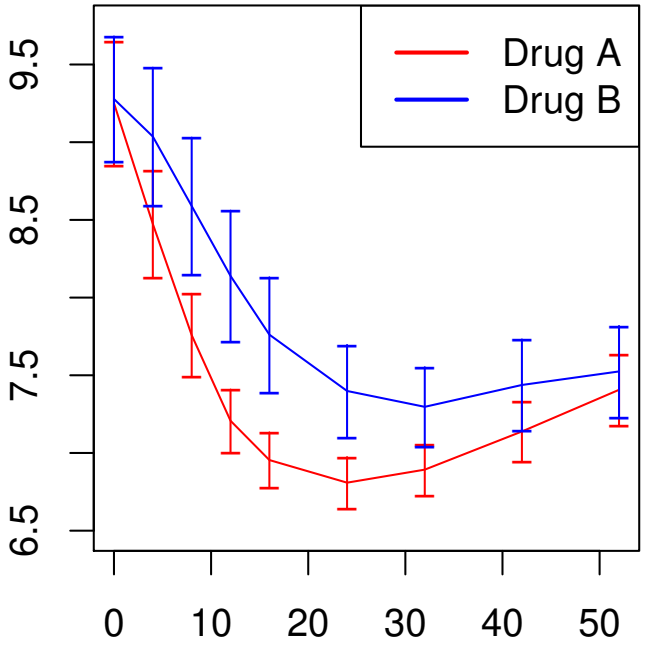
HbA1c means for 747 subjects



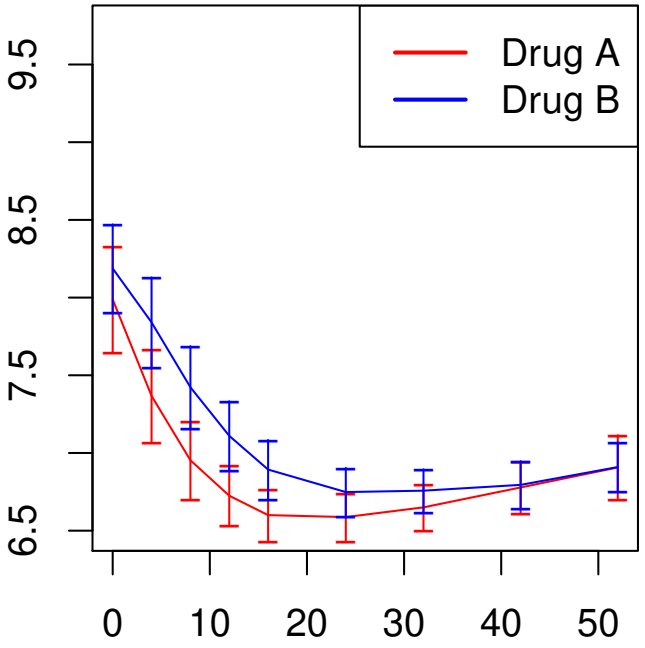
GUIDE tree with 95% bootstrap CIs (Loh et al., 2016)



