Recursive Partitioning and Applications

Heping Zhang

Department of Epidemiology and Public Health
Yale University School of Medicine

June–July, 2011

This PDF slides cover Chapters 1–4, 6, 8, and 9 in the book of Heping Zhang and Burton Singer entitled “Recursive Partitioning and Applications" published by Springer in 2010. The instructors and students are assumed to have some acess to the book.

Outline

Regression Model

- Many scientific problems reduce to modeling the relationship between two sets of variables. Regression methodology is designed to quantify these relationships.
  - linear regression for continuous data
  - logistic regression for binary data
  - proportional hazard regression for censored survival data
  - mixed-effect regression for longitudinal data
- These parametric (or semiparametric) regression methods may not lead to faithful data descriptions when the underlying assumptions are not satisfied.

Recursive Partitioning Based Methods

- Nonparametric regression has evolved to relax or remove the restrictive assumptions.
- Recursive partitioning provides a useful alternative to the parametric regression methods.
  - Classification and Regression Trees (CART)
  - Multivariate Adaptive Regression Splines (MARS)
  - Forest
  - Survival Trees

Areas of Applications

- financial firms
  - banking crises (Cashin and Duttagupta 2008)
  - credit cards (Altman 2002; Frydman, Altman and Kao 2002; Kumar and Ravi 2008)
  - investments (Pace 1995 and Brennan, Parameswaran et al. 2001)
- manufacturing and marketing companies (Levin, Zahavi, and Olitsky 1995; Chen and Su 2008)
- pharmaceutical industries (Chen et al. 1998)
- engineering research
  - natural language speech recognition (Bahl et al. 1989)
  - musical sounds (Wieczorkowska 1999)
  - text recognition (Desilva and Hull 1994)
  - tracking roads in satellite images (Geman and Jedynak 1996)

Areas of Applications

- astronomy (Owens, Griffiths, and Ratnatunga 1996)
- computers and the humanities (Shmulevich et al. 2001)
- chemistry (Chen, Rusinko, and Young 1998)
- environmental entomology (Hebertson and Jenkins 2008)
- forensics (Appavu and Rajaram 2008)
- polar biology (Terhune et al. 2008).

Biomedical Applications

- Is this patient with chest pain suffering a heart attack?
- Does he simply have a strained muscle?

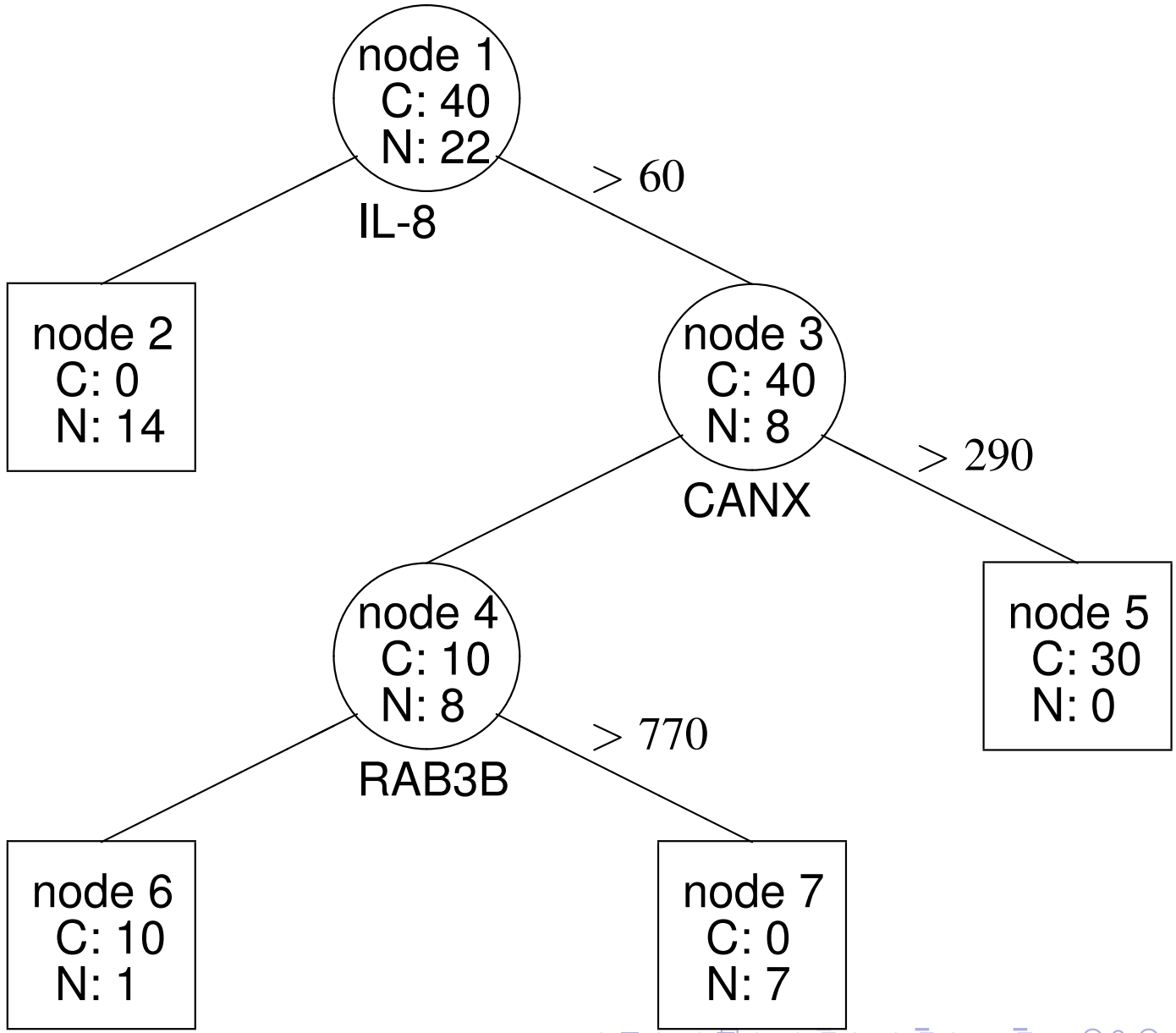
Examples

- Chest Pain
  - Goldman et al. (1982, 1996): Build an expert computer system that could assist physicians in emergency rooms to classify patients with chest pain into relatively homogeneous groups within a few hours of admission using the clinical factors available.
  - The authors included 10,682 patients with acute chest pain in the derivation data set and 4,676 in the validation data set.
- Coma
  - Levy et al. (1985): Predict the outcome from coma caused by cerebral hypoxia-ischemia
  - they studied 210 patients with cerebral hypoxia-ischemia and considered 13 factors including age, sex, verbal and motor responses, and eye opening movement.

Gene Expression

- Zhang et al. (2001) analyzed a data set from the expression profiles
- 2,000 genes in 22 normal and 40 colon cancer tissues (Alon et al. 1999).

Gene Expression



Statistical Problem

- an outcome variable, Y, and a set of p predictors, $x_1, \ldots, x_p$.
- establish a relationship between Y and the x's
- $\boldsymbol{P}\{Y = y \,|\, x_1, \ldots, x_p\}$,
- parametric models
  - $$\frac{\exp(\beta_0 + \sum_{i=1}^{p} \beta_i x_i)}{1 + \exp(\beta_0 + \sum_{i=1}^{p} \beta_i x_i)},$$
  - $$\frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right], \mu = \beta_0 + \sum_{i=1}^{p} \beta_i x_i.$$

Statistical Problem

Table: Correspondence Between the Uses of Classic Approaches and Recursive Partitioning Technique in This Book

Type of response	Parametric methods	Recursive partitioning technique
Continuous	Ordinary linear regression	Regression trees and adaptive splines
Binary	Logistic regression	Classification trees and forests
Censored	Proportion hazard regression	Survival trees
Longitudinal	Mixed-effects models	Regression trees and adaptive splines
Multiple discrete	Exponential, marginal, and frailty models	Classification trees

Yale Pregnancy Outcome Study

- PI: Dr. Michael B. Bracken at Yale University.
- Population: women who made a first prenatal visit to a private obstetrics or midwife practice, health maintenance organization, or hospital clinic in the greater New Haven, Connecticut, area, between May 12, 1980, and March 12, 1982, and who anticipated delivery at the Yale–New Haven Hospital.
- Sample size: a subset of 3,861 women whose pregnancies ended in a singleton live birth.
- Outcome: preterm delivery

Yale Pregnancy Outcome Study

Table: A List of Candidate Predictor Variables

Variable name	Label	Type	Range/levels
Maternal age	x_1	Continuous	13–46
Marital status	x_2	Nominal	Currently married, divorced, separated, widowed, never married
Race	x_3	Nominal	White, Black, Hispanic, Asian, others
Marijuana use	x_4	Nominal	Yes, no
Times of using marijuana	x_5	Ordinal	$>= 5$, 3–4, 2, 1 (daily), 4–6, 1–3 (weekly), 2–3, 1, < 1 (monthly)
Years of education	x_6	Continuous	4–27
Employment	x_7	Nominal	Yes, no
Smoker	x_8	Nominal	Yes, no
Cigarettes smoked	x_9	Continuous	0–66
Passive smoking	x_{10}	Nominal	Yes, no
Gravidity	x_{11}	Ordinal	1–10
Hormones/DES used by mother	x_{12}	Nominal	None, hormones, DES, both, uncertain
Alcohol (oz/day)	x_{13}	Ordinal	0–3
Caffeine (mg)	x_{14}	Continuous	12.6–1273
Parity	x_{15}	Ordinal	0–7

The Elements of Tree

- root node: the circle on the top.
- internal node
- terminal nodes
- left and right daughter nodes
- offspring nodes
- split

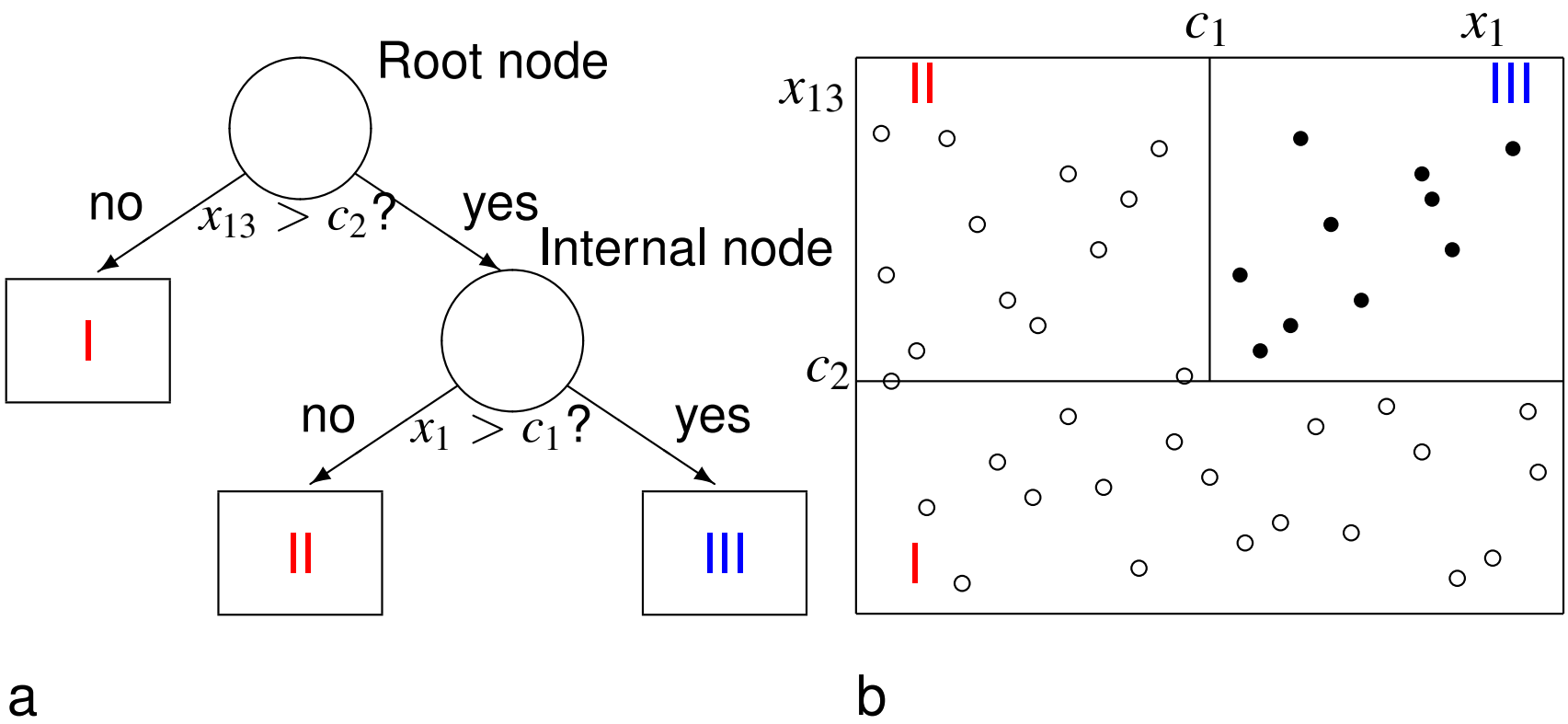
Interpretation of Tree

- What are the contents of the nodes?
- Why and how is a parent node split into two daughter nodes?
- When do we declare a terminal node?

Interpretation of Tree

- The root node contains a sample of subjects from which the tree is grown–learning sample.
- The root node contains all 3,861 pregnant women.
- All nodes in the same layer constitute a partition of the root node.
- Every node in a tree is merely a subset of the learning sample.

An Example of Tree



Aim of Recursive Partitioning

- Produce the terminal nodes that are homogeneous
- They contain either dots or circles

Splitting a Node

- Consider the variable x_1 (age)
- 32 distinct age values in the range of 13 to 46
- 32-1=31 allowable splits
- For an ordinal predictor, the number of allowable splits is one fewer than the number of its distinctly observed values.

Allowable Splits

Table: Race

Left daughter node	Right daughter node
White	Black, Hispanic, Asian, others
Black	White, Hispanic, Asian, others
Hispanic	White, Black, Asian, others
Asian	White, Black, Hispanic, others
White, Black	Hispanic, Asian, others
White, Hispanic	Black, Asian, others
White, Asian	Black, Hispanic, others
Black, Hispanic	White, Asian, others
Black, Asian	White, Hispanic, others
Hispanic, Asian	White, Black, others
Black, Hispanic, Asian	White, others
White, Hispanic, Asian	Black, others
White, Black, Asian	Hispanic, others
White, Black, Hispanic	Asian, others
White, Black, Hispanic, Asian	Others

Splits of a Nominal Variable

- Race has 5 levels
- $2^{5-1} - 1 = 15$ allowable splits
- any nominal variable that has k levels contributes $2^{k-1} - 1$ allowable splits

Allowable Splits

- The 15 predictors yield 347 possible splits
- How do we select one or several preferred splits from the pool of allowable splits?
- We need to define a selection criterion

Goodness of Split

The goodness of a split must weigh the homogeneities (or the impurities) in the two daughter nodes.
Consider the question "Is $x_1 > c$?"

		Term	Preterm	
Left Node (τ_L)	$x_1 \leq c$	n_{11}	n_{12}	$n_{1.}$
Right Node (τ_R)	$x_1 > c$	n_{21}	n_{22}	$n_{2.}$
		$n_{.1}$	$n_{.2}$	

Entropy Impurity

- Left node

$$i(\tau_L) = -\frac{n_{11}}{n_{1.}} \log\left(\frac{n_{11}}{n_{1.}}\right) - \frac{n_{12}}{n_{1.}} \log\left(\frac{n_{12}}{n_{1.}}\right).$$

- Right node

$$i(\tau_R) = -\frac{n_{21}}{n_{2.}} \log\left(\frac{n_{21}}{n_{2.}}\right) - \frac{n_{22}}{n_{2.}} \log\left(\frac{n_{22}}{n_{2.}}\right).$$

- The goodness of a split, s, is measured by

$$\Delta I(s,\tau) = i(\tau) - \boldsymbol{P}\{\tau_L\} i(\tau_L) - \boldsymbol{P}\{\tau_R\} i(\tau_R).$$

τ is the parent node of τ_L and τ_R. $\boldsymbol{P}\{\tau_L\}$ and $\boldsymbol{P}\{\tau_R\}$ are the proportions of the observations assigned to the left and right daughter nodes, respectively.

Goodness of Split for Age Split at 35

	Term	Preterm	
Left Node (τ_L)	3521	198	3719
Right Node (τ_R)	135	7	142
	3656	205	3861

$$i(\tau_L) = -\frac{3521}{3719}\log(\frac{3521}{3719}) - \frac{198}{3719}\log(\frac{198}{3719}) = 0.2079.$$

$$i(\tau_R) = 0.1964, i(\tau) = 0.20753.$$

$$\Delta I(s,\tau) = 0.00001.$$

The Goodness of Allowable Age Splits

Split value	Impurity Left node	Impurity Right node	$1000\Delta I$
13	0.00000	0.20757	0.01
14	0.00000	0.20793	0.14
15	0.31969	0.20615	0.17
16	0.27331	0.20583	0.13
17	0.27366	0.20455	0.23
18	0.31822	0.19839	1.13
19	0.30738	0.19508	1.40
20	0.28448	0.19450	1.15
21	0.27440	0.19255	1.15
22	0.26616	0.18965	1.22
23	0.25501	0.18871	1.05
24	0.25747	0.18195	1.50
25	0.24160	0.18479	0.92
26	0.23360	0.18431	0.72
27	0.22750	0.18344	0.58
28	0.22109	0.18509	0.37

Split value	Impurity Left node	Impurity Right node	$1000\Delta I$
29	0.21225	0.19679	0.06
30	0.20841	0.20470	0.00
31	0.20339	0.22556	0.09
32	0.20254	0.23871	0.18
33	0.20467	0.23524	0.09
34	0.20823	0.19491	0.01
35	0.20795	0.19644	0.01
36	0.20744	0.21112	0.00
37	0.20878	0.09804	0.18
38	0.20857	0.00000	0.37
39	0.20805	0.00000	0.18
40	0.20781	0.00000	0.10
41	0.20769	0.00000	0.06
42	0.20761	0.00000	0.03
43	0.20757	0.00000	0.01

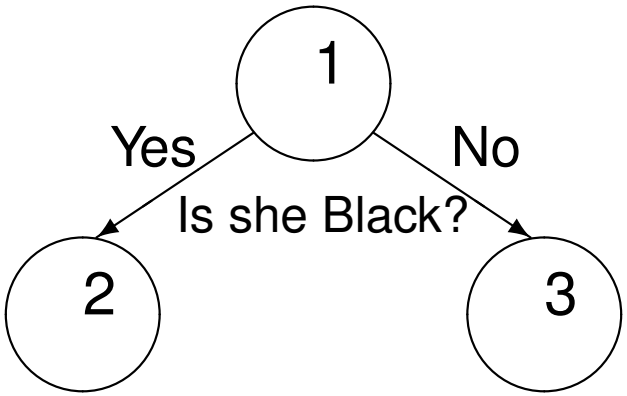
The Largest Goodness of Split from All Predictors

Variable	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
$1000\Delta I$	1.5	2.8	4.0	0.6	0.6	3.2	0.7	0.6
Variable	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	
$1000\Delta I$	0.7	0.2	1.8	1.1	0.5	0.8	1.2	

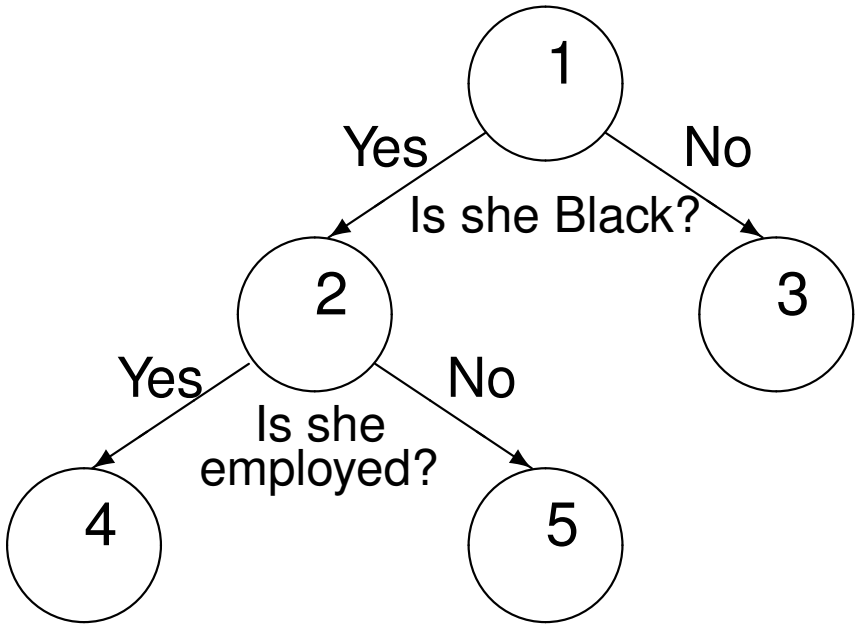
Things to Notice

- The greatest reduction in the impurity comes from the age split at 24.
- What about the age split at age 19, stratifying the study sample into teenagers and adults?
- This best age split is used to compete with the best splits from the other 14 predictors.
- The best of the best comes from the race variable with $1000\Delta I = 4.0$, i.e., $\Delta I = 0.004$.
- This best split divides the root node according to whether a pregnant woman is Black or not.

Top Splits



a



b

Recursive Partitioning

- After splitting the root node, we continue to divide its two daughter nodes.
- The partition of node 2 uses only 710 Black women, and the remaining 3,151 non-Black women are put aside.
- The pool of allowable splits is nearly intact except that race does not contribute any more splits, as everyone is now Black.
- The total number of allowable splits decreases from 347 to at least 332.
- An offspring node may use the same splitting variable as its ancestors.
- The number of allowable splits decreases as the partitioning continues.

Important Issues

- If all candidate variables are equally plausible substantively, then generate separate trees using each of the variables to continue the splitting process.
- If only one or two of the candidate variables is interpretable in the context of the classification problem at hand, then select them for each of two trees to continue the splitting process.

Terminal Nodes

The recursive partitioning process may proceed until the tree is saturated in the sense that the offspring nodes subject to further division cannot be split.

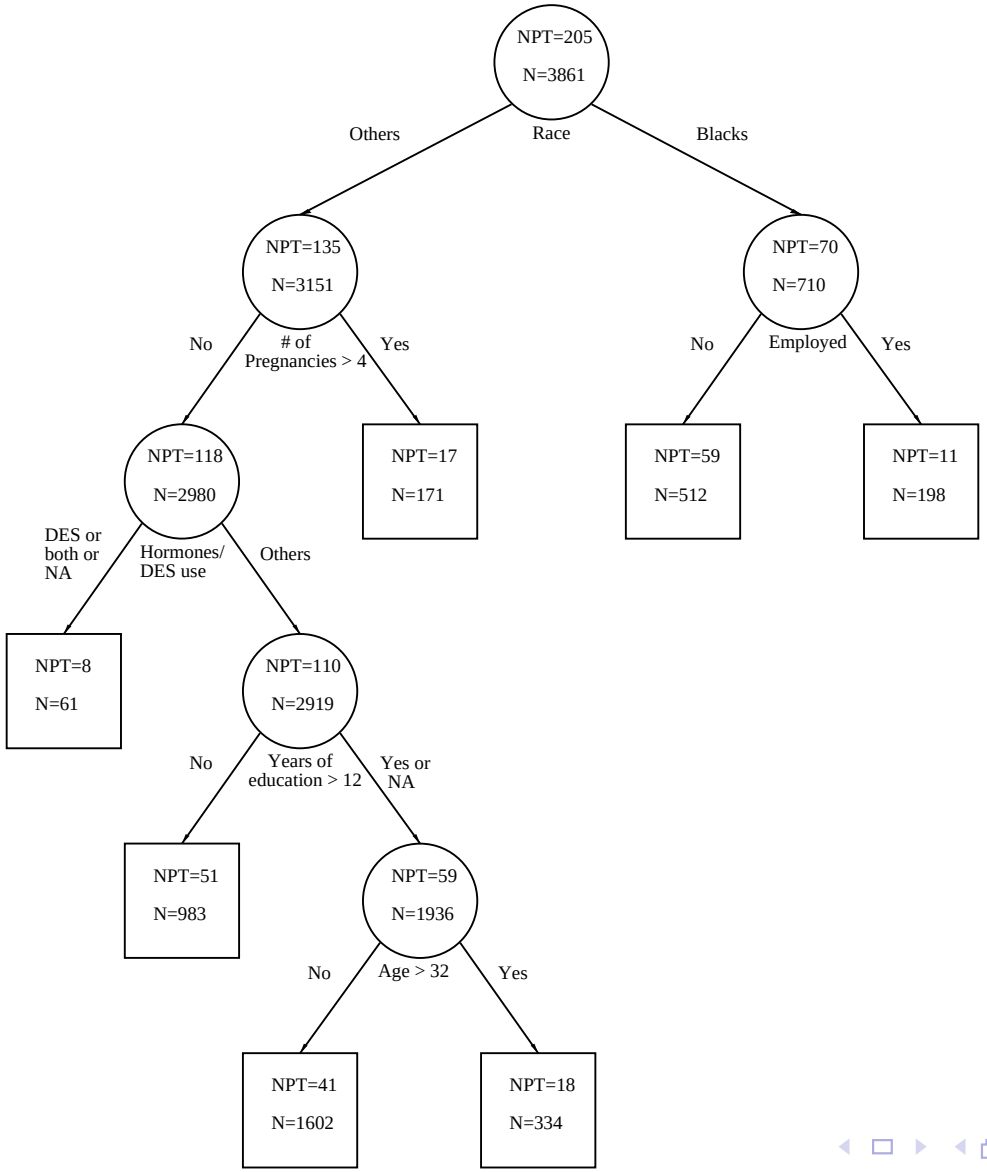
- there is only one subject in a node.
- the total number of allowable splits for a node drops as we move from one layer to the next.
- the number of allowable splits eventually reduces to zero
- the nodes are terminal when they are not divided further

Stopping Rules and Tree Pruning

The saturated tree is usually too large to be useful.

- the terminal nodes are so small that we cannot make sensible statistical inference.
- this level of detail is rarely scientifically interpretable.
- a minimum size of a node is set *a priori*.
- stopping rules
  - Automatic Interaction Detection(AID) (Morgan and Sonquist 1963) declares a terminal node based on the relative merit of its best split to the quality of the root node
- Breiman et al. (1984, p. 37) argued that depending on the stopping threshold, the partitioning tends to end too soon or too late.
- pruning
  - find a subtree of the saturated tree that is most “predictive" of the outcome and least vulnerable to the noise in the data.

The computer-selected tree structure

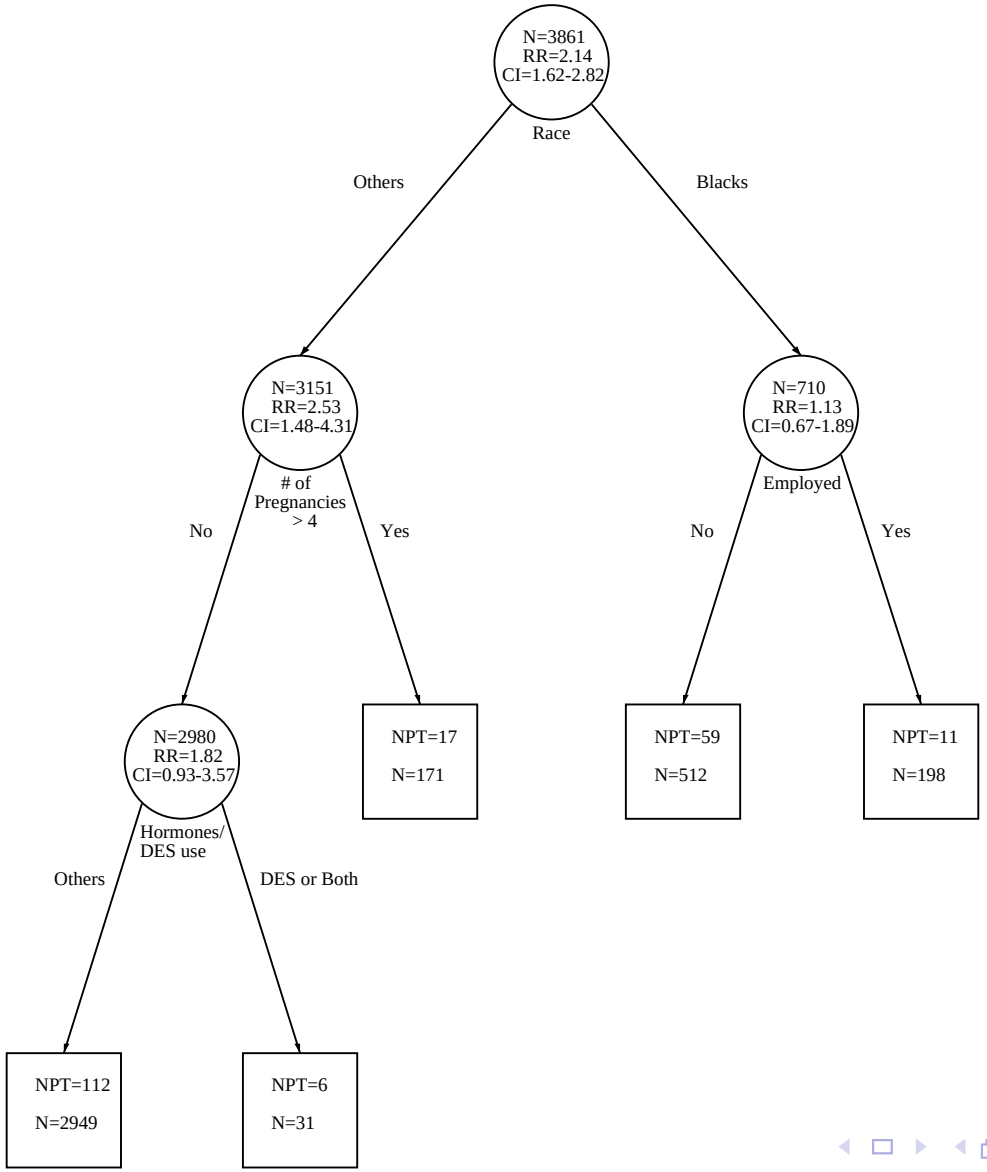


Interpretation

Let us examine the left node in the third layer

- 2,980 non-Black women who had no more than four pregnancies
- The split for this group of women is based on their mothers' use of hormones and/or DES
- If their mothers used hormones and/or DES, or the answers were not reported, they are assigned to the left daughter node.
- The right daughter node consists of those women whose mothers did not use hormones or DES, or who reported uncertainty about their mothers' use.
- Women with the "uncertain" answer and the missing answer are assigned to different sides of the parent node.
- We need to manually change the split.
- Numerically, the goodness of split, Δ, changes from 0.00176 to 0.00148.

Revised tree structure



Risk profile

Who were at risk of preterm delivery?

- non-Black women who have four or fewer prior pregnancies and whose mothers used DES and/or other hormones are at highest risk
- 19.4% of these women have preterm deliveries as opposed to 3.8% whose mothers did not use DES
- among Black women who are also unemployed, 11.5% had preterm deliveries, as opposed to 5.5% among employed Black women
- employment status may just serve as a proxy for more biological circumstances

Regression Model

Logistic regression is a standard approach to the analysis of binary data. For every study subject i we assume that the response Y_i has the Bernoulli distribution

$$P\{Y_i = y_i\} = \theta_i^{y_i}(1-\theta_i)^{1-y_i}, \quad y_i = 0, 1, \ i = 1, \ldots, n,$$

where the parameters

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$$

must be estimated from the data. Here, a prime denotes the transpose of a vector or matrix.

Link Function

To model these data, we generally attempt to reduce the n parameters in $\boldsymbol{\theta}$ to fewer degrees of freedom. The unique feature of logistic regression is to accomplish this by introducing the logit link function:

$$\theta_i = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})},$$

where

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$$

is the new $(p+1)$-vector of parameters to be estimated and $(x_{i1}, \ldots, x_{ip})$ are the values of the p covariates included in the model for the ith subject $(i = 1, \ldots, n)$.

Likelihood

To estimate $\boldsymbol{\beta}$, we make use of the likelihood function

$$
\begin{aligned}
& L(\boldsymbol{\beta}; \mathbf{y}) \\
&= \prod_{i=1}^{n} \left[\frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})} \right]^{y_i} \left[\frac{1}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})} \right]^{1-y_i} \\
&= \frac{\prod_{y_i=1} \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{\prod_{i=1}^{n} [1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})]}.
\end{aligned}
$$

By maximizing $L(\boldsymbol{\beta}; \mathbf{y})$, we obtain the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$.

Odds Ratio

The odds that the ith subject has an abnormal condition is

$$\frac{\theta_i}{1-\theta_i} = \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}).$$

Consider two individuals i and k for whom $x_{i1} = 1$, $x_{k1} = 0$, and $x_{ij} = x_{kj}$ for $j = 2, \ldots, p$. Then, the odds ratio for subjects i and k to be abnormal is

$$\frac{\theta_i/(1-\theta_i)}{\theta_k/(1-\theta_k)} = \exp(\beta_1).$$

Taking the logarithm of both sides, we see that β_1 is the log odds ratio of the response resulting from two such subjects when their first covariate differs by one unit and the other covariates are the same.

Revisit of the Pregnancy Example

Three predictors, x_2 (marital status), x_3 (race), and x_{12} (hormones/DES use), are nominal and have five levels.

Dummy Variable for Marital Status

Let

$$z_1 = \begin{cases} 1 & \text{if a subject was currently married,} \\ 0 & \text{otherwise,} \end{cases}$$

$$z_2 = \begin{cases} 1 & \text{if a subject was divorced,} \\ 0 & \text{otherwise,} \end{cases}$$

$$z_3 = \begin{cases} 1 & \text{if a subject was separated,} \\ 0 & \text{otherwise,} \end{cases}$$

$$z_4 = \begin{cases} 1 & \text{if a subject was widowed,} \\ 0 & \text{otherwise.} \end{cases}$$

Dummy Variable for Race

$$
\begin{aligned}
z_5 &= \begin{cases} 1 & \text{for a Caucasian,} \\ 0 & \text{otherwise,} \end{cases} \\
z_6 &= \begin{cases} 1 & \text{for an African-American,} \\ 0 & \text{otherwise,} \end{cases} \\
z_7 &= \begin{cases} 1 & \text{for a Hispanic,} \\ 0 & \text{otherwise,} \end{cases} \\
z_8 &= \begin{cases} 1 & \text{for an Asian,} \\ 0 & \text{otherwise,} \end{cases}
\end{aligned}
$$

Dummy Variable for Hormones or DES

$$z_9 = \begin{cases} 1 & \text{if a subject's mother did not use hormones or DES,} \\ 0 & \text{otherwise,} \end{cases}$$

$$z_{10} = \begin{cases} 1 & \text{if a subject's mother used hormones only,} \\ 0 & \text{otherwise,} \end{cases}$$

$$z_{11} = \begin{cases} 1 & \text{if a subject's mother used DES only,} \\ 0 & \text{otherwise,} \end{cases}$$

$$z_{12} = \begin{cases} 1 & \text{if a subject's mother used both hormones and DES,} \\ 0 & \text{otherwise.} \end{cases}$$

Variable Selection

Table: MLE for an Initially Selected Model

Selected variable	Degrees of freedom	Coefficient Estimate	Standard Error	p-value
Intercept	1	-2.172	0.6912	0.0017
x_1(age)	1	0.046	0.0218	0.0356
z_6(Black)	1	0.771	0.2296	0.0008
x_6(educ.)	1	-0.159	0.0501	0.0015
z_{10}(horm.)	1	1.794	0.5744	0.0018

The model selection is based on the observations with complete information in all predictors even though fewer predictors are considered in later steps.

Variable Selection

x_7 (employment) and x_8 (smoking) were not selected and had most of the missing data, and hence removed from the selection.