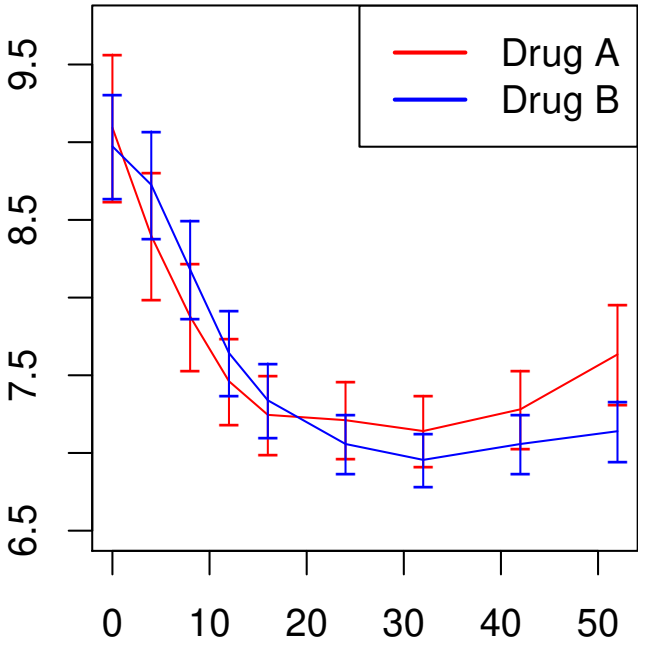
Node 2



Node 6



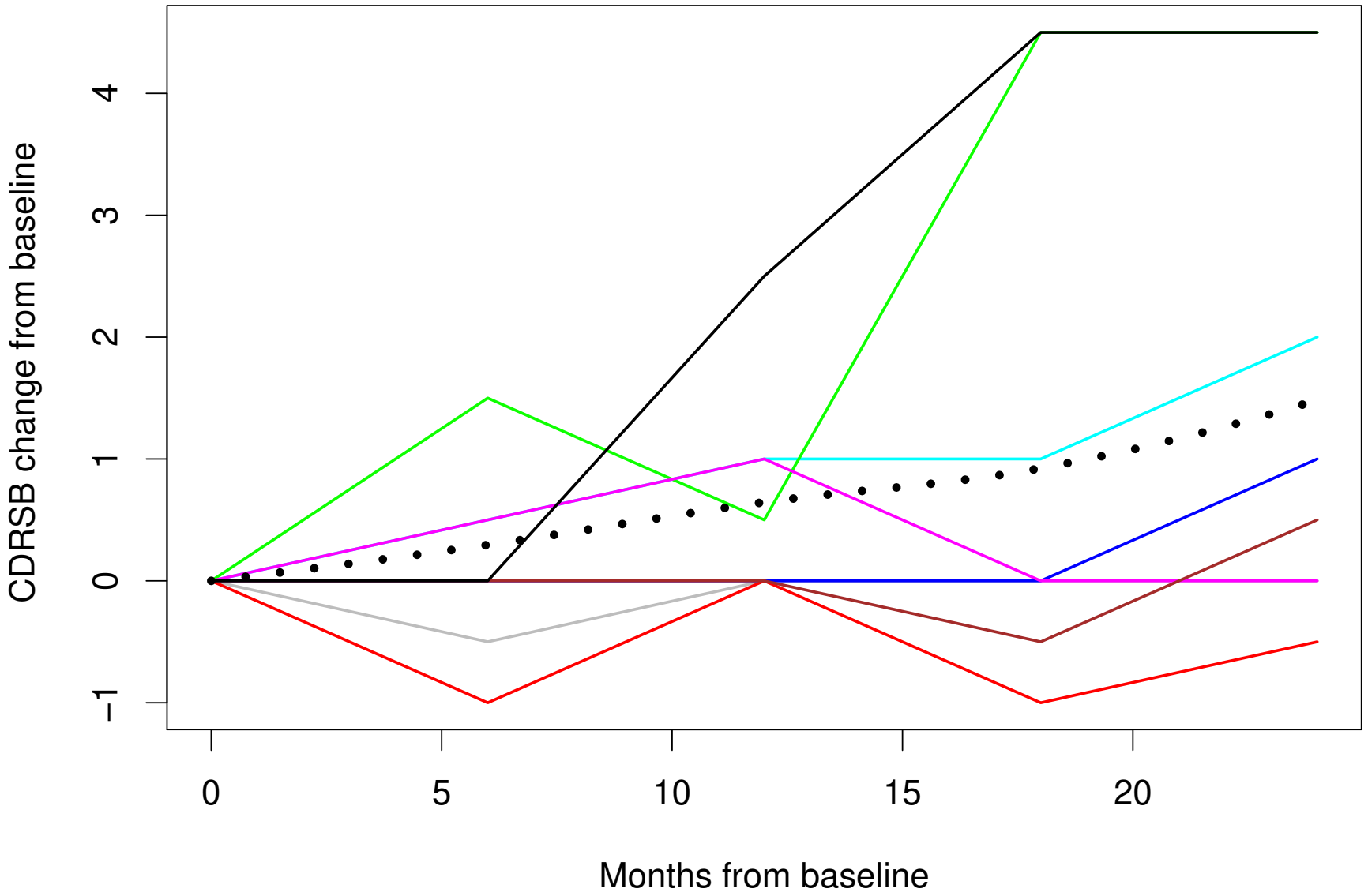
Node 7



Clustering longitudinal responses: Alzheimer's (ADNI) data

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

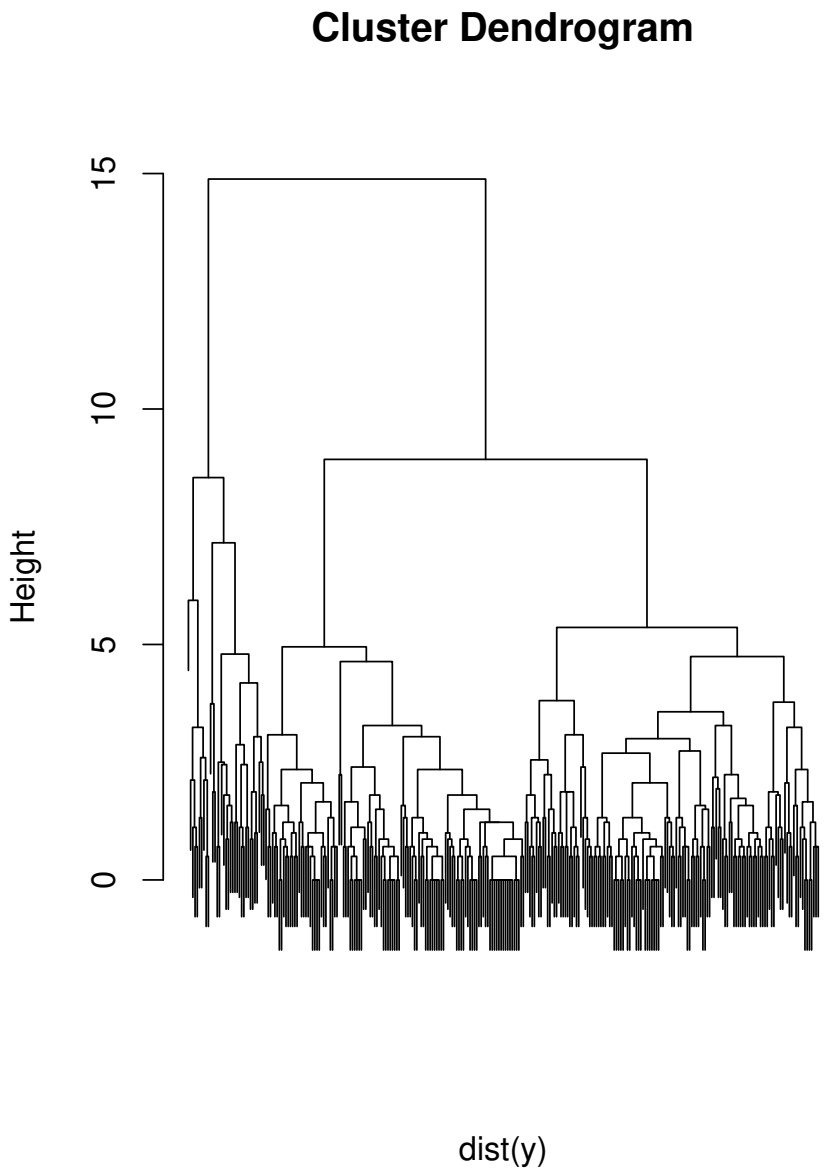
Sample paths with smoothed mean

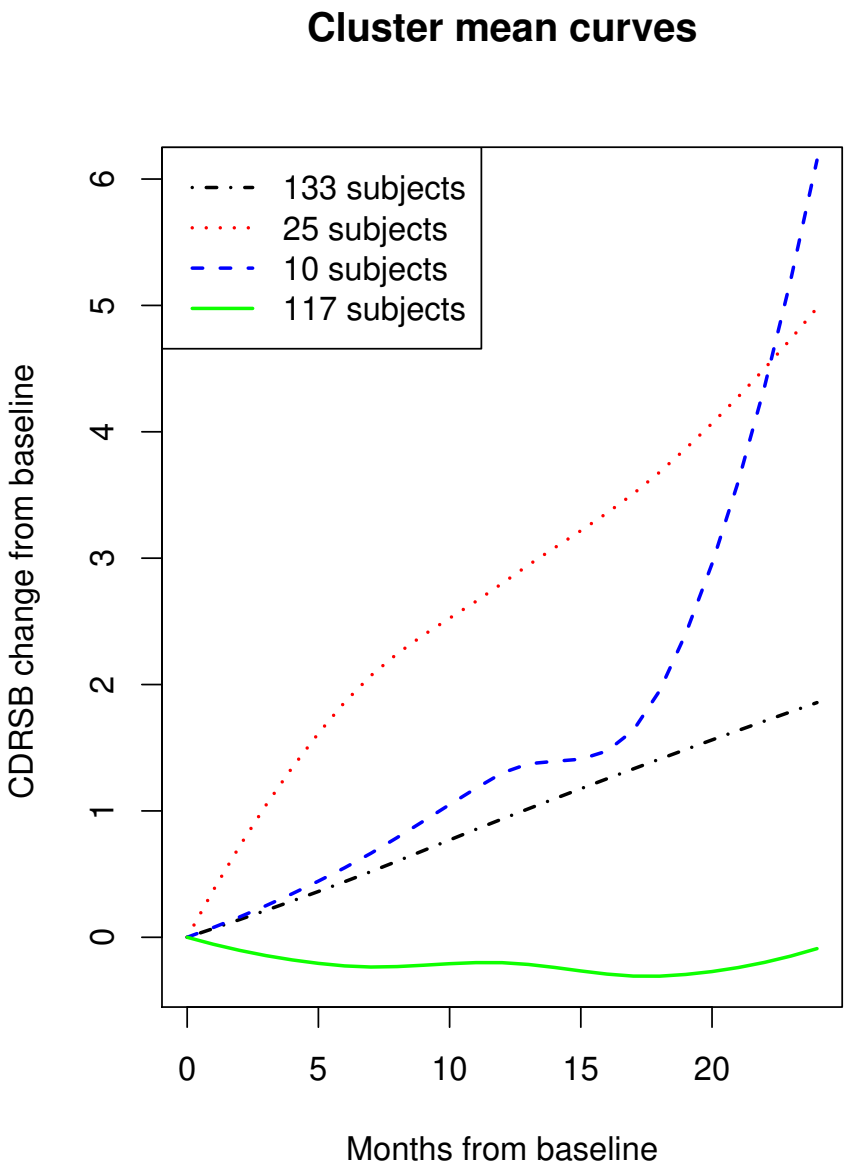


Two methods to group subjects into clusters

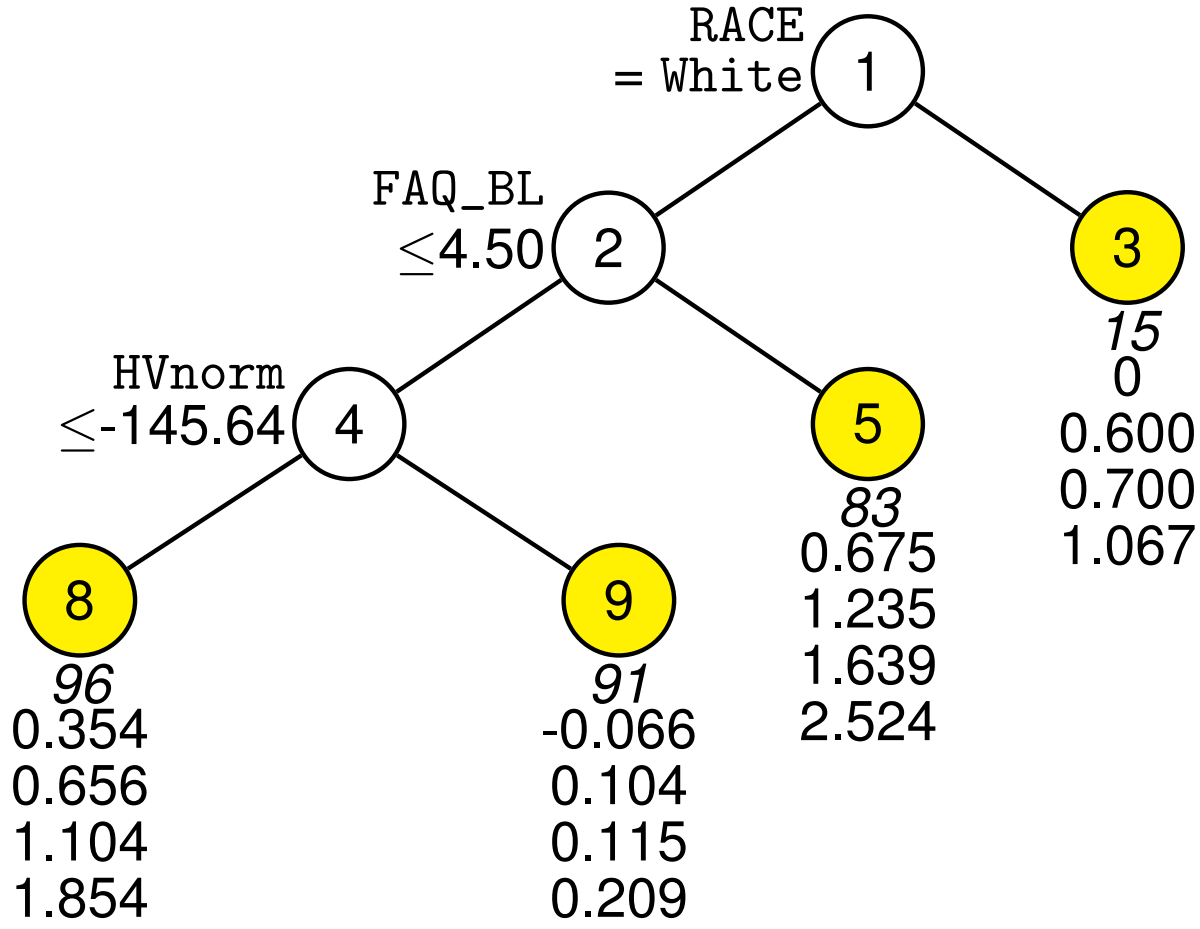
1. Standard clustering methods —
   (a) uses only responses but not covariates
   (b) requires pre-specification of number of clusters
   (c) does not show relationship of clusters to covariates — ANOVA or chi-squared tests typically used to test for associations with covariates
2. GUIDE regression tree —
   (a) uses responses and covariates
   (b) uses cross-validation to determine number of clusters
   (c) cluster-covariate associations obtained directly — no need for ANOVA or chi-squared tests

Hierarchical clustering for 285 completers



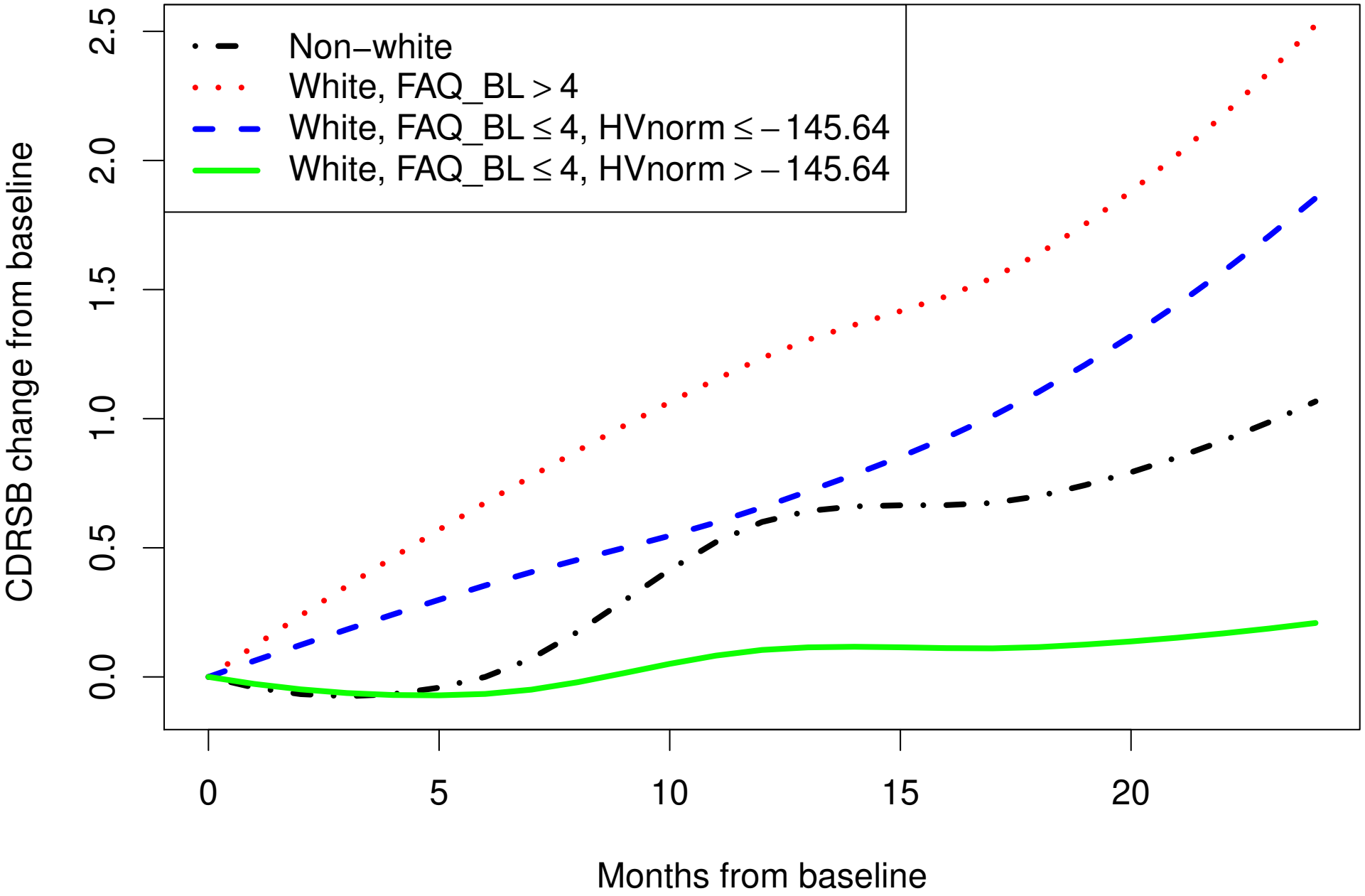


Subgroups for change in CDRSB from baseline (Loh and Zheng, 2013)



Sample size beside node; CDRSB change at 6, 12, 18 and 24 mths below;
FAQ_BL = Functional Activities Questionnaire at baseline (lower is better)
HVnorm = normalized Hippocampal volume (higher is better)

Subgroup mean paths



Nonrandomized treatment: Right heart catheterization (RHC)

- Doctors believe that direct measurement of cardiac function by right heart catheterization (RHC) for some critically ill patients yields better outcomes
- Relative risk of death is higher in elderly and patients with acute myocardial infarction who received RHC
- Benefit of RHC has not been demonstrated in a randomized clinical trial, because physicians refuse to allow their patients to be randomized
- 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 `t3d30` (censored 30-day survival time) and `survtime` (days till death or last contact)

Demographics & outcomes [#missing in brackets]

swang1	Right heart catheterization (RHC) [0]
age	Age in years [0]
sex	Sex (female/male) [0]
wtkilo1	Weight in kilograms [515]
edu	Years of Education [0]
race	Race [0]
income	Income bracket ($<$11k, 11–25k, 25–50k, $>$50k) [0]
ninsclas	Medical insurance (Medicaid, Medicare, Medicare & Medicaid, no insurance, private, private & Medicare) [0]
t3d30	Days from admission to death within 30 days [0]
dth30	Death indicator for t3d30 (0=no, 1=yes) [0]
survtime	Days from admission to death or last contact day [0]
death	Death indicator for survtime (0=no, 1=yes) [0]
transhx	Transfer ($>$ 24 hours) from another hospital (no/yes) [0]

Disease variables [#missing in brackets]

cat1	Primary disease category (9 levels) [0]
cat2	Secondary disease category (6 levels) [2798]
ca	Cancer (3 levels) [0]
card	Cardiovascular diagnosis [0]
gastr	Gastrointestinal diagnosis [0]
hema	Hematologic diagnosis [0]
meta	Metabolic diagnosis [0]
neuro	Neurological diagnosis [0]
ortho	Orthopedic diagnosis [0]
renal	Renal diagnosis [0]
resp	Respiratory diagnosis [0]
seps	Sepsis diagnosis [0]
trauma	Trauma diagnosis [0]

Medical history [#missing in brackets]

amihx	Definite myocardial infarction (no/yes) [0]
cardiohx	Acute MI, peripheral vascular disease, severe cardiovascular symptoms [0]
chfhx	Congestive heart failure (no/yes) [0]
chrpulhx	Chronic or severe pulmonary disease (no/yes) [0]
dementhx	Dementia, stroke or cerebral infarction, Parkinson's disease (no/yes) [0]
gibledhx	Upper GI bleeding (no/yes) [0]
liverhx	Cirrhosis, hepatic failure (no/yes) [0]
malighx	Solid tumor, metastatic disease, chronic leukemia/myeloma, acute leukemia, lymphoma (no/yes) [0]
immunhx	Immunosuppression, organ transplant, HIV positivity, diabetes mellitus, connective tissue disease(no/yes) [0]
psychhx	Psychiatric history, active psychosis or severe depression (no/yes) [0]
renalhx	Chronic renal disease, chronic hemodialysis or peritoneal dialysis (no/yes) [0]

Admission variables [#missing in brackets]

alb1	Albumin [0]
bili1	Bilirubin [0]
crea1	Serum creatinine [0]
hema1	Hematocrit [0]
hrt1	Heart rate [159]
meanbp1	Mean blood pressure [80]
pot1	Serum potassium [0]
pafi1	PaO2/(0.01*FiO2) [0]
paco21	Partial pressure of arterial carbon dioxide [0]
ph1	Serum ph [0]
resp1	Respiration rate [136]
scoma1	Glasgow coma score [0]
sod1	Serum sodium [0]
temp1	Temperature (Celsius) [0]
urin1	Urine output [3028]
wblc1	White blood cell count [0]

PaO2 is partial pressure of arterial oxygen; FiO2 is fraction of inspired oxygen

Admission variables (cont'd.)

aps1	APACHE III score ignoring coma [0]
adld3p	Katz Activities of Daily Living Scale [3016]
das2d3pc	DASI (Duke Activity Status Index) [0]
dnr1	DNR (do-not-resuscitate) status [0]
surv2md1	Estimated probability of 2-month survival [0]