Table: MLE for a Revised Model

Selected variable	Degrees of freedom	Coefficient Estimate	Standard Error	p-value
Intercept	1	-2.334	0.4583	0.0001
x_6(educ.)	1	-0.076	0.0313	0.0151
z_6(Black)	1	0.705	0.1688	0.0001
x_{11}(grav.)	1	0.114	0.0466	0.0142
z_{10}(horm.)	1	1.535	0.4999	0.0021

Final Model

Table: MLE for the Final Model

Selected variable	Degrees of freedom	Coefficient Estimate	Standard Error	p-value
Intercept	1	-2.344	0.4584	0.0001
x_6(educ.)	1	-0.076	0.0313	0.0156
z_6(Black)	1	0.699	0.1688	0.0001
x_{11}(grav.)	1	0.115	0.0466	0.0137
z_{10} (horm.)	1	1.539	0.4999	0.0021

Comparison of the Initial and Final Fits

Selected variable	Degrees of freedom	Coefficient Estimate	Standard Error	p-value
Intercept	1	-2.334	0.4583	0.0001
x_6(educ.)	1	-0.076	0.0313	0.0151
z_6(Black)	1	0.705	0.1688	0.0001
x_{11}(grav.)	1	0.114	0.0466	0.0142
z_{10}(horm.)	1	1.535	0.4999	0.0021
Intercept	1	-2.344	0.4584	0.0001
x_6(educ.)	1	-0.076	0.0313	0.0156
z_6(Black)	1	0.699	0.1688	0.0001
x_{11}(grav.)	1	0.115	0.0466	0.0137
z_{10} (horm.)	1	1.539	0.4999	0.0021

Interactions

- Two-way interactions between the selected variables were examined.
- The backward stepwise procedure was run again.
- No interaction terms were significant at the level of 0.05.
- The final model does not include any interaction.

Interpretation

- The odds ratio for a Black woman (z_6) to deliver a premature infant is doubled relative to that for a White woman, because the corresponding odds ratio equals $\exp(0.699) \approx 2.013$..
- The use of DES by the mother of the pregnant woman (z_{10}) has a significant and enormous effect on the preterm delivery.
- Years of education (x_6), however, seems to have a small, but significant, protective effect.
- Finally, the number of previous pregnancies (x_{11}) has a significant, but low-magnitude negative effect on the preterm delivery.

Interpretation

- The odds ratio for a Black woman (z_6) to deliver a premature infant is doubled relative to that for a White woman, because the corresponding odds ratio equals $\exp(0.699) \approx 2.013$.
- The use of DES by the mother of the pregnant woman (z_{10}) has a significant and enormous effect on the preterm delivery.
- Years of education (x_6), however, seems to have a small, but significant, protective effect.
- Finally, the number of previous pregnancies (x_{11}) has a significant, but low-magnitude negative effect on the preterm delivery.

Impact of Missing Data

Selected variable	Degrees of freedom	Coefficient Estimate	Standard Error	p-value
Intercept	1	-2.172	0.6912	0.0017
x_1(age)	1	0.046	0.0218	0.0356
z_6(Black)	1	0.771	0.2296	0.0008
x_6(educ.)	1	-0.159	0.0501	0.0015
z_{10}(horm.)	1	1.794	0.5744	0.0018
Intercept	1	-2.344	0.4584	0.0001
x_6(educ.)	1	-0.076	0.0313	0.0156
z_6(Black)	1	0.699	0.1688	0.0001
x_{11}(grav.)	1	0.115	0.0466	0.0137
z_{10} (horm.)	1	1.539	0.4999	0.0021

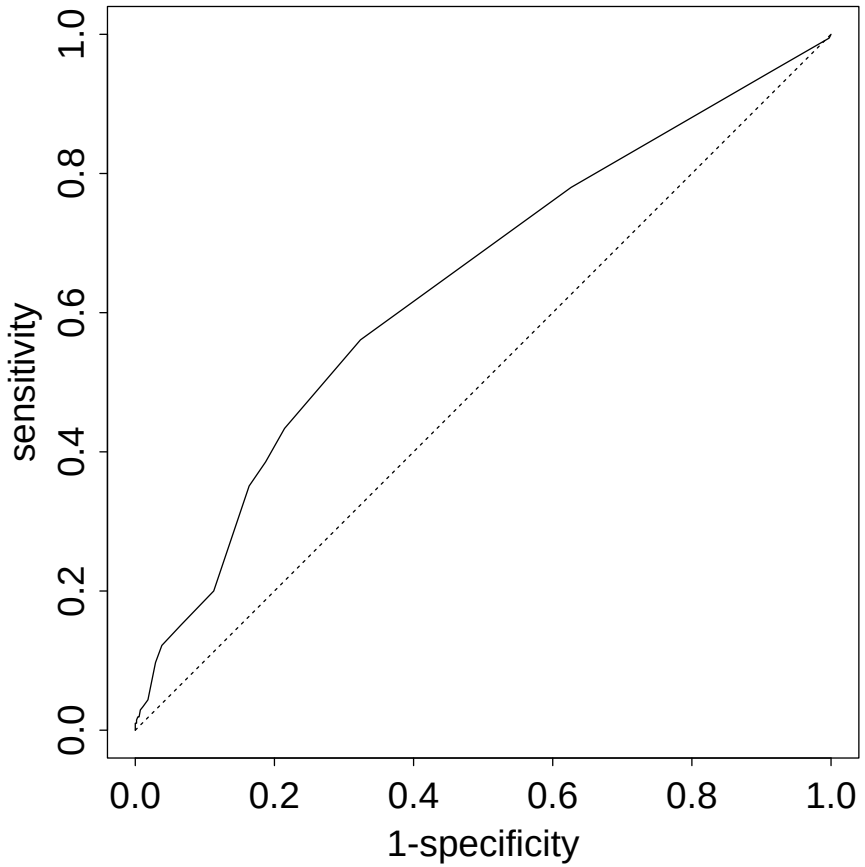
Impact of Missing Data

- Missing data may lead to serious loss of information.
- We may end up with imprecise or even false conclusions.
- Variables change in the selected models.
- The estimated coefficients can be notably different.

Predictive Performance

- false-positive errors
- false-negative errors
- receiver operating characteristic (ROC) curve plots true-positive probability (y-axis) against false-positive probability (x-axis)
- true positive probability: sensitivity
- true negative probability: specificity

The computer-selected tree structure



Node Impurity

- Intuitively, the least impure node should have only one class of outcome (i.e., $\boldsymbol{P}\{Y = 1 \,|\, \tau\} = 0$ or 1), and its impurity is zero.
- Node τ is most impure when $\boldsymbol{P}\{Y = 1 \,|\, \tau\} = \frac{1}{2}$.
- The impurity function has a concave shape and can be formally defined as

$$i(\tau) = \phi(\{Y = 1 \,|\, \tau\}),$$

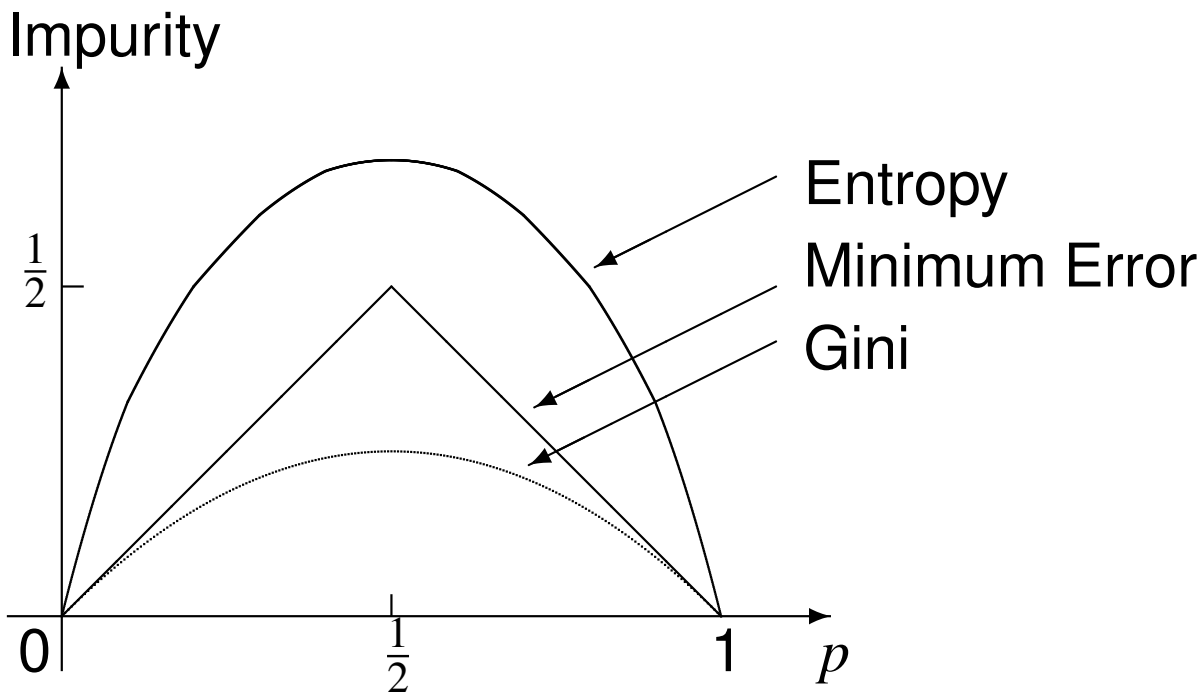
where the function ϕ has the properties (i) $\phi \geq 0$ and (ii) for any $p \in (0, 1)$, $\phi(p) = \phi(1 - p)$ and $\phi(0) = \phi(1) < \phi(p)$.

Common Choices of Node Impurity

- $\phi(p) = \min(p, 1-p)$, (Bayes or the minimum error)
- $\phi(p) = -p\log(p) - (1-p)\log(1-p)$, (entropy)
- $\phi(p) = p(1-p)$, (Gini index)

where $0 \log 0 = 0$.
Devroye et al. (1996, p. 29) call these ϕ's the F-errors.

Impurity Functions



Entropy and Likelihood

- Suppose that Y in node τ_L follows a binomial distribution with a frequency of θ, namely,

$$\boldsymbol{P}\{Y = 1 \,|\, \tau_L\} = \theta.$$

- The log-likelihood function from the $n_{1\cdot}$ observations in node τ_L is

$$n_{11} \log(\theta) + n_{12} \log(1 - \theta).$$

- The maximum of this log-likelihood function is

$$n_{11} \log\left(\frac{n_{11}}{n_{1\cdot}}\right) + n_{12} \log\left(\frac{n_{12}}{n_{1\cdot}}\right),$$

  which is proportional to the entropy.

Determination of Terminal Nodes

- For a tree $\mathcal{T}$ we define

$$R(\mathcal{T}) = \sum_{\tau \in \tilde{\mathcal{T}}} \boldsymbol{P}\{\tau\} r(\tau),$$

  where $\tilde{\mathcal{T}}$ is the set of terminal nodes of $\mathcal{T}$.
- $r(\tau)$ measures a certain quality of node τ. It is similar to the sum of the squared residuals in the linear regression.
- The purpose of pruning is to select the best subtree, $\mathcal{T}^*$, of an initially saturated tree, $\mathcal{T}_0$, such that $R(\mathcal{T})$ is minimized.

Misclassification Cost

- Let $c(i|j)$ be a unit misclassification cost that a class j subject is classified as a class i subject.
- When $i = j$, we have the correct classification and the cost should naturally be zero, i.e., $c(i|i) = 0$.
- Without loss of generality we can set $c(1|0) = 1$.
- The clinicians and the statisticians need to work together to gauge the relative cost of $c(0|1)$.

Misclassification Cost

	Assumed	Node number				
	Class	1	2	3	4	5
$c(0\|1)$	1	3656	640	3016	187	453
1	0	205	70	135	11	59
10	0	2050	700	1350	110	590
18	0	3690	1260	2430	198	1062

Class Assignment

Node τ is assigned class j if

$$\sum_i [c(j|i)\boldsymbol{P}\{Y = i \,|\, \tau\}] \leq \sum_i [c(1-j|i)\boldsymbol{P}\{Y = i \,|\, \tau\}].$$

For example, when $c(0|1) = 10$, it means that one false-negative error counts as many as ten false-positive ones. The cost is 3656 if the root node is assigned class 1. It becomes $225 \times 10 = 2250$ if the root node is assigned class 0. Therefore, the root node should be assigned class 0 for $2250 < 3656$.

Use of $r(\tau)$ for Splitting?

It is usually difficult to assign the cost function before any tree is grown. As a matter of fact, the assignment can still be challenging even when a tree profile is given.

Resubstitution Estimates of Misclassification Cost

Unit cost: $c(0|1) = 10)$

Node number	Node class	Weight $\boldsymbol{P}\{\tau\}$	Within-node cost $r(\tau)$	Cost $R^s(\tau)$
1	0	$\frac{3861}{3861}$	$\frac{10*205}{3861}$	$\frac{2050}{3861} = 0.531$
2	1	$\frac{710}{3861}$	$\frac{1*640}{710}$	$\frac{640}{3861} = 0.166$
3	0	$\frac{3151}{3861}$	$\frac{10*135}{3151}$	$\frac{1350}{3861} = 0.35$
4	0	$\frac{198}{3861}$	$\frac{10*11}{198}$	$\frac{110}{3861} = 0.028$
5	1	$\frac{506}{3861}$	$\frac{1*453}{506}$	$\frac{453}{3861} = 0.117$

Caveat of Resubstitution Estimates

- Let $R^s(\tau)$ denote the resubstitution estimate of the misclassification cost for node τ.
- The resubstitution estimates generally underestimate the cost.
- If we have an independent data set, we can assign the new subjects to various nodes of the tree and calculate the cost based on these new subjects. This cost tends to be higher than the resubstitution estimate, because the split criteria are somehow related to the cost, and as a result, the resubstitution estimate of misclassification cost is usually overoptimistic.
- In some applications, such an independent data set, called a test sample or validation set, is available.

Cost–Complexity

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

where α (≥ 0) is the complexity parameter and $|\tilde{\mathcal{T}}|$ is the number of terminal nodes in $\mathcal{T}$.
The use of tree cost-complexity allows us to construct a sequence of nested "essential" subtrees from any given tree $\mathcal{T}$ so that we can examine the properties of these subtrees and make a selection from them.

Cost–Complexity

- Let $\mathcal{T}_0$, be the five-node tree. The cost for $\mathcal{T}_0$ is $0.350 + 0.028 + 0.117 = 0.495$ and its complexity is 3. Thus, its cost-complexity is $0.495 + 3\alpha$ for a given complexity parameter α.
- Is there a subtree of $\mathcal{T}_0$ that has a smaller cost-complexity?

Theorem

(Breiman et al. 1984, Section 3.3) *For any value of the complexity parameter α, there is a unique smallest subtree of $\mathcal{T}_0$ that minimizes the cost-complexity.*

Cost–Complexity

- We cannot have two subtrees of the smallest size and of the same cost-complexity.
- This smallest subtree is referred to as the optimal subtree with respect to the complexity parameter.
- When $\alpha = 0$, the optimal subtree is $\mathcal{T}_0$ itself.
- What are the other subtrees and their complexities?

Cost–Complexity

- We can always choose α large enough that the corresponding optimal subtree is the single-node tree.
- When $\alpha \geq 0.018$, $\mathcal{T}_2$ (the root node tree) becomes the optimal subtree, because

$$R_{0.018}(\mathcal{T}_2) = 0.531 + 0.018 * 1 = 0.495 + 0.018 * 3 = R_{0.018}(\mathcal{T}_0)$$

and

$$R_{0.018}(\mathcal{T}_2) = 0.531 + 0.018 * 1 < 0.516 + 0.018 * 2 = R_{0.018}(\mathcal{T}_1).$$

- Although $R_{0.018}(\mathcal{T}_2) = R_{0.018}(\mathcal{T}_0)$, $\mathcal{T}_2$ is the optimal subtree, because it is smaller than $\mathcal{T}_0$.
- This calculation confirms the theorem that we do not have two subtrees of the smallest size and of the same cost-complexity.

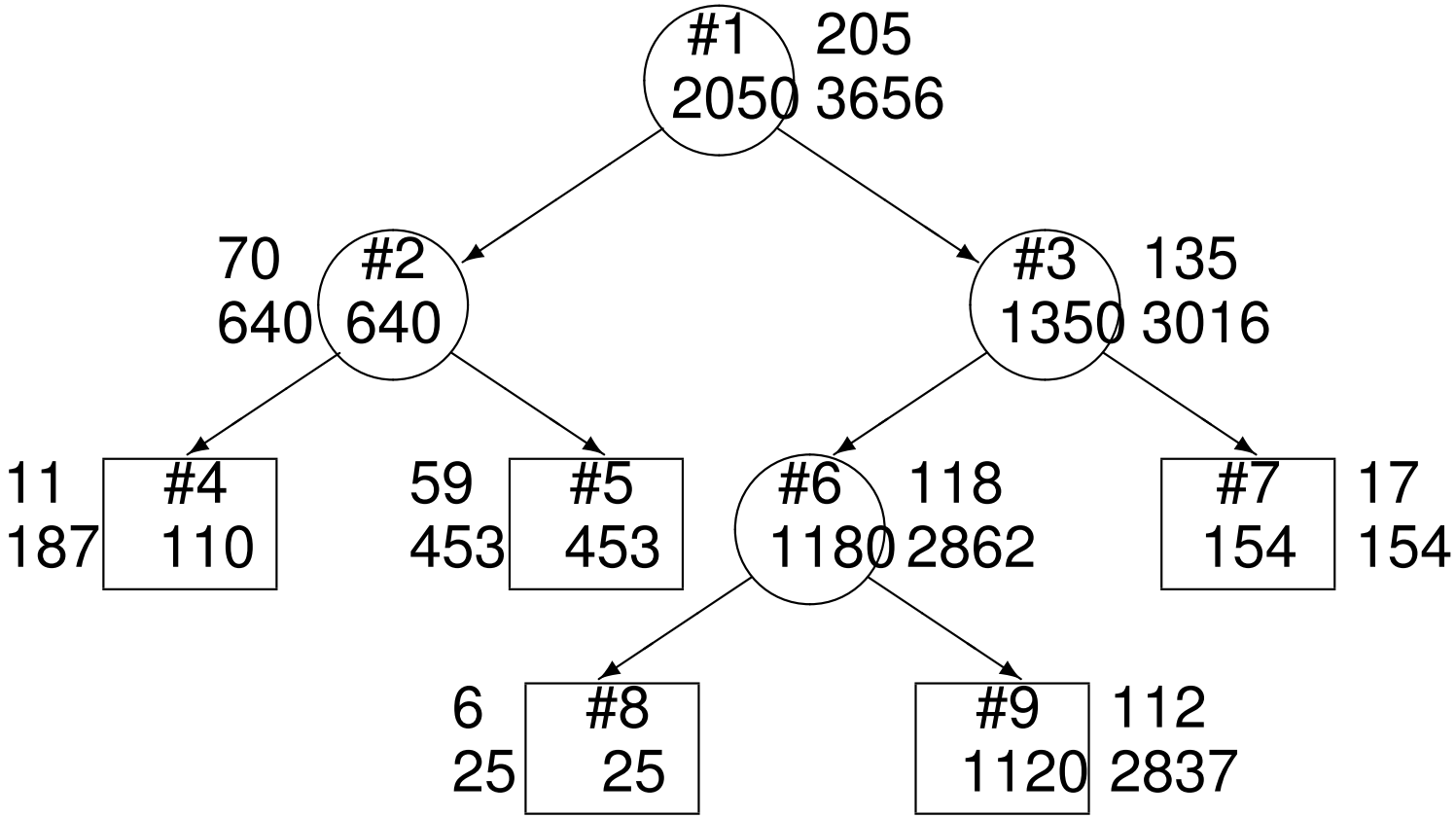
Cost–Complexity

- $\mathcal{T}_1$ is not an optimal subtree for any α.
- $\mathcal{T}_0$ is the optimal subtree for any $\alpha \in [0, 0.018)$ and $\mathcal{T}_2$ is the optimal subtree when $\alpha \in [0.018, \infty)$.
- Not all subtrees are optimal with respect to a complexity parameter.
- Although the complexity parameter takes a continuous range of values, we have only a finite number of subtrees.
- An optimal subtree is optimal for an interval range of the complexity parameter, and the number of such intervals has to be finite.