Potential outcome model

- Suppose each subject can be exposed to two alternative “treatments”, e.g., without RHC vs with RHC
- Potential outcomes of subject i are $Y_i(0)$ and $Y_i(1)$, $i = 1, 2, \ldots, n$, for treatments 0 and 1, resp.
- Subject causal effect is $\{Y_i(1) - Y_i(0)\}$
- Impossible to observe both $Y_i(0)$ and $Y_i(1)$
- Population average treatment effect is ATE = $E\{Y(1) - Y(0)\}$

Necessary assumptions for estimating ATE

- Probabilistic assignment assumption:

$$0 < P(T_i = 1 \,|\, X_i = x) < 1 \text{ for all } x$$

- Unconfoundedness assumption (conditional independence):

$$\{(Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \,|\, X_i = x \text{ for all } x$$

- Let $\mu(t, x) = E(Y_i(t) \,|\, T_i = t, X_i = x)$ and $\hat{\mu}(t, x)$ be its estimate
- Regression-based estimator of ATE is

$$n^{-1} \sum_{i=1}^{n} \{\hat{\mu}(1, X_i) - \hat{\mu}(0, X_i)\}$$

Methods

Matching. Group treatment and control subjects so that they are similar w.r.t. all X variables—inefficient because unmatched data are discarded

Propensity score. Let $\pi(X_i) = P(T_i = 1 \,|\, X_i)$. Then $T_i \perp\!\!\!\perp X_i \,|\, \pi(X_i)$ and

$$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp T_i \,|\, \pi(X_i)$$

ATE can be estimated by the Horvitz-Thompson formula:

$$n^{-1} \sum_{i=1}^{n} \left\{ \frac{T_i Y_i}{\hat{\pi}(X_i)} - \frac{(1 - T_i) Y_i}{1 - \hat{\pi}(X_i)} \right\}$$

or inverse propensity weighting (IPW) formula:

$$\frac{\sum_{i=1}^{n} T_i Y_i / \hat{\pi}(X_i)}{\sum_{i=1}^{n} T_i / \hat{\pi}(X_i)} - \frac{\sum_{i=1}^{n} (1 - T_i) Y_i / (1 - \hat{\pi}(X_i))}{\sum_{i=1}^{n} (1 - T_i) / (1 - \hat{\pi}(X_i))}$$

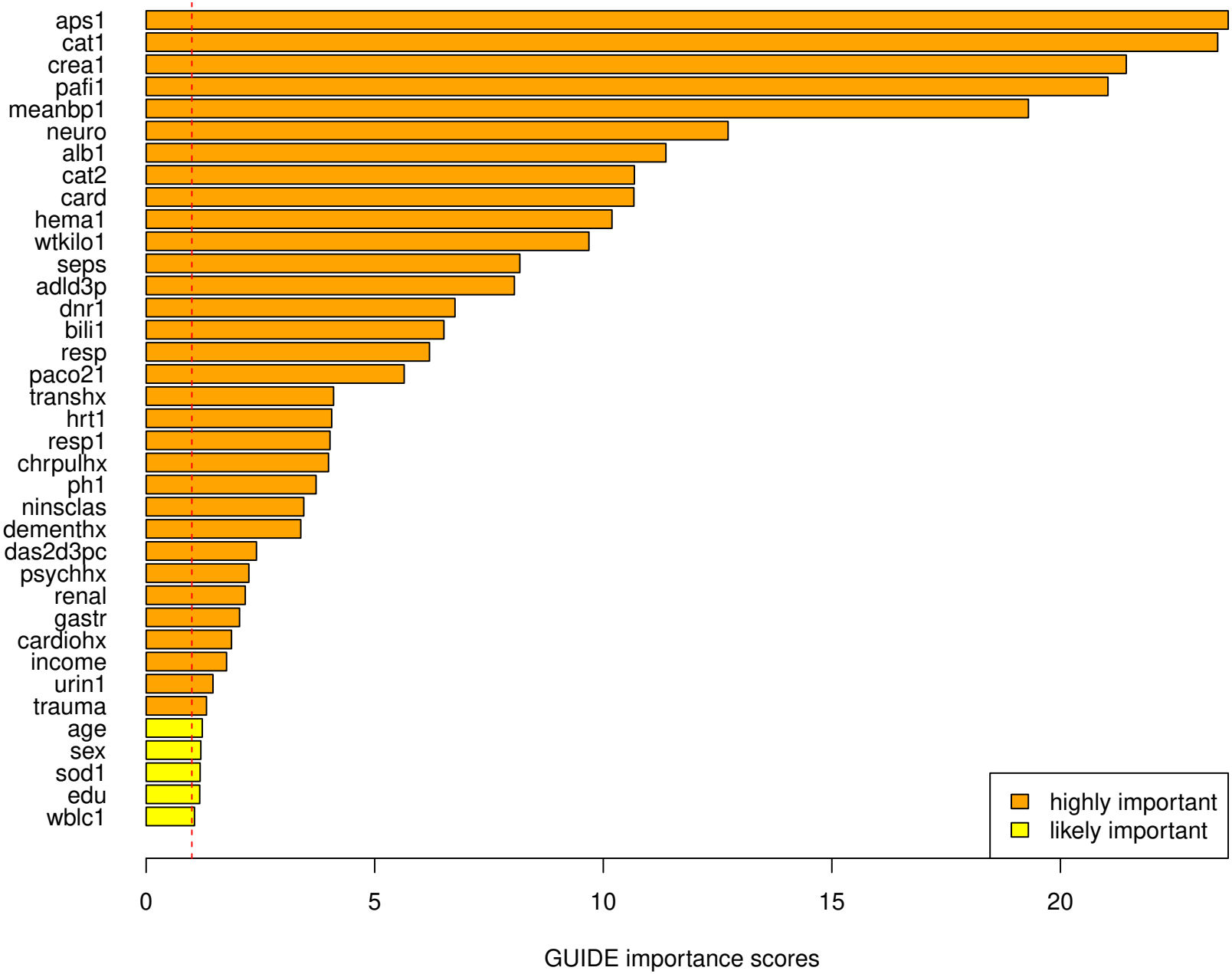
Difficulties with propensity score method

1. $\pi(X_i)$ is unknown and must be estimated (often using logistic regression)
2. Method may be sensitive to misspecification of model for $\pi(X_i)$
3. On one hand, "unmeasured confounder" assumption should be easier to satisfy if number (p) of X variables is large
4. On the other hand, logistic regression is more likely to encounter computational problems, such as quasi-complete separation, if p is large
5. Logistic regression requires all X variables to be nonmissing, but missing values occur frequently in practice
6. Usual approach: impute missing X values and then apply logistic regression but this requires "missing at random" assumptions on *all* X
7. Better approach: replace logistic regression with nonparametric methods that do not need imputation and can deal with large p