Nested Optimal Subtrees

- We derive the first positive threshold parameter, α_1, for this tree by comparing the resubstitution misclassification cost of an internal node to the sum of the resubstitution misclassification costs of its offspring terminal nodes.
- Note the sum of the resubstitution misclassification costs of its offspring terminal nodes denoted by $R^s(\tilde{\mathcal{T}}_\tau)$ for a node τ.

Nested Optimal Trees



Cost–Complexity Parameter

Node	$R^s(\tau)$	$R^s(\tilde{\mathcal{T}}_\tau)$	$\lvert\tilde{\mathcal{T}}_\tau\rvert$	α
9	0.290	0.290	1	
8	0.006	0.006	1	
7	0.040	0.040	1	
6	0.306	0.296	2	0.010
5	0.117	0.117	1	
4	0.028	0.028	1	
3	0.350	0.336	3	0.007
2	0.166	0.145	2	0.021
1	0.531	0.481	5	0.013
		Minimum		0.007

Cost–Complexity Parameter

- The cost of node 3 per se is $R^s(3) = 1350/3861 = 0.350$.
- It is the ancestor of terminal nodes 7, 8, and 9. The units of misclassification cost within these three terminal nodes are respectively 154, 25, and 1120. Hence, $R^s(\tilde{\mathcal{T}}_3) = (154 + 25 + 1120)/3861 = 0.336$.
- The difference between $R^s(3)$ and $R^s(\tilde{\mathcal{T}}_3)$ is $0.350 - 0.336 = 0.014$.
- The difference in complexity between node 3 alone and its three offspring terminal nodes is $3 - 1 = 2$.
- On average, an additional terminal node reduces the cost by $0.014/2 = 0.007$.

Consequence of Pruning

- If we cut the offspring nodes of the root node, we have the root-node tree whose cost-complexity is $0.531 + \alpha$.
- For it to have the same cost-complexity as the initial nine-node tree, we need $0.481 + 5\alpha = 0.531 + \alpha$, giving $\alpha = 0.013$.
- How about changing node 2 to a terminal node?
  - The initial nine-node tree is compared with a seven-node subtree, consisting of nodes 1 to 3, and 6 to 9.
  - For the new subtree to have the same cost-complexity as the initial tree, we find $\alpha = 0.021$.
- In fact, for any internal node, $\tau \notin \tilde{\mathcal{T}}$, the value of α is precisely

$$\frac{R^s(\tau) - R^s(\tilde{\mathcal{T}}_\tau)}{|\tilde{\mathcal{T}}_\tau| - 1}.$$

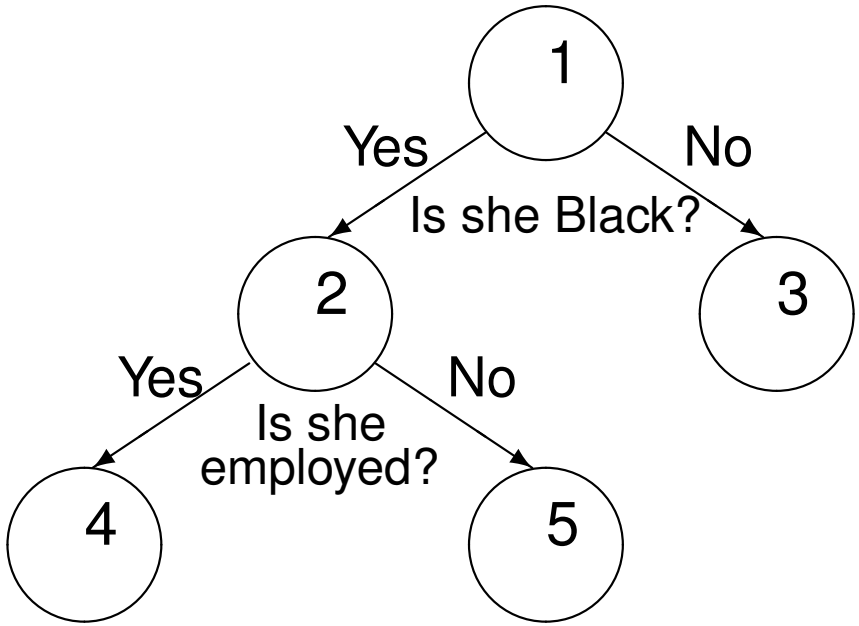
- The first positive threshold parameter, α_1, is the smallest α over the $|\tilde{\mathcal{T}}| - 1$ internal nodes.

A Pruned Tree

Using α_1 we change an internal node τ to a terminal node when

$$R^s(\tau) + \alpha_1 \leq R^s(\tilde{\mathcal{T}}_\tau) + \alpha_1 |\tilde{\mathcal{T}}_\tau|$$

until this is not possible. This pruning process results in the optimal subtree corresponding to α_1.



Nested Optimal Subtrees

- After pruning the tree using the first threshold, we seek the second threshold complexity parameter, α_2.
- We knew from our previous discussion that $\alpha_2 = 0.018$ and its optimal subtree is the root-node tree. No more thresholds need to be found from here, because the root-node tree is the smallest one.
- In general, suppose that we end up with m thresholds, $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_m$, and let $\alpha_0 = 0$.
- Let the corresponding optimal subtrees be $\mathcal{T}_{\alpha_0} \supset \mathcal{T}_{\alpha_1} \supset \mathcal{T}_{\alpha_2} \supset \cdots \supset \mathcal{T}_{\alpha_m}$, where $\mathcal{T}_{\alpha_1} \supset \mathcal{T}_{\alpha_2}$ means that $\mathcal{T}_{\alpha_2}$ is a subtree of $\mathcal{T}_{\alpha_1}$.

Nested Optimal Subtrees

Theorem

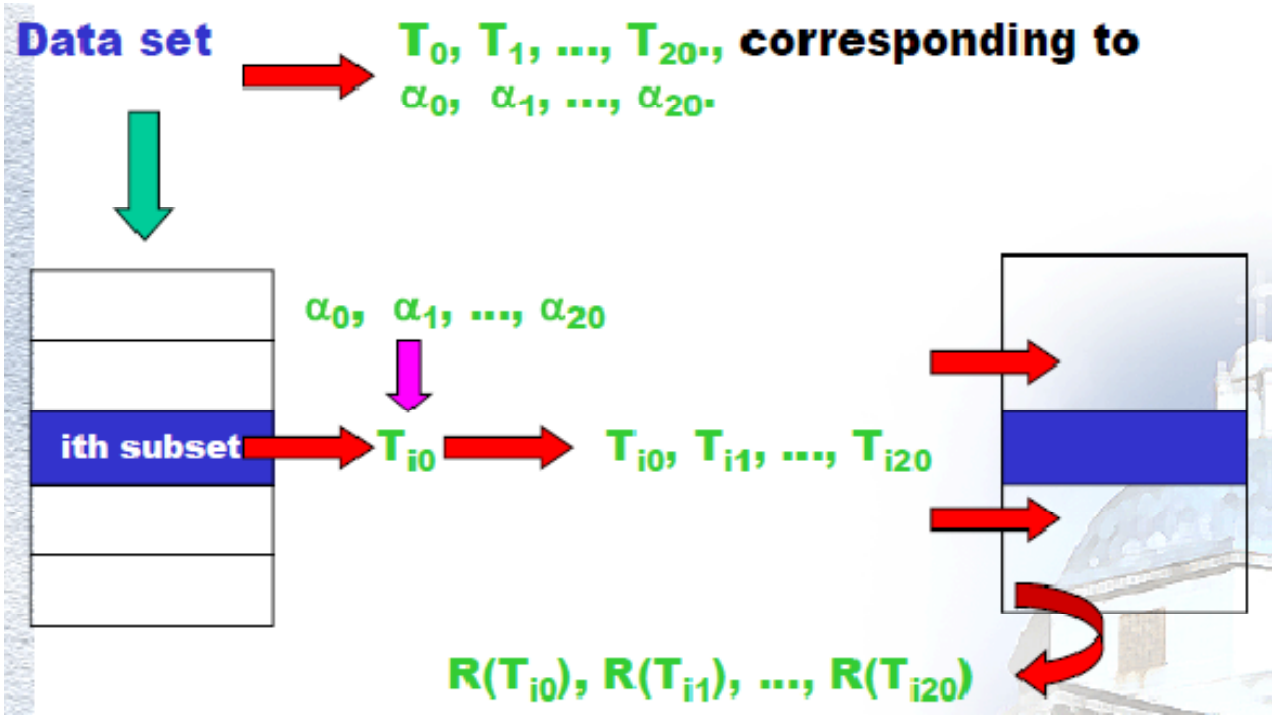
If $\alpha_1 > \alpha_2$, the optimal subtree corresponding to α_1 is a subtree of the optimal subtree corresponding to α_2.

- What's next?
- We need a good estimate of $R(\mathcal{T}_{\alpha_k})$ $(k = 0, 1, \ldots, m)$, namely, the misclassification costs of the subtrees.
- We will select the one with the smallest misclassification cost.

Select the Optimal Subtree

- When a test sample is available, estimating $R(\mathcal{T})$ for any subtree $\mathcal{T}$ is straightforward, because we only need to apply the subtrees to the test sample.
- Difficulty arises when we do not have a test sample.
- The cross-validation process is generally used by creating artificial test samples.
- Divide the entire study sample into a number of pieces, usually 5, 10, or 25 corresponding to 5-, 10-, or 25-fold cross-validation, respectively.

Cross-validation



Cross-validation

- Randomly divide the 3861 women into five groups: 1 to 5. Group 1 has 773 women and each of the rest contains 772 women.
- Let $\mathcal{L}_{(-i)}$ be the sample set including all but those subjects in group i, $i = 1, \ldots, 5$.
- Using the 3088 women in $\mathcal{L}_{(-1)}$, produce another large tree, say $\mathcal{T}_{(-1)}$, in the same way as we did using all 3861 women.
- Take each α_k from the sequence of complexity parameters as has already been derived above and obtain the optimal subtree, $\mathcal{T}_{(-1),k}$, of $\mathcal{T}_{(-1)}$ corresponding to α_k.
- We have a sequence of the optimal subtrees of $\mathcal{T}_{(-1)}$, i.e., $\{\mathcal{T}_{(-1),k}\}_0^m$.
- Using group 1 as a test sample relative to $\mathcal{L}_{(-1)}$, we have an unbiased estimate, $R^{ts}(\mathcal{T}_{(-1),k})$, of $R(\mathcal{T}_{(-1),k})$.

Cross-validation

- Because $\mathcal{T}_{(-1),k}$ is related to $\mathcal{T}_{\alpha_k}$ through the same α_k, $R^{ts}(\mathcal{T}_{(-1),k})$ can be regarded as a cross-validation estimate of $R(\mathcal{T}_{\alpha_k})$.
- Using $\mathcal{L}_{(-i)}$ as the learning sample and the data in group i as the test sample, we also have $R^{ts}(\mathcal{T}_{(-i),k})$, $(i = 2, \ldots, 5)$ as the cross-validation estimate of $R(\mathcal{T}_{\alpha_k})$.
- The final cross-validation estimate, $R^{cv}(\mathcal{T}_{\alpha_k})$, of $R(\mathcal{T}_{\alpha_k})$ follows from averaging $R^{ts}(\mathcal{T}_{(-i),k})$ over $i = 1, \ldots, 5$.

Cross-validation

- The subtree corresponding to the smallest R^{cv} is obviously desirable.
- The cross-validation estimates generally have substantial variabilities.
- Breiman et al. (1984) proposed a revised strategy to select the final tree, which takes into account the standard errors of the cross-validation estimates.
  - Let SE_k be the standard error for $R^{cv}(\mathcal{T}_{\alpha_k})$.
  - Suppose that $R^{cv}(\mathcal{T}_{\alpha_{k*}})$ is the smallest among all $R^{cv}(\mathcal{T}_{\alpha_k})$'s.
  - The revised selection rule selects the smallest subtree whose cross-validation estimate is within a prespecified range of $R^{cv}(\mathcal{T}_{\alpha_{k*}})$, which is usually defined by one unit of SE_{k*}. This is the so-called 1-SE rule.
- Empirical evidence suggests that the tree selected with the 1-SE rule is often superior to the one selected with the 0-SE rule.

Cross-validation

- Every subject in the entire study sample was used once as a testing subject and was assigned a class membership $m+1$ times through the sequence of $m+1$ subtrees built upon the corresponding learning sample.
- Let $C_{i,k}$ be the misclassification cost incurred for the ith subject while it was a testing subject and the classification rule was based on the kth subtree, $i = 1, \ldots, n$, $k = 0, 1, \ldots, m$.
- $R^{cv}(\mathcal{T}_{\alpha_k}) = \sum_{j=0,1} \boldsymbol{P}\{Y = j\}\bar{C}_{k|j}$, where $\bar{C}_{k|j}$ is the average of $C_{i,k}$ over the set S_j of the subjects whose response is j (i.e., $Y = j$).

Cross-validation

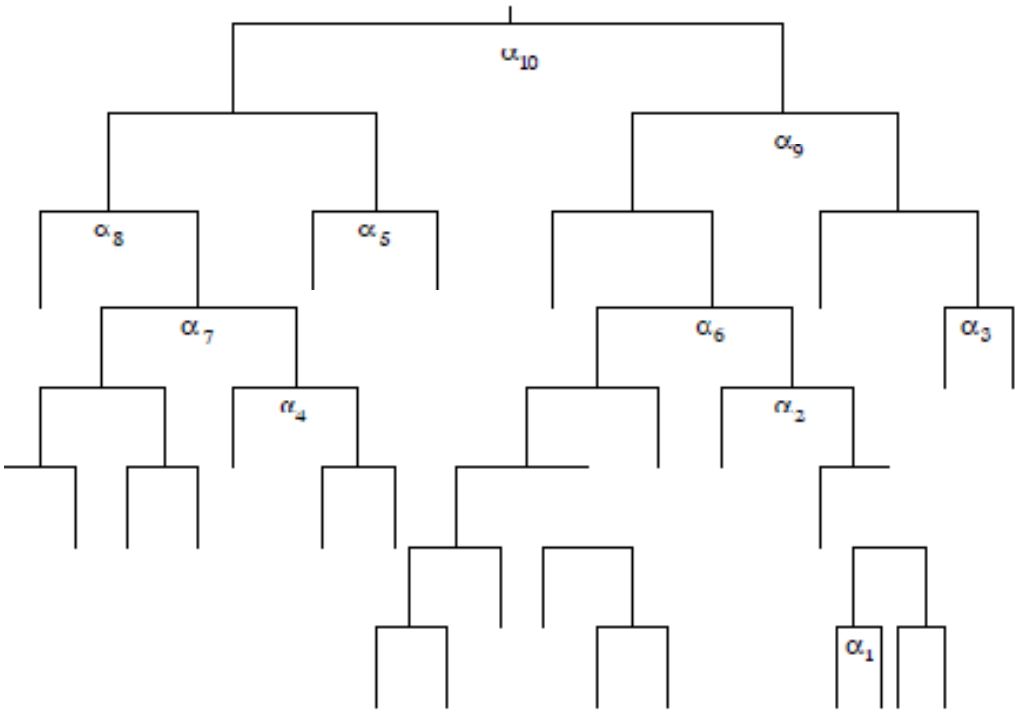
- $C_{i,k}$'s are likely to be correlated with each other, because $C_{i,k}$ is the cost from the same subject (the ith one) while the subtree (the kth one) varies.
- For convenience, however, they are treated as if they were not correlated.
- The sample variance of each $\bar{C}_{k|j}$ is

$$\frac{1}{n_j^2}\left(\sum_{i\in S_j} C_{i,k}^2 - n_j \bar{C}_{k|j}^2\right).$$

- The heuristic standard error for $R^{cv}(\mathcal{T}_{\alpha_k})$ is given by

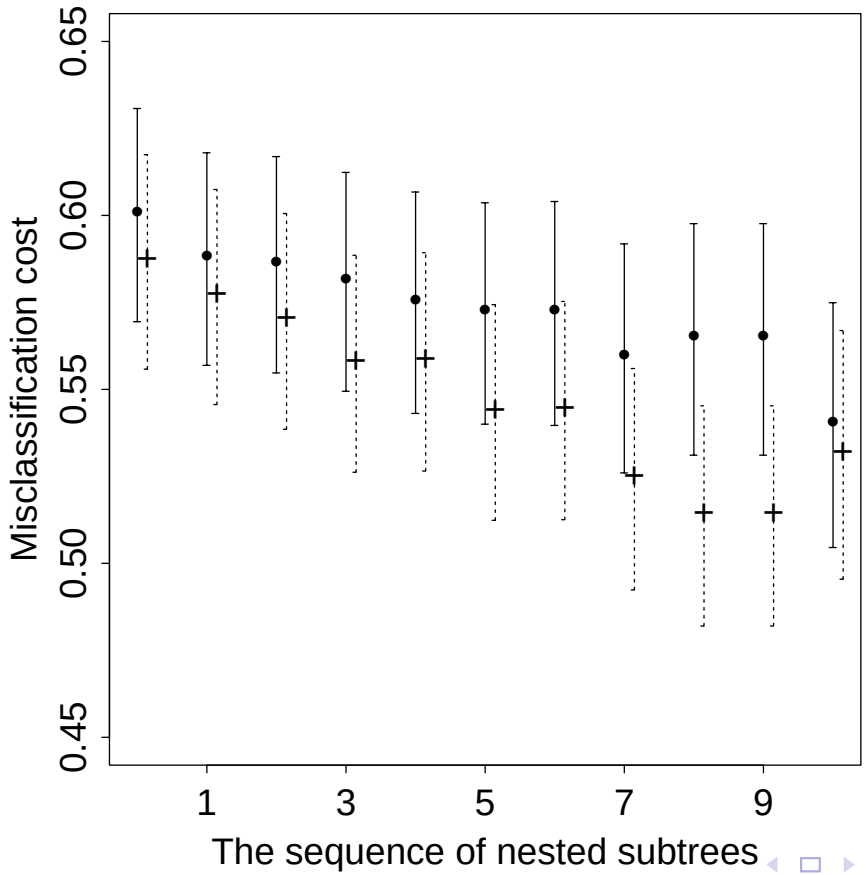
$$\mathrm{SE}_k = \left\{\sum_{j=0,1}\left(\frac{\boldsymbol{P}\{Y=j\}}{n_j}\right)^2 \left(\sum_{i\in S_j} C_{i,k}^2 - n_j \bar{C}_{k|j}^2\right)\right\}^{1/2}.$$

An Initial with $C(0|1) = 10$



Cross-validation estimates of MC

5- and 10-fold estimates are respectively indicated by • and +. Also plotted along the estimates are the intervals of length of two SEs..