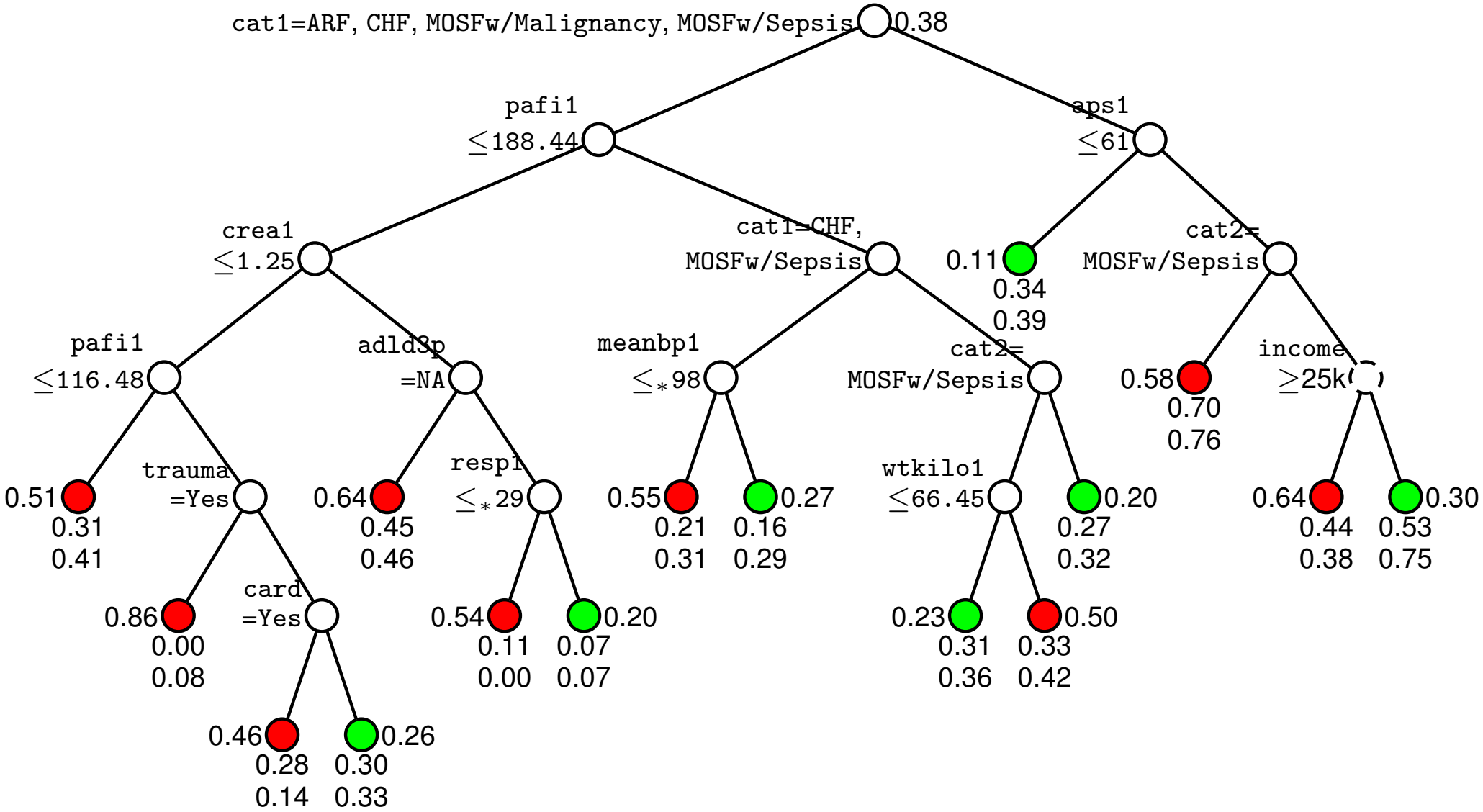
Naïve analysis

	Death rates		Unadjusted
	with RHC	without RHC	treatment effect
Within 30 days	0.381	0.306	0.075
Within 6 months	0.680	0.630	0.050

Important variables for estimating propensity scores

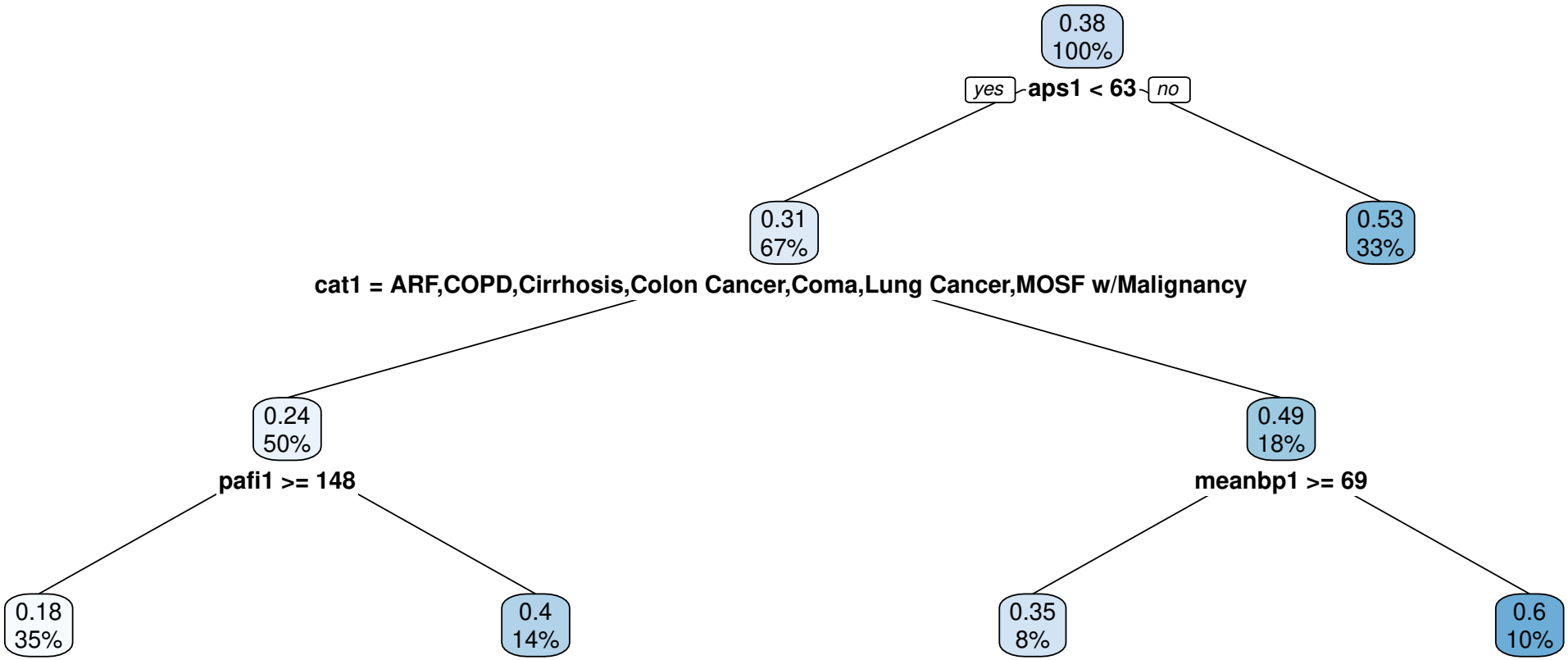


GUIDE propensity score tree for P(RHC)