Interpretation

- The 1-SE rule selects the root-node subtree.
- The risk factors considered here may not have enough predictive power to stand out and pass the cross-validation.
- This statement is obviously relative to the selected unit cost $C(0|1) = 10$.
- When we used $C(0|1) = 18$ and performed a 5-fold cross-validation, the final tree was different.

An Alternative Pruning Approach

- The choice of the penalty for a false-negative error, $C(0|1) = 10$, is vital to the selection of the final tree structure.
- In many secondary analyses, however, the purpose is mainly to explore the data structure and to generate hypotheses.
- It would be convenient to proceed with the analysis without assigning the unit of misclassification cost.
- Sometimes we cannot hold trees to a fixed algorithm.

An Alternative Pruning Approach

- After the large tree $\mathcal{T}$ is grown, assign a statistic S_τ to each internal node τ from the bottom up.
- Align these statistics in increasing order as

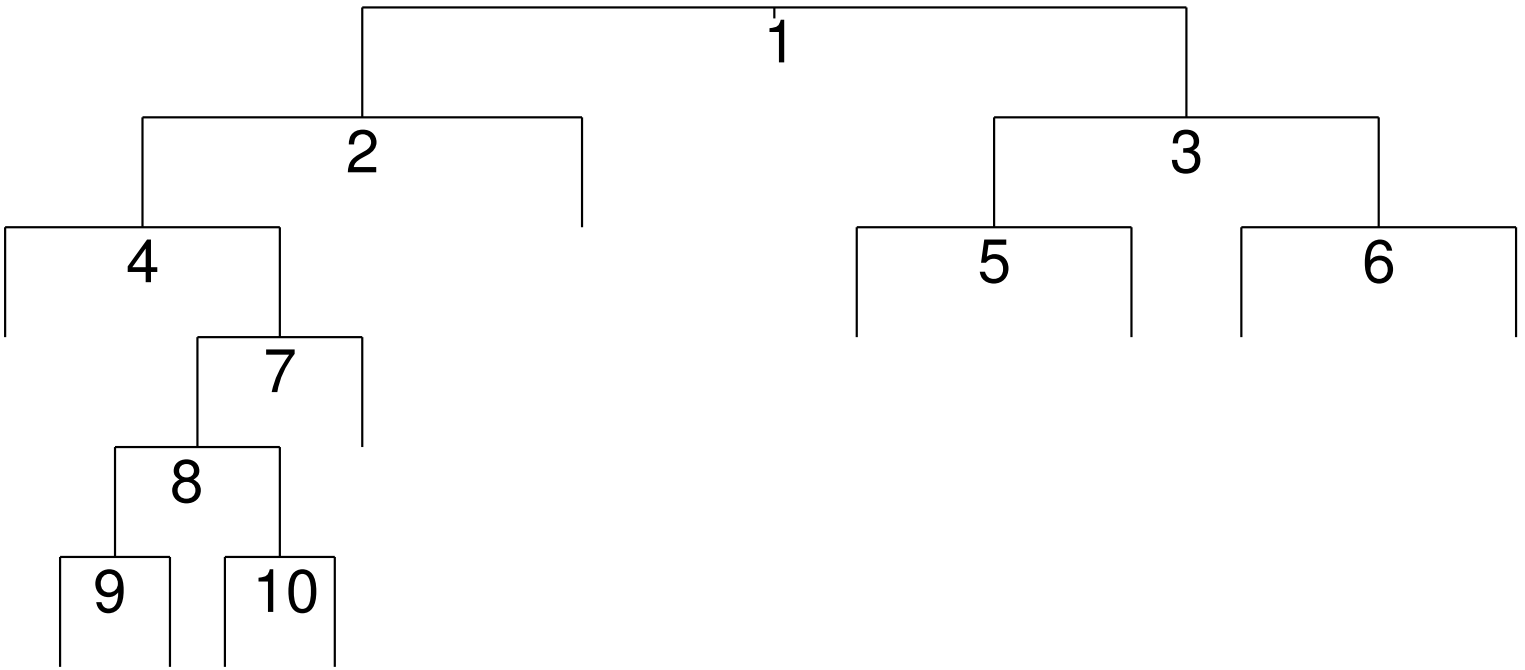
$$S_{\tau_1} \leq S_{\tau_2} \leq \cdots \leq S_{\tau_{|\tilde{\mathcal{T}}|-1}}.$$

- Select a threshold level and change an internal node to a terminal one if its statistic is less than the threshold level.

An Alternative Pruning Approach

- Locate the smallest S_τ over all internal nodes and prune the offspring of the highest node(s) that reaches this minimum.
- What remains is the first subtree.
- Repeat the same process until the subtree contains the root node only.
- As the process continues, a sequence of nested subtrees, $\mathcal{T}_1, \ldots, \mathcal{T}_m$, will be produced. To select a threshold value, we make a plot of $\min_{\tau \in \mathcal{T}_i - \tilde{\mathcal{T}}_i} S_\tau$ versus $|\tilde{\mathcal{T}}_i|$, i.e., the minimal statistic of a subtree against its size.
- Look for a possible "kink" in this plot where the pattern changes.

A Roughly Pruned Tree



Maximum Statistic

	Term	Preterm	
Left Node	640	70	710
Right Node	3016	135	3151
	3656	205	3861

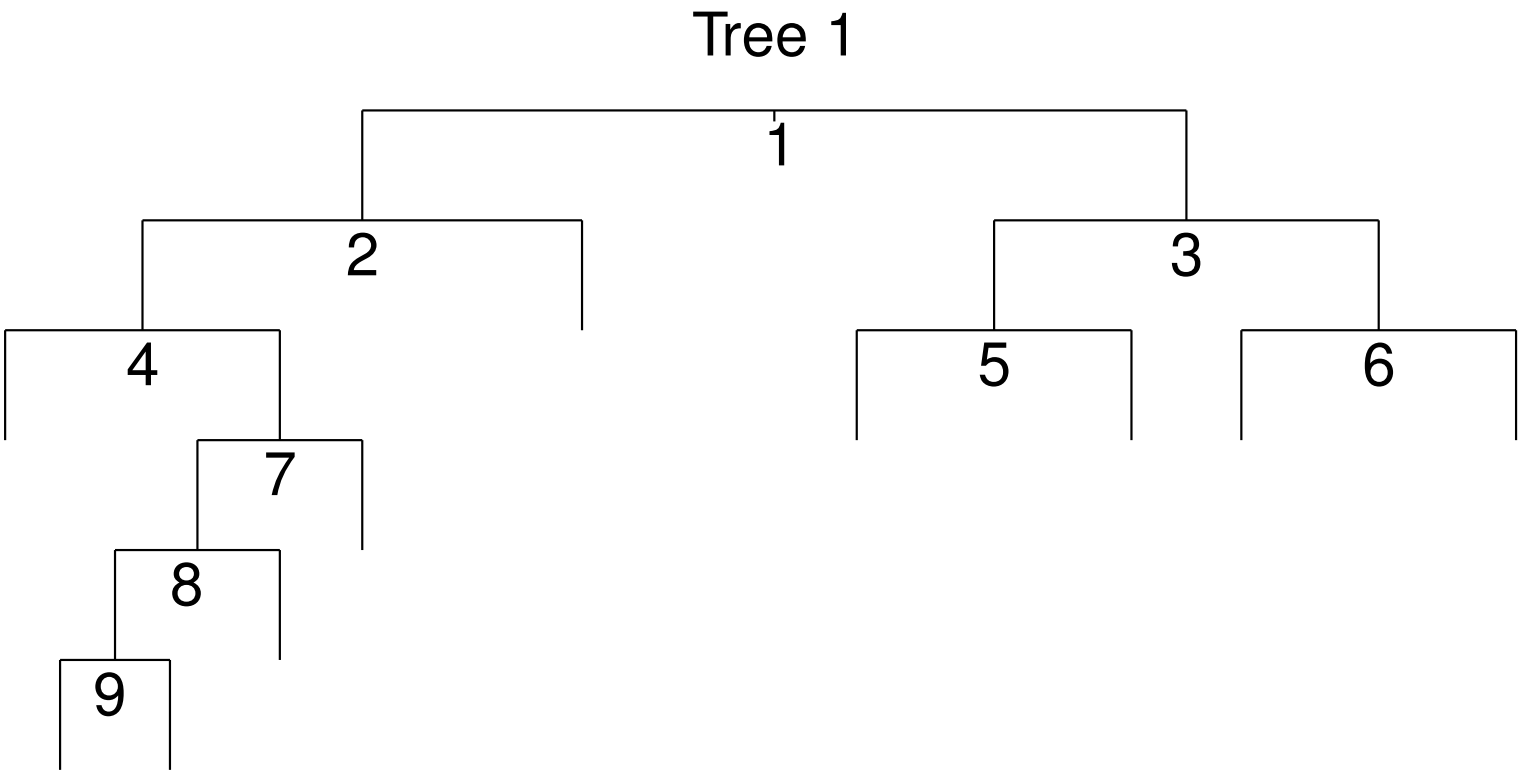
- Relative risk (RR) of preterm as $(70/710)/(135/3151) = 2.3$.
- The standard error for the log RR is approximately $\sqrt{1/70 - 1/710 + 1/135 - 1/3151} = 0.141$.
- The Studentized log relative risk is $0.833/0.141 = 5.91$.

Statistics for Internal Nodes

Node #	1	2	3	4	5
Raw Statistic	5.91	2.29	3.72	1.52	3.64
Max. Statistic	5.91	2.29	3.72	1.94	3.64
Node #	6	7	8	9	10
Raw Statistic	1.69	1.47	1.35	1.94	1.60
Max. Statistic	1.69	1.94	1.94	1.94	1.60

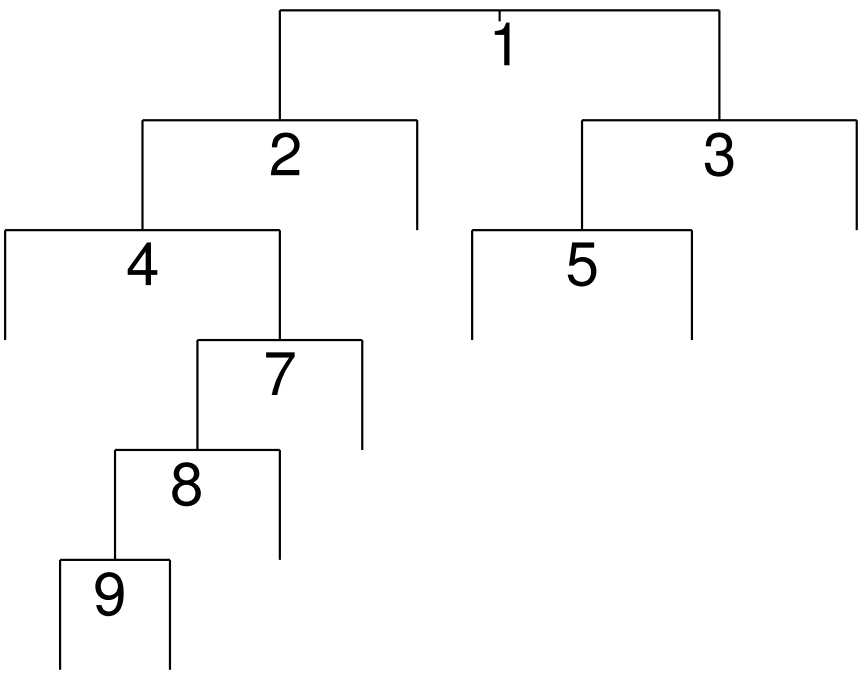
- For each internal node we replace the raw statistic with the maximum of the raw statistics over its offspring internal nodes if the latter is greater.
- For instance, the raw value 1.52 is replaced with 1.94 for node 4;
- The maximum statistic has seven distinct values: 1.60, 1.69, 1.94, 2.29, 3.64, 3.72, and 5.91, each of which results in a subtree.
- We have a sequence of eight nested subtrees.

The First Subtree

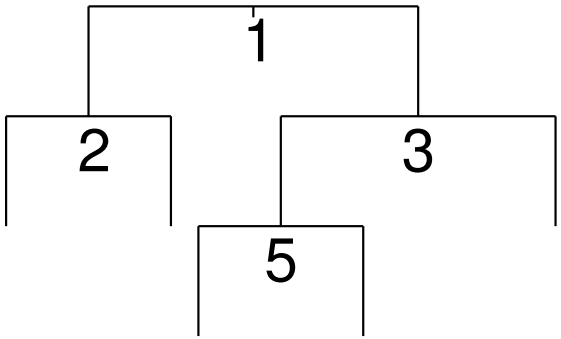


The Next Two Subtrees

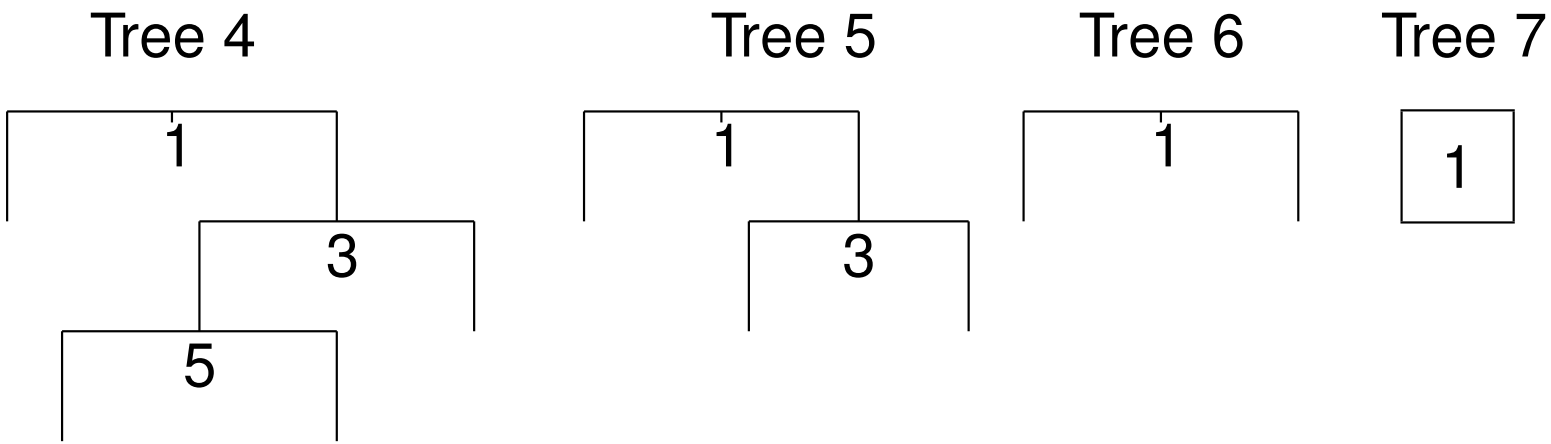
Tree 2



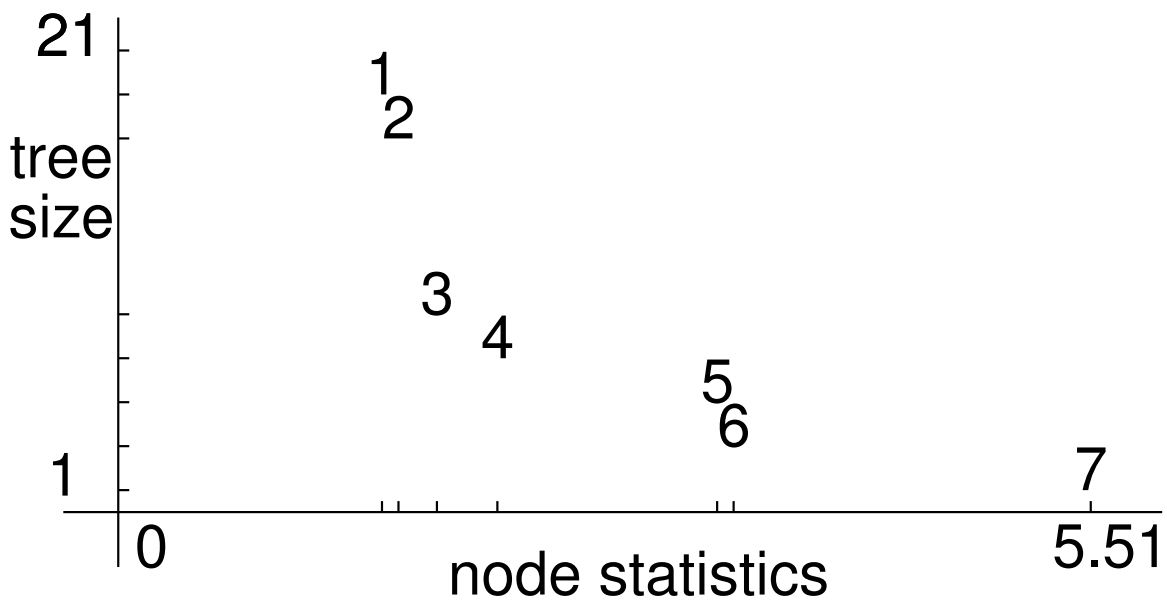
Tree 3



The Next Four Subtrees

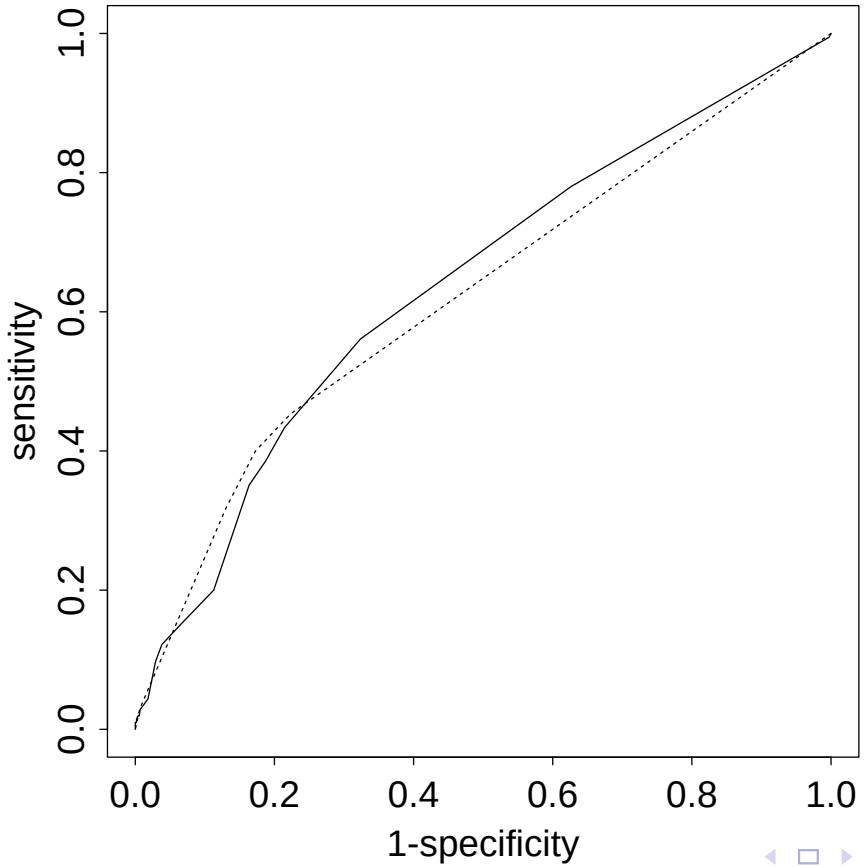


Tree Size vs Internal Node Statistics



Tree-Based vs Logistic Regression

The area under the curve is 0.622 for the tree-based model and 0.637 for the logistic model

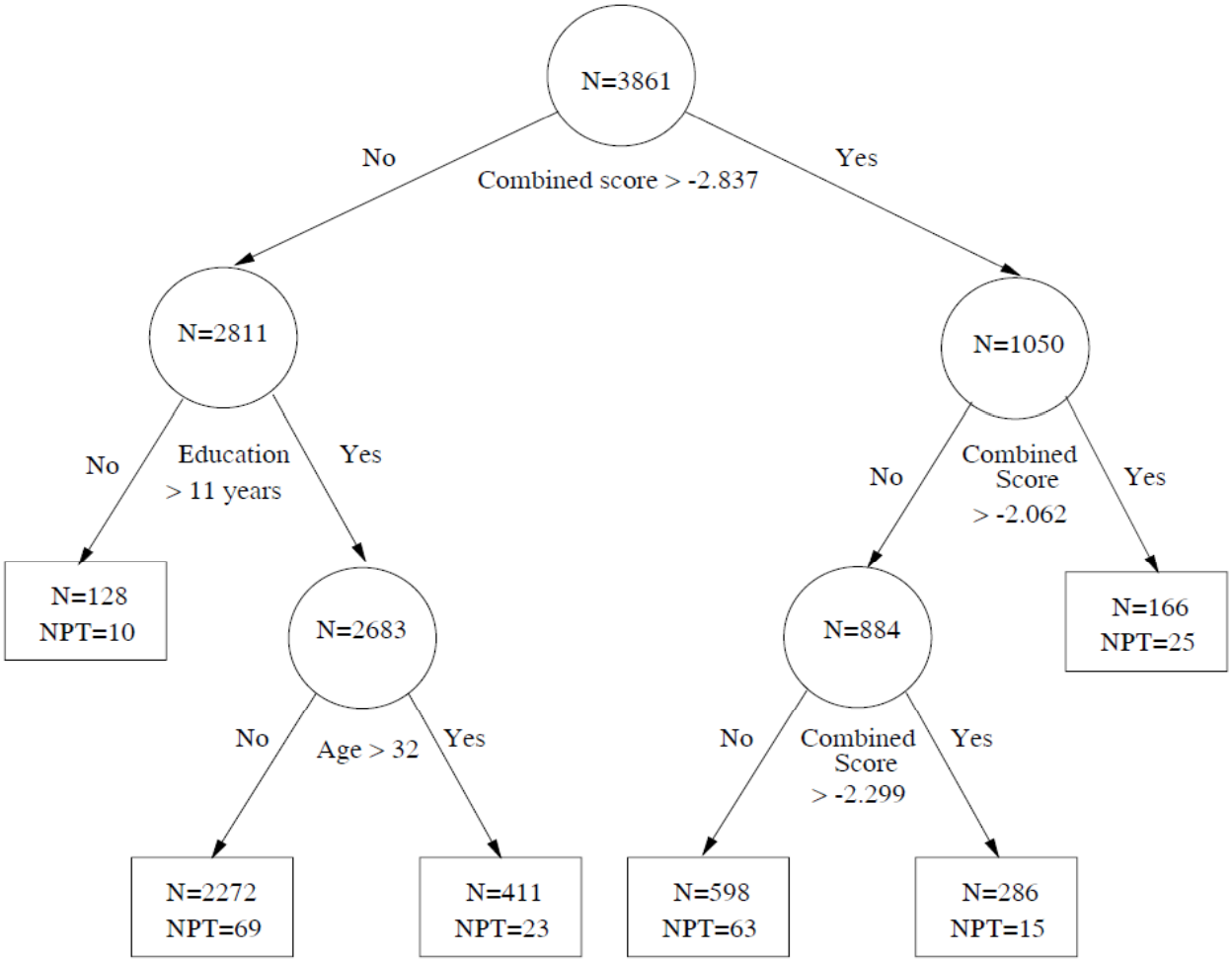


Use of Both Tree-Based and Logistic Regression: Approach I

- Take the linear equation derived from the logistic regression as a new predictor.
- In the present application, the new predictor is defined as $x_{16} = -2.344 - 0.076x_6 + 0.699z_6 + 0.115x_{11} + 1.539z_{10}$.

The Final Tree

The equation from the logistic regression is used.



Conclusion

- Education shows a protective effect, particularly for those with college or higher education.
- Age has merged as a risk factor. In the fertility literature, whether a women is at least 35 years old is a common standard for pregnancy screening.
- The risk of delivering preterm babies is not monotonic with respect to the combined score x_{16}.
- The risk is lower when $-2.837 < x_{16} \leq -2.299$ than when $-2.299 < x_{16} \leq -2.062$.

Use of Both Tree-Based and Logistic Regression: Approach II

- Run the logistic regression after a tree is grown.
- Create five dummy variables, each of which corresponds to one of the five terminal nodes.

Variable label	Specification
z_{13}	Black, unemployed
z_{14}	Black, employed
z_{15}	non-Black, ≤ 4 pregnancies, DES not used
z_{16}	non-Black, ≤ 4 pregnancies, DES used
z_{17}	non-Black, > 4 pregnancies

Use of Both Tree-Based and Logistic Regression: Approach II

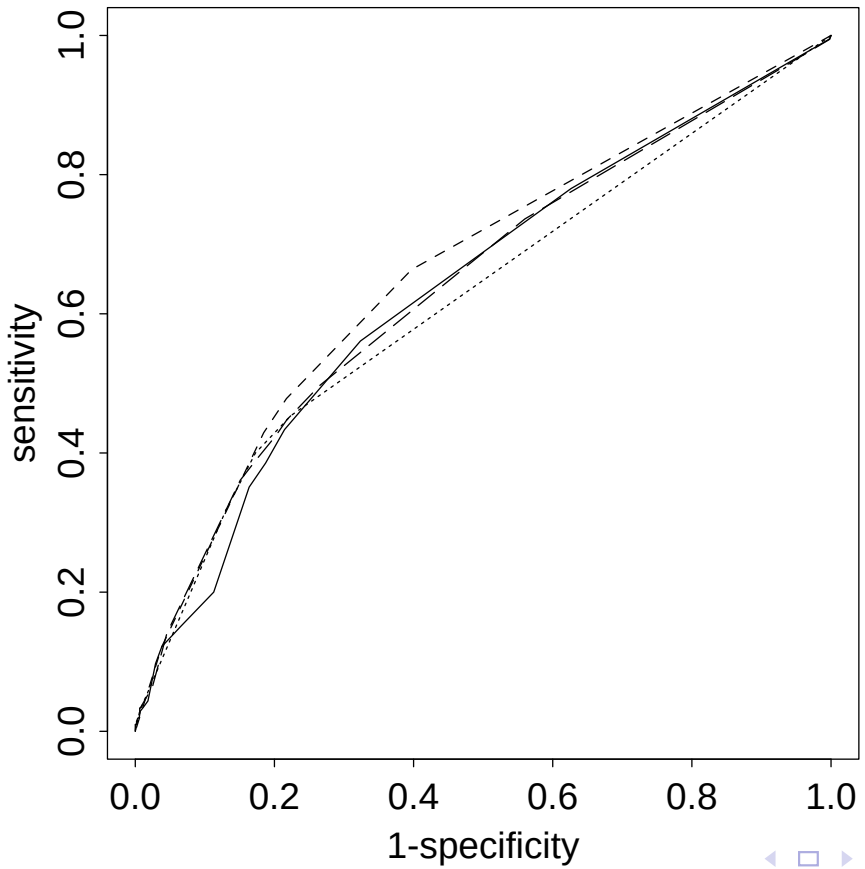
- Include these five dummy variables, z_{13} to z_{17}, in addition to the 15 predictors, x_1 to x_{15}.
- Rebuild a logistic regression model.

$$\hat{\theta} = \frac{\exp(-1.341 - 0.071x_6 - 0.885z_{15} + 1.016z_{16})}{1 + \exp(-1.341 - 0.071x_6 - 0.885z_{15} + 1.016z_{16})}.$$

  - It is very similar to the previous equation.
  - The variables z_{15} and z_{16} are an interactive version of z_6, x_{11}, and z_{10}.
  - The coefficient for x_6 is nearly intact.
  - The area under the new curve is 0.642, which is narrowly higher than 0.639.

Comparison of ROC Curves

dotted: tree; solid: logistic regression model; short-dashed: hybrid I; long-dashed: hybrid II



Missing Data

- Surrogate splits (Breiman et al. 1984, Section 5.3)
- Imputation (Little and Rubin 1987).
- Missings Together (Clark and Pregibon 1992)

Missing Data

- For missings together and imputation, no need to change the tree algorithm.
- For imputation, missing data can be imputed and entered into trees as observed.
- For missings together, we create a new “level” for missing values.
  - Simple to implement and understand.
  - Easy to trace where the subjects with missing information.

Surrogate Splits

- Surrogate splits attempt to utilize the information in other predictors to assist us in splitting when the splitting variable, say, race, is missing.
- The idea to look for a predictor that is most similar to race in classifying the subjects.
- One measure of similarity between two splits suggested by Breiman et al. (1984) is the coincidence probability that the two splits send a subject to the same node.

Coincidence Probability

- The 2×2 table below compares the split of "is age $>$ 35?" with the selected race split.

	Black	Others
Age $\leq$ 35	702	8
Age $>$ 35	3017	134

- 702+134=836 of 3861 subjects are sent to the same node, and hence $836/3861 = 0.217$ can be used as an estimate for the coincidence probability of these two splits.

Coincidence Probability

- In general, prior information should be incorporated in estimating the coincidence probability when the subjects are not randomly drawn from a general population, such as in case–control studies.
- We estimate the coincidence probability with

$$\boldsymbol{P}\{Y=0\}M_0(\tau)/N_0(\tau)+\boldsymbol{P}\{Y=1\}M_1(\tau)/N_1(\tau),$$

where $N_j(\tau)$ is the total number of class j subjects in node τ and $M_j(\tau)$ is the number of class j subjects in node τ that are sent to the same daughters by the two splits; here $j=0$ (normal) and 1(abnormal). $\boldsymbol{P}\{Y=0\}$ and $\boldsymbol{P}\{Y=1\}$ are the priors to be specified. Usually, $\boldsymbol{P}\{Y=1\}$ is the prevalence rate of a disease under investigation and $\boldsymbol{P}\{Y=0\}=1-\boldsymbol{P}\{Y=1\}$.

The Best Surrogate Split

For any split s^*, split s' is the best surrogate split of s^* when s' yields the greatest coincidence probability with s^* over all allowable splits based on different predictors.

Surrogate Split

- It is possible that the predictor that yields the best surrogate split may also be missing.
- We have to look for the second best, and so on.
- If our purpose is to build an automatic classification rule (e.g., Goldman et al., 1982, 1996), it is not difficult for a computer to keep track of the list of surrogate splits.
- However, the same task may not be easy for humans.
- Surrogate splits are rarely published in the literature.

Surrogate Split

- There is no guarantee that surrogate splits improve the predictive power of a particular split as compared to a random split. In such cases, the surrogate splits should be discarded.
- If surrogate splits are used, the user should take full advantage of them. They may provide alternative tree structures that in principle can have a lower misclassification cost than the final tree, because the final tree is selected in a stepwise manner and is not necessarily a local optimizer in any sense.

Tree Stability

- If we take a random sample of 3861 with replacement from the Yale Pregnancy Outcome Study, what is the chance that we come to the same tree as the original one?
- This chance is not so great, as all stepwise model selections potentially suffer from the same problem.
- While the trees structures are instable, the trees could provide very similar classifications and predictions.

Tree for Treatment Effectiveness

In a typical randomized clinical trial, different treatments (say two treatments) are compared in a study population, and the effectiveness of the treatments is assessed by averaging the effects over the treatment arms. However, it is possible that the on-average inferior treatment is superior in some of the patients. The trees provide a useful framework to explore this possibility by identifying patient groups within which the treatment effectiveness varies the greatest among the treatment arms.

Splitting Criterion

- We need to replace the impurity with the Kullback–Leibler divergence (Kullback and Leibler 1951).
- Let $p_{y,i}(t) = P(Y = y|t, \text{Trt} = i)$ be the probability that the response is y when a patient in node t received the i-th treatment. Then, the Kullback–Leibler divergence within node t is $\sum_y p_{y,1} \log(p_{y,1}/p_{y,2})$.
- Note that the Kullback–Leibler divergence is not symmetric with respect to the role of $p_{y,1}$ and $p_{y,2}$, but it is easy to symmetrize it as follows:

$$D_{KL}(t) = \sum_y p_{y,1} \log(p_{y,1}/p_{y,2}) + \sum_y p_{y,2} \log(p_{y,2}/p_{y,1}).$$

- A simpler and more direct measure is the difference

$$DIFF(t) = \sum_y (p_{y,1} - p_{y,2})^2.$$

Splitting Criterion

It is noteworthy that neither D_{KL} nor $DIFF$ is a distance metric and hence does not possess the property of triangle inequality. Consequently, the result does not necessarily improve as we split a parent node into offspring nodes.

Limitations of Trees

- Tree-based data analyses are readily interpretable.
- Tree-based methods have their limitations.
  - Tree structure is prone to instability even with minor data perturbations.
  - To leverage the richness of a data set of massive size, we need to broaden the classic statistical view of “one parsimonious model" for a given data set.
  - Due to the adaptive nature of the tree construction, theoretical inference based on a tree is usually not feasible. Generating more trees may provide an empirical solution to statistical inference.

Random Forests

- Forests have emerged as an ideal solution.
- A forest refers to a constellation of any number of tree models. Such an approach is also referred to as an *ensemble*.
- A forest consists of hundreds or thousands of trees, so it is more stable and less prone to prediction errors as a result of data perturbations (Breiman 1996, 2001).
- While each individual tree is not a good model, combining them into a committee improves their value.
- Trees in a forest should not be pruned; otherwise it would be counterproductive to pool “good" models into a committee.

Random Forests

- Suppose we have n observations and p predictors.

1. Draw a bootstrap sample.
2. Apply recursive partitioning to the bootstrap sample. At each node, randomly select q of the p predictors and restrict the splits based on the random subset of the q variables. Here, q should be much smaller than p.
3. Let the recursive partitioning run to the end and generate a tree.
4. Repeat Steps 1 to 3 to form a forest. The forest-based classification is made by majority vote from all trees.

Random Forests

- If Step 2 is skipped, the above algorithm is called *bagging* (bootstraping and aggregating) (Breiman 1996).
- Bagging should not be confused with another procedure called *boosting* (Freund and Schapire 1996).
  - One of the boosting algorithms is *Adaboost*, which makes use of two sets of intervening weights.
  - One set, w, weighs the classification error for each observation.
  - The other, β, weighs the voting of the class label.
  - Boosting is an iterative procedure, and at each iteration, a model (e.g., a tree) is built.
  - It begins with an equal w-weight for all observations.
  - Then, the β-weights are computed based on the w-weighted sum of error, and w-weights are updated with β-weights.
  - With the updated weights, a new model is built and the process continues.

Forest Size

- How many trees do we need in a forest?
- Because of so many trees in a forest, it is impractical to present a forest or interpret a forest.
- Zhang and Wang (2009): a tree is removed if its removal from the forest has the minimal impact on the overall prediction accuracy.
  - Calculate the prediction accuracy of forest F, denoted by p_F.
  - For every tree, denoted by T, in forest F, calculate the prediction accuracy of forest F_{-T} that excludes T, denoted by $p_{F_{-T}}$.
  - Let Δ_{-T} be the difference in prediction accuracy between F and F_{-T}: $\Delta_{-T} = p_F - p_{F_{-T}}$.
  - The tree T^p with the smallest Δ_T is the least important one and hence subject to removal: $T^p = \arg\min_{T \in F}(\Delta_{-T})$.

Optimal Size Subforest

- Let $h(i), i = 1, \ldots, N_f - 1$, denote the performance trajectory of a subforest of i trees, where N_f is the size of the original random forest.
- If there is only one realization of $h(i)$, they select the optimal size i_{opt} of the subforest by maximizing $h(i)$ over $i = 1, \ldots, N_f - 1$: $i_{opt} = \underset{i=1,\ldots,N_f-1}{\arg\max} \ (h(i))$.
- If there are M realizations of $h(i)$, they select the optimal size subforest by using the 1-se.