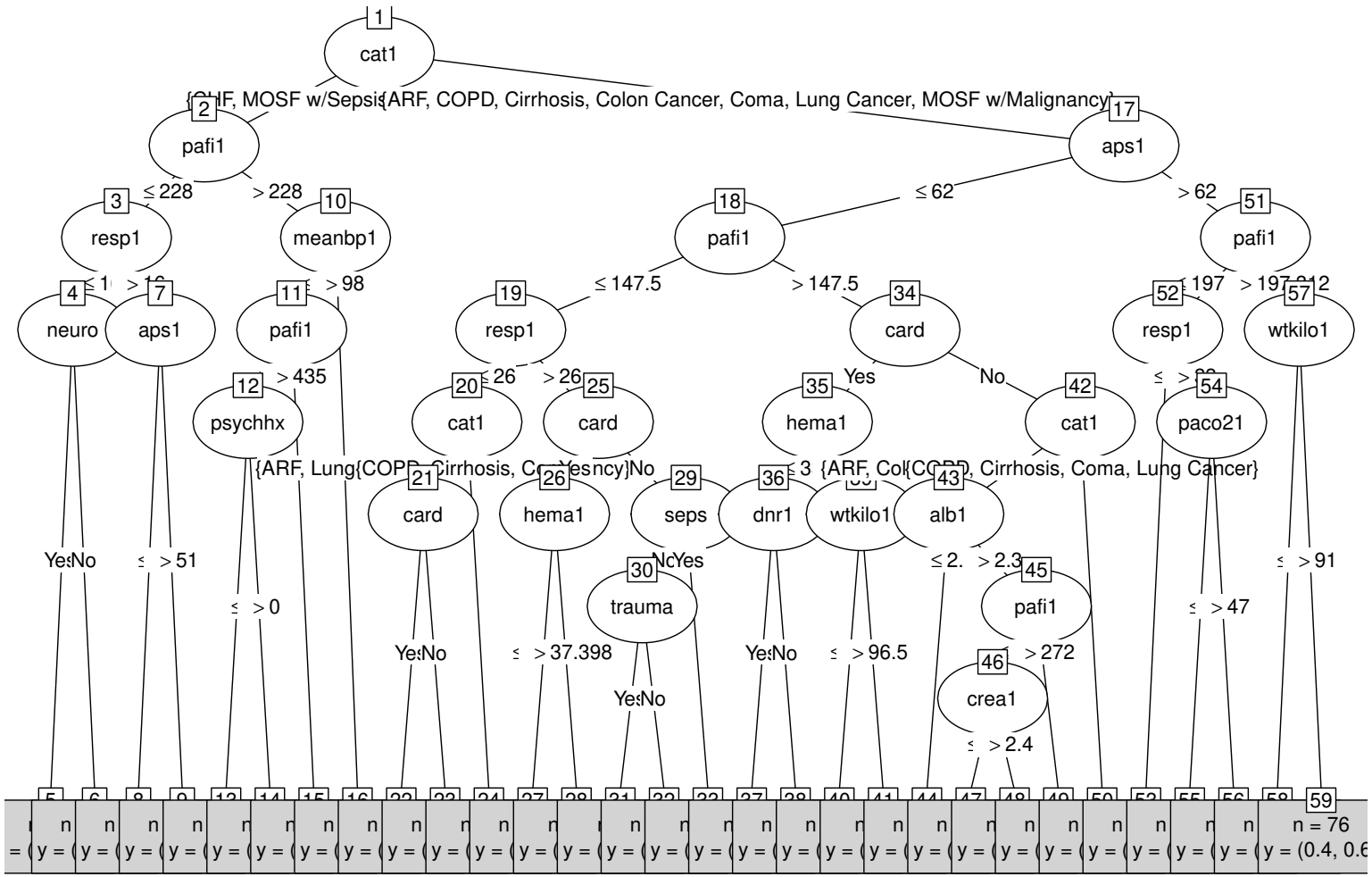


Mortality rates for NoRHC, RHC below and P(RHC) beside nodes; ATE=0.052, 0.042 for P(death) within 1, 6 months

RPART propensity score tree for P(RHC)
ATE = 0.042, 0.024 for P(death) within 1, 6 months



CTREE (party) propensity score tree for P(RHC)
ATE = 0.049, 0.036 for P(death) within 1, 6 months



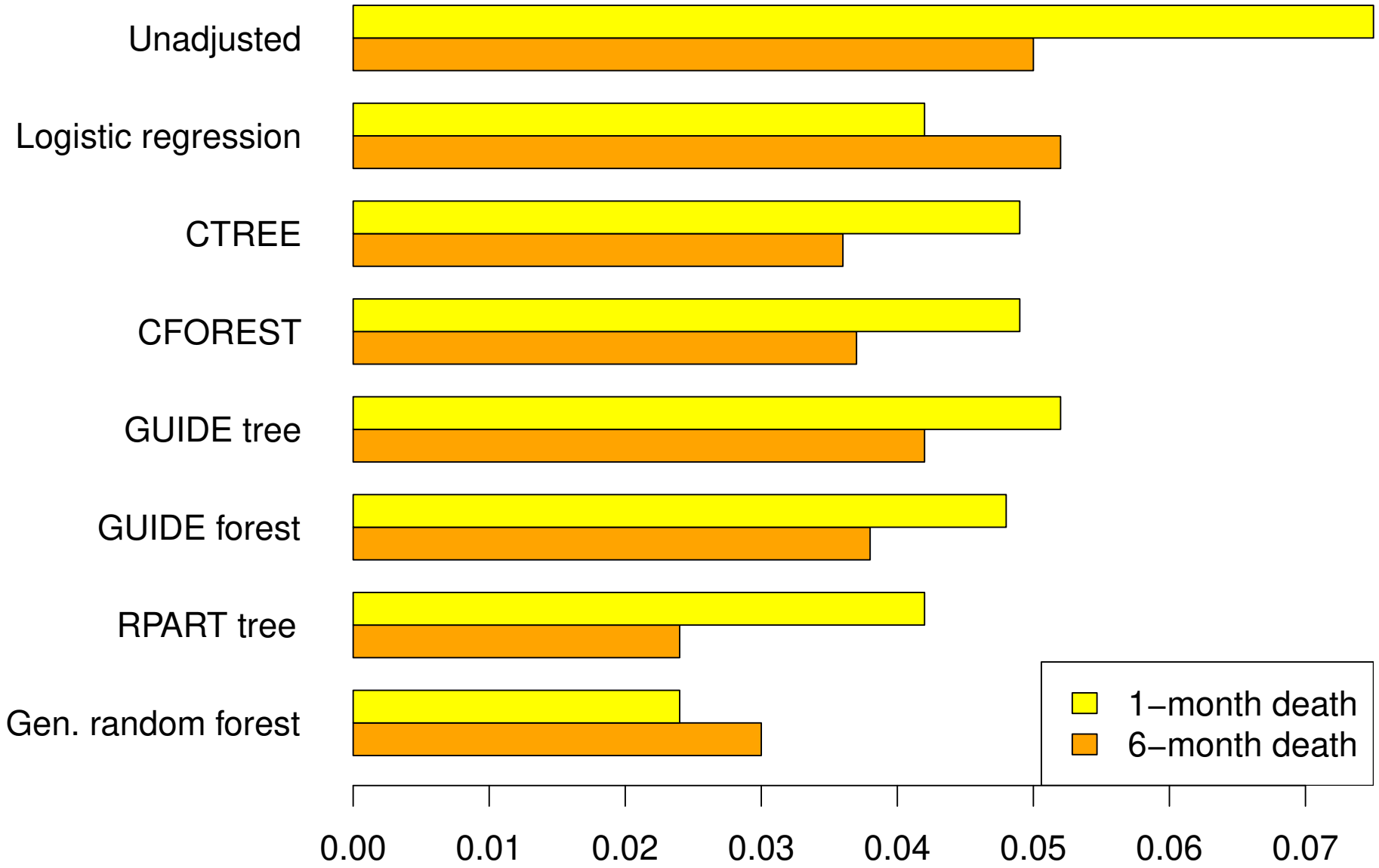
Estimates of average treatment effect for 1-month and 6-month mortality