Optimal Size Subforest

- Compute the average $\overline{h}(i)$ and its standard error $\hat{\sigma}(i)$:
  $\overline{h}(i) = \frac{1}{M}\sum_{j=1,\ldots,M} h_j(i), i = 1,\ldots,N_f - 1,$
  $\hat{\sigma}(i) = \mathrm{var}(h_1(i),\ldots,h_M(i)), i = 1,\ldots,N_f - 1.$
- Find the i_m that maximizes the average $\overline{h}(i)$ over $i = 1,\ldots,N_f - 1$:
  $i_m = \underset{i=1,\ldots,N_f-1}{\arg\max}\ (\overline{h}(i)).$
- Choose the smallest subforest such that its corresponding $\overline{h}$ is within one standard error (se) of $\overline{h}(i_m)$ as the optimal subforest size i_{opt}: $i_{opt} = \min_{i=1,\ldots,M}(h(i) > (\overline{h}(i_m) - \hat{\sigma}(i_m)).$

Breast Cancer Prognosis

- van de Vijver et al. (2002): the microarray data set of a cohort of 295 young patients with breast cancer, containing expression profiles from 70 previously selected genes.
- The responses of all patients are defined by whether the patients remained disease-free five years after their initial diagnoses or not.
- To begin the process, an initial forest is constructed using the whole data set as the training data set.
- One bootstrap data set is used for execution and the out-of-bag (oob) samples for evaluation.
- Replicating the bootstrapping procedure 100 times, Zhang and Wang (2009) found that the sizes of the optimal subforests fall in a relatively narrow range, of which the 1st quartile, the median, and the 3rd quartile are 13, 26, and 61, respectively. This allows them to choose the smallest optimal subforest with the size of 7.

Comparison of Prediction Performance

Method	Error rate	Predicted outcome	Observed outcome Good	Poor
Initial random forest	26.0%	Good	141	17
		Poor	53	58
Optimal subforest	26.0%	Good	146	22
		Poor	48	53
Published classifier	35.3%	Good	103	4
		Poor	91	71

Importance Score

- Unlike a tree, a forest is generally too overwhelming to interpret.
- Summarize or quantify the information in the forest, for example, by identifying “important" predictors in the forest.
- If important predictors can be identified, a random forest can also serve as a method of variable (feature) selection.
- We can utilize other simpler methods such as classification trees by focusing on the important predictors.
- How do we know a predictor is important?

Gini Importance

- During the course of building a forest, whenever a node is split based on variable k, the reduction in Gini index from the parent node to the two daughter nodes is added up for variable k.
- Do this for all trees in the forest, giving rise to a simple variable importance score.
- Although Breiman noted that Gini importance is often very consistent with the permutation importance measure, others found it undesirable for being in favor of predictor variables with many categories (see, e.g., Strobl et al. 2007).

Depth Importance

- Chen et al. (2007) introduced an importance index that is similar to Gini importance score, but considers the location of the splitting variable as well as its impact.
- Whenever node t is split based on variable k, let $L(t)$ be the depth of the node and $S(k,t)$ be the χ^2 test statistic from the variable, then $2^{-L(t)}S(k,t)$ is added up for variable k over all trees in the forest.
- The depth is 1 for the root node, 2 for the offspring of the root node, and so forth.
- This depth importance measure was found useful in identifying genetic variants for complex diseases, although it is not clear whether it also suffers from the same end-cut preference problem.

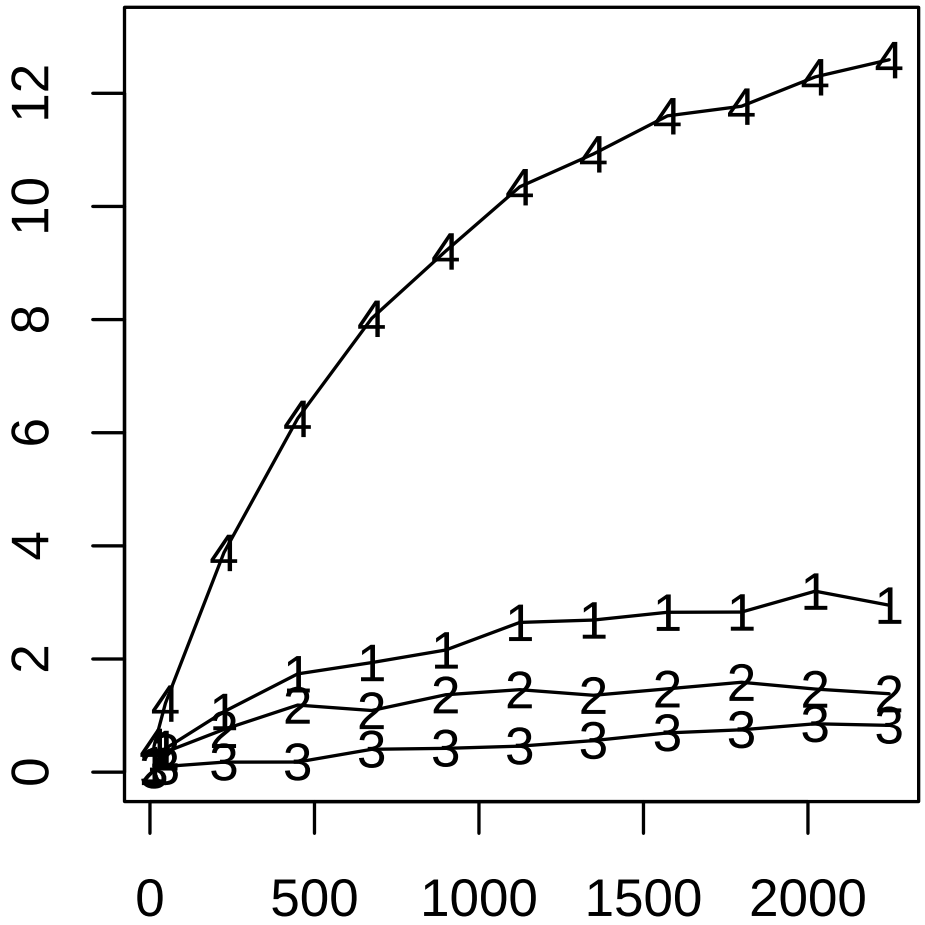
Permutation Importance

- Also referred to as the variable importance.
- For each tree in the forest, we count the number of votes cast for the correct class.
- We randomly permute the values of variable k in the oob cases and recount the number of votes cast for the correct class in the oob cases with the permuted values of variable k.
- Average the differences between the number of votes for the correct class in the variable-k-permuted oob data from the number of votes for the correct class in the original oob data, over all trees in the forest.

Permutation Importance

- Arguably the most commonly used choice.
- Not necessarily positive, and does not have an upper limit.
- Both the magnitudes and relative rankings of the permutation importance for predictors can be unstable when the number, p, of predictors is large relative to the sample size.
- The magnitudes and relative rankings of the permutation importance for predictors vary according to the number of trees in the forest and the number, q, of variables that are randomly selected to split a node.

Permutation Importance



Permutation Importance

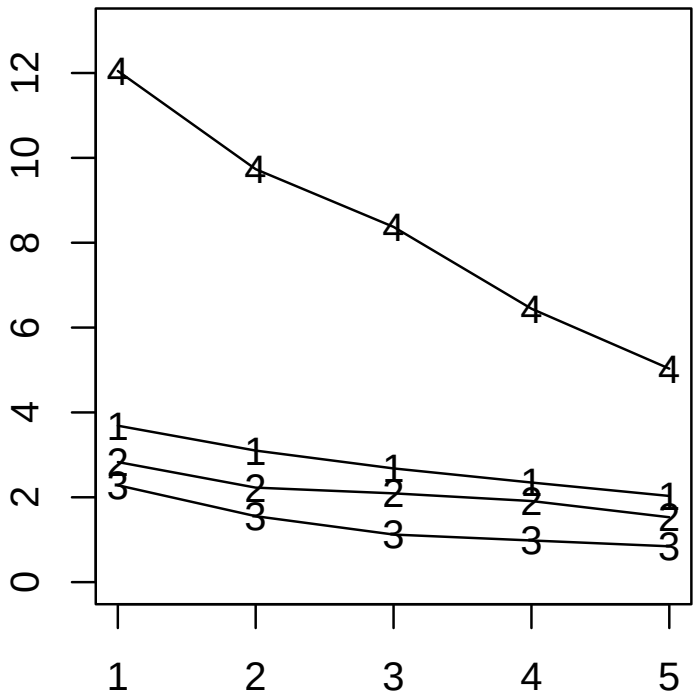
- there are conflicting numerical reports with regard to the possibility that the permutation importance overestimates the variable importance of highly correlated variables.
- Genuer et al. (2008): specifically addressed this issue with simulation studies and concluded that the magnitude of the importance for a predictor steadily decreases when more variables highly correlated with the predictor are included in the data set.

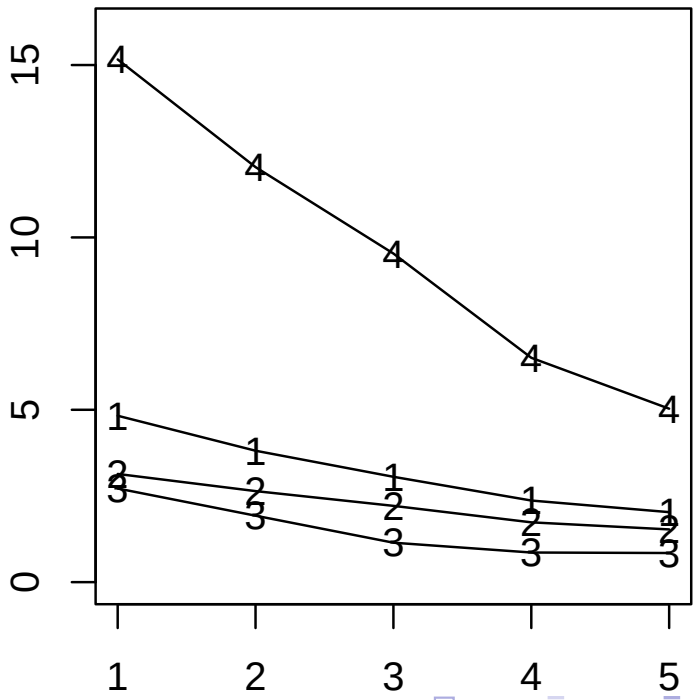
Permutation Importance

- Began with the four selected genes.
- Identified the genes whose correlations with any of the four selected genes are at least 0.4.
- Those correlated genes are divided randomly in five sets of about same size.
- We added one, two, . . . , and five sets of them sequentially together with the four selected genes as the predictors.

Permutation Importance

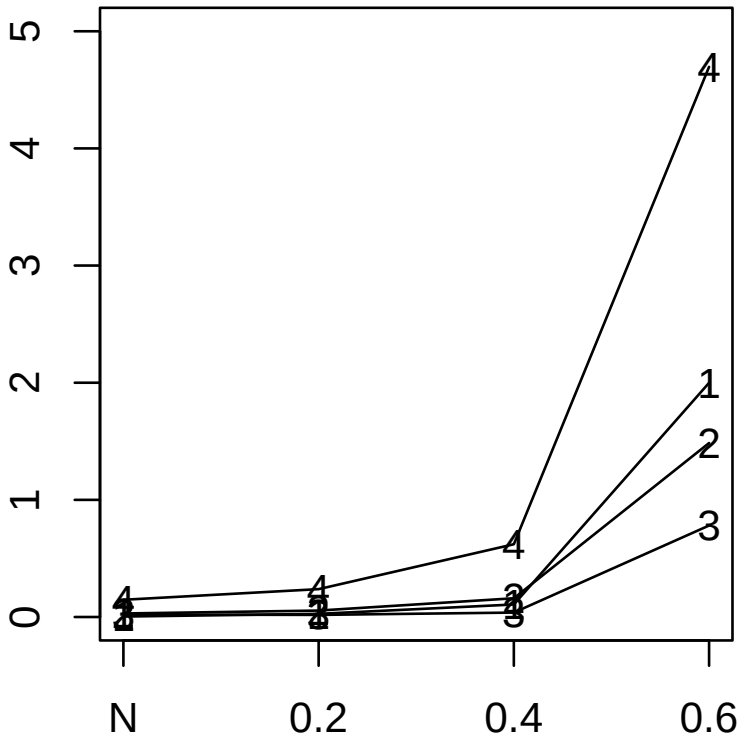
- The x-axis is the number of correlated sets of genes and the y-axis the importance score.
- The forest size is set at 1000.
- q equals the square root of the forest size for the left panel and 8 for the right panel.
- The rankings of the predictors are preserved.

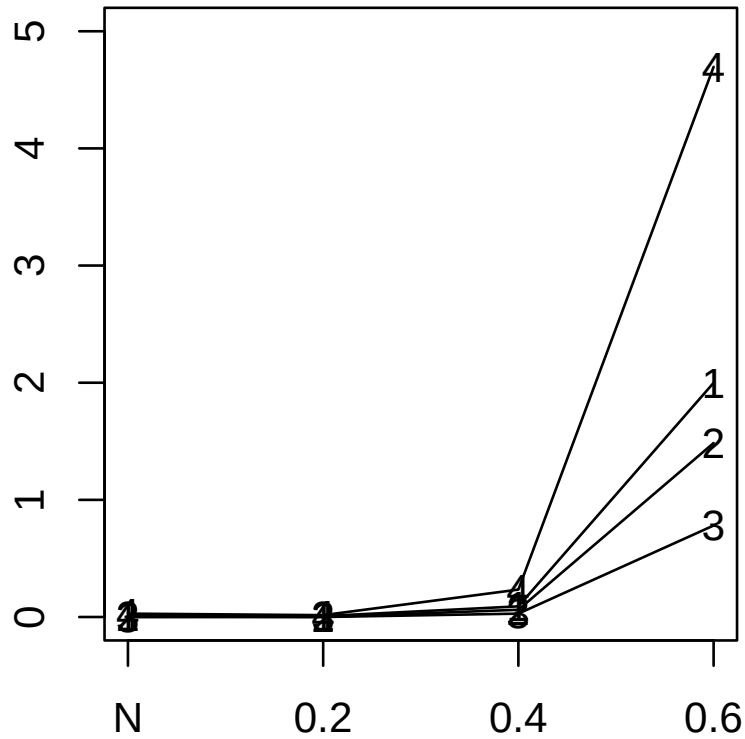




Permutation Importance

- Included genes that are correlated with any of the correlated gene at least 0.6, 0.4, and 0.2.
- The ranking is more relevant than the magnitude





Maximum Conditional Importance

- Wang et al. (2010): introduced a maximal conditional chi-square (MCC) importance by taking the maximum chi-square statistic resulting from all splits in the forest that use the same predictor.
- MCC can distinguish causal predictors from noise.
- MCC can assess interactions.
  - Consider the interaction between two predictors x_i and x_j.
  - For x_i, suppose its MCC is reached in node t_i of a tree within a forest. Whenever x_j splits an ancestor of node t_i, we count one and otherwise zero.
  - The final frequency, f, can give us a measure of interaction between x_i and x_j.
  - Through the replication of the forest construction we can estimate the frequency and its precision.

Maximum Conditional Importance

- They generated 100 predictors independently, each of them is the sum of two i.i.d. binary variables (0 or 1).
- For the first 16 predictors, the underlying binary random variable has the success probability of 0.282.
- For the remaining 84, they draw a random number between 0.01 and 0.99 as the success probability of the underlying binary random variable.
- The first 16 predictors serve as the risk variables and the remaining 84 the noise variables.

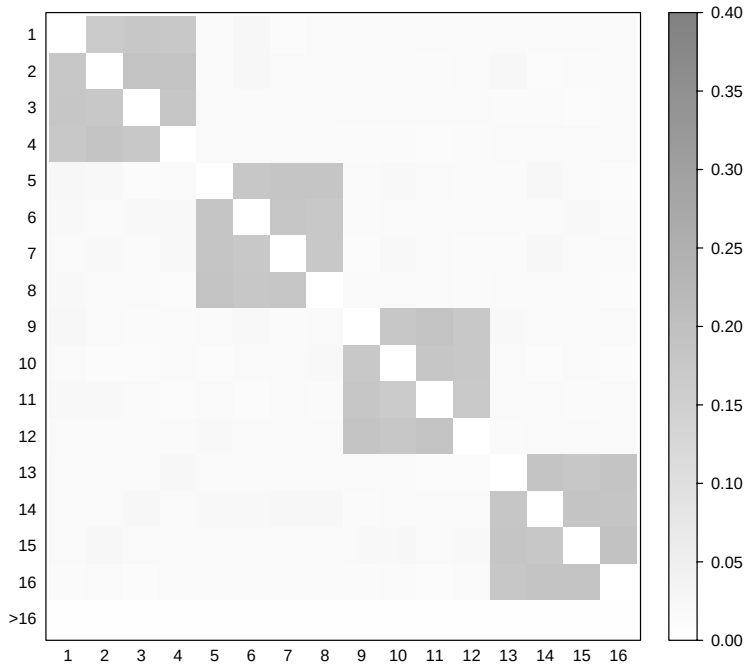
Maximum Conditional Importance

- The outcome variable is generated as follows.
- The 16 risk variables are divided equally into four groups, and without loss of generality, say sequentially.
- Once these 16 risk variables are generated, we calculate the following probability on the basis of which the response variable is generated: $w = 1 - \Pi(1 - \Pi q_k)$ where the first product is with respect to the four groups, the second product is with respect to the first predictors inside each group, and $q_0 = 1.2 \times 10^{-8}$, $q_1 = 0.79$, and $q_2 = 1$. The subscript k equals the randomly generated value of the respective predictor.

Maximum Conditional Importance

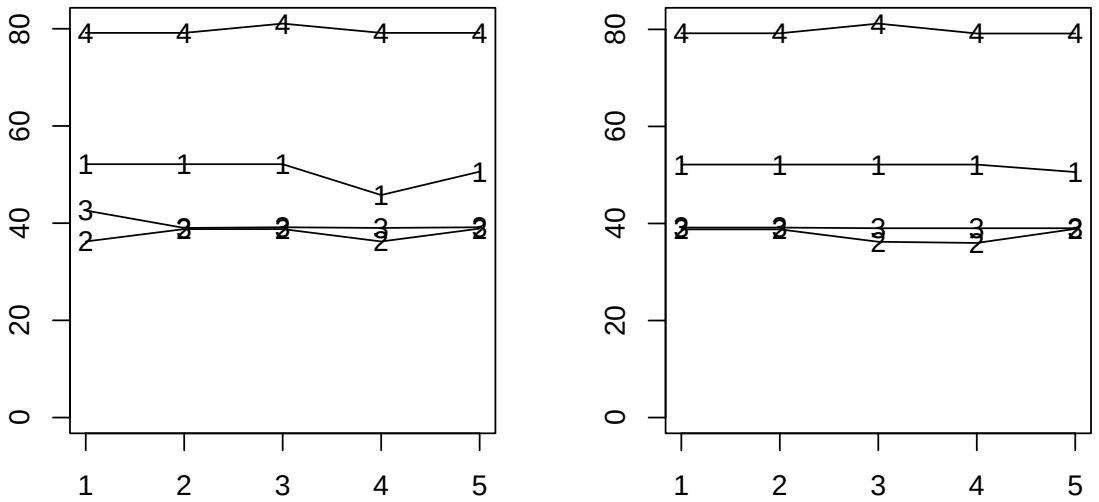
- Generate the first 200 possible controls and the first 200 possible cases.
- This completes the generation of one data set.
- Replicate the entire process 1000 times.