Method	1-month	6-month
Unadjusted	0.075	0.050
GUIDE tree	0.052	0.042
CTREE tree	0.049	0.036
CFOREST	0.049	0.037
GUIDE forest	0.048	0.038
Logistic regression[a]	0.042	0.052
RPART tree	0.042	0.024
Generalized random forest[b]	0.024	0.030

[a]Missing values imputed with means and missingness dummy variables added

[b]GRF does not allow categorical variables; they are dummy-coded here (Athey et al., 2019)

Estimates of average treatment effect



GUIDE variable importance scores

- Grow a tree with four levels of splits
- If X_i is not constant in a node t, let $W(X_i, t)$ be the Wilson-Hilferty 1-df chi-squared value of X_i at t
- If X_i is constant at t, define $W(X_i, t) = 0$
- Define the unadjusted importance 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-adjusted importance scores

- 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
  3. Let $v_b^*(X_i)$ be the unadjusted score of X_i
- Define $\bar{v}(X_i) = B^{-1} \sum_b v_b^*(X_i)$
- Bias-adjusted importance score of X_i is

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

$(1-\alpha)$-confidence level threshold

- Define X_i as "unimportant" if it is independent of Y
- Define X_i as "important" if it is not "unimportant"
- To find a cut-off score for the important variables,
  1. Randomly permute B times (default is $B = 300$) the Y values, holding the X values fixed
  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,

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

- X_i is important at confidence level $(1-\alpha)$ if $\texttt{IMP}(X_i) > u^*(\alpha)$
- Loh and Zhou (2021) gives details and comparisons with 11 other methods

Concluding remarks: interpretability & missing data

Model interpretability

- Machine learning methods are often criticized for being hard to interpret
- Linear and logistic regression are considered easy to interpret, but regression coefficients are easy to misinterpret if collinearity is present

Missing data

Methods not designed for missing data need data trimming or imputation

- Multiple imputation (Rubin, 1987) involves iteratively fitting a model to every X variable with missing values—a harder task than fitting a model to the original Y variable
- Imputation requires unverifiable assumptions on model structure and missingness mechanism (e.g., MAR) for each predictor variable

Concluding remarks: model stability

- A method is “unstable” if a small change in the data can cause large changes in the model
- Popular myth: “Trees are unstable but forests are not” (Breiman, 1996)
- Truth: all models are unstable to varying degrees
  - Linear and logistic coefficients change if any observation is changed
  - Tree models are robust against extreme outliers because each one affects only one terminal node and no observation has high leverage
  - Random forest and methods that require a random seed (such as LASSO) are inherently unstable because they are randomized—their predictions change with the random seed even if data are unchanged

Concluding remarks: GUIDE properties

1. No parametric model assumptions
2. No problems with collinearity or large numbers of variables
3. Tree selected by cross-validation estimates of prediction error
4. No missing-value imputation nor assumptions on missing-value mechanisms
5. Multiple missing-value codes allowed
6. Importance scores and thresholds available for variable selection
7. Tree model straightforward to interpret but it is just one model
   - Other models available, e.g., split root node on 2nd best X, 3rd best X, etc., and linear, nearest-neighbor or kernel discriminant models in nodes
   - Random forests of pruned and unpruned GUIDE trees also available

Concluding remarks: comparative reviews

1. Brief introductory survey: Loh (2011)
2. Detailed historical review: Loh (2014)
3. GUIDE for subgroup identification: Loh et al. (2019a); Loh and Zhou (2020)
4. GUIDE compared with missing data methods: Loh et al. (2019b, 2020)
5. GUIDE compared with 11 other importance score methods: Loh and Zhou (2021)

References

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *Annals of Statistics*, 47:1148–1178.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24:2350–2383.

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.

Harrison, S. L., Fazio-Eynullayeva, E., Lane, D. A., Underhill, P., and Lip, G. Y. H. (2020). Comorbidities associated with mortality in 31,461 adults with COVID-19 in the United States: A federated electronic medical record analysis. *PLOS Medicine*, 17(9):1–11.

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

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

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

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., Cao, L., and Zhou, P. (2019a). Subgroup identification for precision medicine: a comparative review of thirteen methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(5):e1326.

Loh, W.-Y., Eltinge, J., Cho, M. J., and Li, Y. (2019b). 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. (2019c). 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., Zhang, Q., Zhang, W., and Zhou, P. (2020). Missing data, imputation and regression trees. *Statistica Sinica*, 30:1697–1722.

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

Loh, W.-Y. and Zhou, P. (2020). The GUIDE approach to subgroup identification. In Ting, N., Cappelleri, J. C., Ho, S., and Chen, D.-G., editors, *Design and analysis of Subgroups with Biopharmaceutical Applications*, Emerging Topics in Statistics and Biostatistics. Springer. In press.

Loh, W.-Y. and Zhou, P. (2021). Variable importance scores. *Journal of Data Science*, 19(4):569–592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York, NY.

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.

Therneau, T., Atkinson, B., and Ripley, B. (2019). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-15.

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.