Interaction Heat Map



The x-axis is the sequence number of the primary predictor and the y-axis the sequence number of the potential interacting predictor. The intensity expresses the frequency when the potential interacting predictor precedes the primary predictor in a forest.

Impact on MCC by Correlated Predictors



Predictors with Uncertainties

- In general, we base our analysis on predictors that are observed with certainty.
- However, this is not always the case.
  - To identify genetic variants for complex diseases, haplotypes are sometimes the predictors.
  - A haplotype is a combination of single nucleotide polymorphisms(SNPs) on a chromatid.
  - Has to be statistically inferred from the SNPs in frequencies.

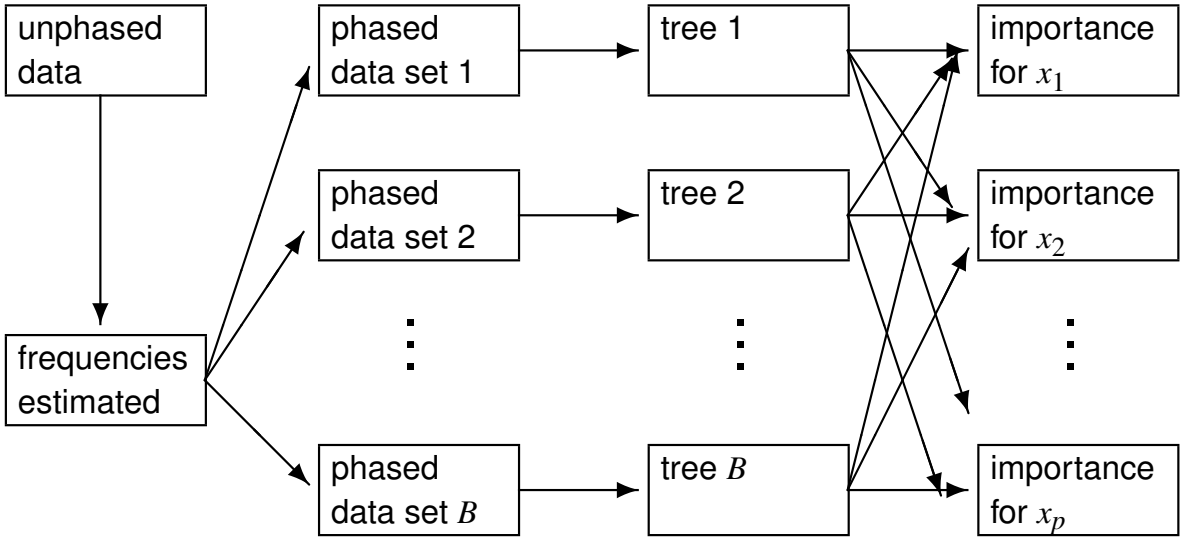
Predictors with Uncertainties

- We assume x_1 is the only categorical variable with uncertainties, and it has K possible levels.
- For the i-th subject, $x_{i1} = k$ with a probability p_{ik} $(\sum_{k=1}^{K} p_{ik} = 1)$.
  - To identify genetic variants for complex diseases, haplotypes are sometimes the predictors.
  - A haplotype is a combination of single nucleotide polymorphisms(SNPs) on a chromatid.
  - Has to be statistically inferred from the SNPs in frequencies.

Predictors with Uncertainties

- In a typical random forest, the “working" data set is a bootstrap sample of the original data set.
- Here, a “working" data set is generated according to the frequencies of x_1 while keeping the other variables intact.
- the data set would be $\{z_{i1}, x_{i2}, \ldots, x_{ip}, y_i\}_{i=1}^{n}$, where z_{i1} is randomly chosen from $1, \ldots, K$, according to the probabilities $(p_{i1}, \ldots, p_{iK})$.
- Once the data set is generated, the rest can be carried out in the same way as for a typical random forest.

Predictors with Uncertainties



Notation

- Let δ indicate whether a subject's survival is observed (one if it is) or censored (zero if it is not).
- Let Y denote the observed time.
- In the absence of censoring, the observed time is the survival time, and hence $Y = T$.
- Otherwise, the observed time is the censoring time, denoted by U.
- $Y = \min(T, U)$ and $\delta = I(Y = T)$, where $I(\cdot)$ is an indicator function.

Data Example

Smoked	Time (days)	Smoked	Time	Smoked	Time
yes	11906+	yes	9389+	yes	4539+
yes	11343+	yes	9515+	yes	10048+
yes	5161	yes	9169	no	8147+
yes	11531+	yes	11403+	yes	11857+
yes	11693+	no	10587	yes	9343+
yes	11293+	yes	6351+	yes	502+
yes	7792	no	11655+	yes	9491+
yes	2482+	no	10773+	yes	11594+
no	7559+	yes	11355+	yes	2397
yes	2569+	yes	2334+	yes	11497+
yes	4882+	yes	9276	yes	703+
yes	10054	no	11875+	no	9946+
yes	11466+	no	10244+	yes	11529+
yes	8757+	no	11467+	yes	4818
yes	7790	yes	11727+	no	9552+
yes	11626+	yes	7887+	yes	11595+
yes	7677+	yes	11503	yes	10396+
yes	6444+	yes	7671+	yes	10529+
yes	11684+	yes	11355+	yes	11334+
yes	10850+	yes	6092	yes	11236+

Survival and Hazard Function

- Survival function
$$S(t) = \boldsymbol{P}\{T > t\}.$$
- Hazard function
$$h(t) = \frac{\lim_{\Delta t \to 0} \boldsymbol{P}\{T \in (t, t+\Delta t)\}/\Delta t}{\boldsymbol{P}\{T > t\}}.$$
- The hazard function is an instantaneous failure rate in the sense that it measures the chance of an instantaneous failure per unit of time given that an individual has survived beyond time t.

Estimate Survival Function

- Parametric Approach: distributions of survival can be assumed.
  - Exponential: $S(t) = \exp(-\lambda t)$ $(\lambda > 0)$, where λ is an unknown constant.
  - Only need to estimate the constant hazard.
  - The full likelihood function

$$L(\lambda) = \prod_{i=1}^{60} [\lambda \exp(-\lambda T_i)]^{\delta_i} [\exp(-\lambda U_i)]^{1-\delta_i}.$$

Estimate Survival Function

- For the sample data, the log likelihood function

$$\begin{aligned} l(\lambda) &= \sum_{i=1}^{60} \{\delta_i [\log(\lambda) - \lambda Y_i] - \lambda(1-\delta_i) Y_i\} \\ &= \log(\lambda) \sum_{i=1}^{60} \delta_i - \lambda \sum_{i=1}^{60} T_i \\ &= 11 \log(\lambda) - \lambda(11906 + 11343 + \cdots + 11236), \end{aligned}$$

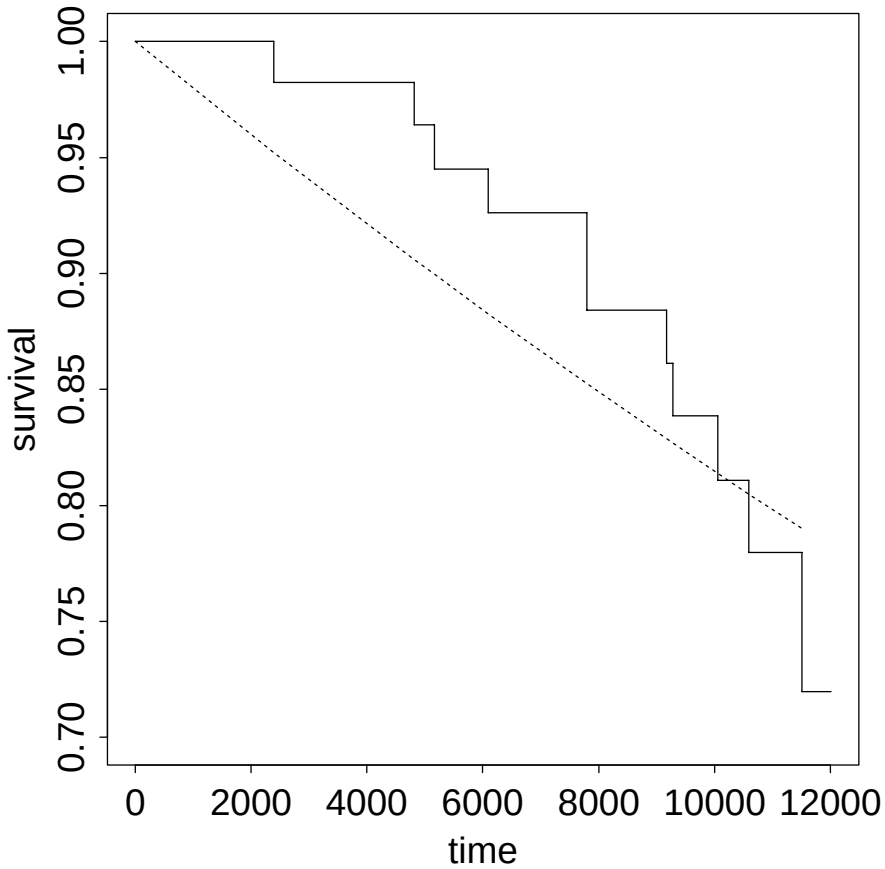
  where 11 is the number of uncensored survival times and the summation is over all observed times.

- The maximum likelihood estimate of the hazard, λ, is $\hat{\lambda} = \frac{11}{527240} = 2.05/10^5$, which is the number of failures divided by the total observed time.

Validate Survival Function

- Compare a parametric fit with the nonparametric Kaplan–Meier Curve.
- Plot the empirical cumulative hazard function against the assumed theoretical cumulative hazard function at times when failures occurred.
  - The cumulative hazard function is defined as $H(t) = \int_0^t h(u)du$.

Exponential and Kaplan–Meier Curves



Cumulative Hazard Functions

Survival time	Risk set K	Failures d	Hazard rate d/K	Cumulative hazard Empirical	Assumed
2397	57	1	0.0175	0.0175	0.0491
4818	53	1	0.0189	0.0364	0.0988
5161	51	1	0.0196	0.0560	0.1058
6092	50	1	0.0200	0.0760	0.1249
7790	44	1	0.0227	0.0987	0.1597
7792	43	1	0.0233	0.1220	0.1597
9169	39	1	0.0256	0.1476	0.1880
9276	38	1	0.0263	0.1740	0.1902
10054	30	1	0.0333	0.2073	0.2061
10587	26	1	0.0385	0.2458	0.2170
11503	13	1	0.0769	0.3227	0.2358

Product Limit Estimate of Survival Function

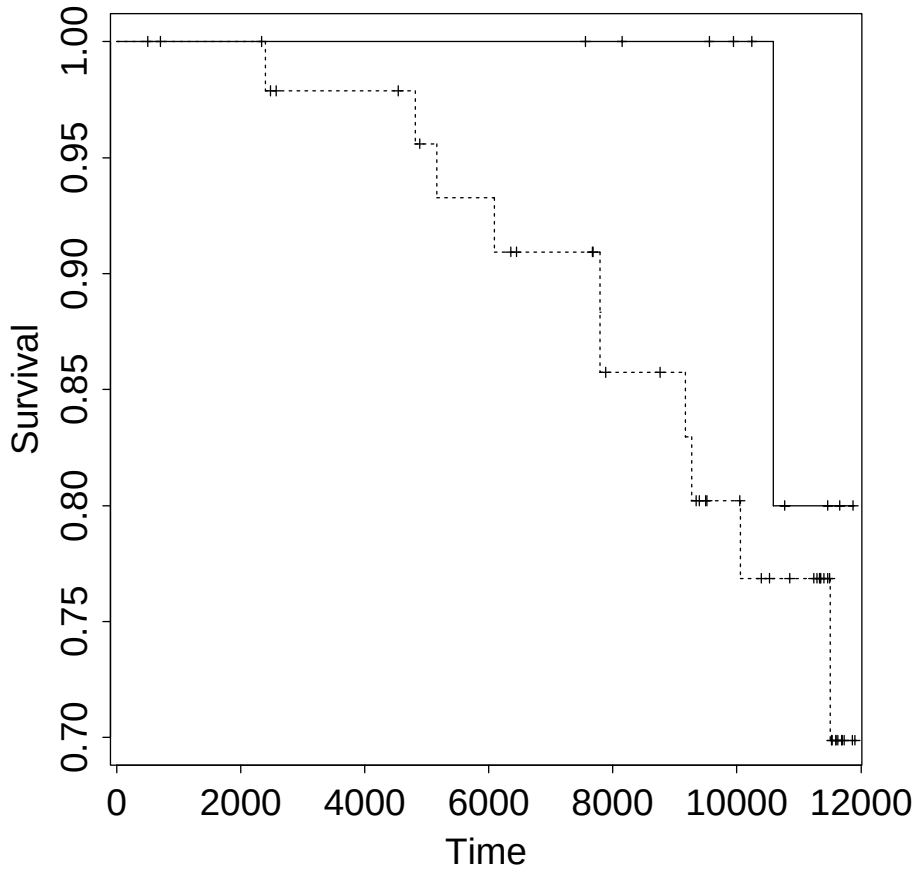
Survival time	Risk set K	Failures d	Ratio $(K-d)/K$	Product $\hat{S}(t)$
2397	57	1	0.982	0.982
4818	53	1	0.981	$0.982 * 0.981 = 0.963$
5161	51	1	0.980	$0.963 * 0.980 = 0.944$
6092	50	1	0.980	$0.944 * 0.980 = 0.925$
7790	44	1	0.977	$0.925 * 0.977 = 0.904$
7792	43	1	0.977	$0.904 * 0.977 = 0.883$
9169	39	1	0.974	$0.883 * 0.974 = 0.860$
9276	38	1	0.974	$0.860 * 0.974 = 0.838$
10054	30	1	0.967	$0.838 * 0.967 = 0.810$
10587	26	1	0.962	$0.810 * 0.962 = 0.779$
11503	13	1	0.923	$0.779 * 0.923 = 0.719$

Cumulative Hazard vs Kaplan–Meier Curve

- The mechanism of producing the Kaplan–Meier curve is similar to the generation of the empirical cumulative hazard function.
- The first three columns are the same.
- The fourth columns differ by one, namely, the proportion of individuals who survived beyond the given time point.

Log-Rank Test

In many clinical studies, a common goal is to compare the survival distributions of various groups.



Log-Rank Test

Peto and Peto (1972): at the distinct failure times, we have a sequence of 2×2 tables.

	Dead	Alive	
Smoking	a_i		n_i
Nonsmoking			
	d_i		K_i

Time T_i	Risk set K_i	Failures d_i	a_i	n_i	E_i	V_i
2397	57	1	1	47	0.825	0.145
4818	53	1	1	43	0.811	0.153
5161	51	1	1	41	0.804	0.158
6092	50	1	1	40	0.800	0.160
7790	44	1	1	35	0.795	0.163
7792	43	1	1	34	0.791	0.165
9169	39	1	1	31	0.795	0.163
9276	38	1	1	30	0.789	0.166
10054	30	1	1	24	0.800	0.160
10587	26	1	0	21	0.808	0.155
11503	13	1	1	11	0.846	0.130

The log-rank test statistic is $LR = \frac{\sum_{i=1}^{k}(a_i - E_i)}{\sqrt{\sum_{i=1}^{k} V_i}}$, where k is the number of distinct failure times, $E_i = \frac{d_i n_i}{K_i}$, and $V_i = \left(\frac{d_i(K_i - n_i)n_i}{K_i(K_i - 1)}\right)\left(1 - \frac{d_i}{K_i}\right)$.

Log-Rank Test

Time T_i	Risk set K_i	Failures d_i	a_i	n_i	E_i	V_i
2397	57	1	1	47	0.825	0.145
4818	53	1	1	43	0.811	0.153
5161	51	1	1	41	0.804	0.158
6092	50	1	1	40	0.800	0.160
7790	44	1	1	35	0.795	0.163
7792	43	1	1	34	0.791	0.165
9169	39	1	1	31	0.795	0.163
9276	38	1	1	30	0.789	0.166
10054	30	1	1	24	0.800	0.160
10587	26	1	0	21	0.808	0.155
11503	13	1	1	11	0.846	0.130

- The log-rank test statistic has an asymptotic standard normal distribution, we test the hypothesis that the two survival functions are the same by comparing LR with the quantiles of the standard normal distribution.
- For our data, $LR = 0.87$, corresponding to a two-sided p-value of 0.38.

Cox Proportional Hazard Regression

- Instead of making assumptions directly on the survival times, Cox (1972) proposed to specify the hazard function.
- Suppose that we have a set of predictors $\mathbf{x} = (x_1, \ldots, x_p)$.
- The Cox proportional hazard model is

$$\lambda(t; \mathbf{x}) = \exp(\mathbf{x}\boldsymbol{\beta})\lambda_0(t),$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters and $\lambda_0(t)$ is an unknown function giving a baseline hazard for $\mathbf{x} = \mathbf{0}$.

Cox Proportional Hazard Regression

- If we take two individuals i and j with covariates $\mathbf{x}_i$ and $\mathbf{x}_j$, the ratio of their hazard functions is $\exp((\mathbf{x}_i - \mathbf{x}_j)\boldsymbol{\beta})$, which is free of time.
- The hazard functions for any two individuals are parallel in time.
- $\lambda_0(t)$ is left to be arbitrary. Thus, the proportional hazard can be regarded as semiparametric.

Conditional Likelihood

- Condition the likelihood on the set of uncensored times.
- At any time t, let $\mathcal{R}(t)$ be the risk set, i.e., the individuals who were at risk right before time t. For each uncensored time T_i, the hazard rate is $h(T_i) = \boldsymbol{P}\{\text{A death in } (T_i, T_i + dt) \mid \mathcal{R}(T_i)\}/dt$.
- Under the proportional hazard model,

$$\boldsymbol{P}\{\text{A death in } (T_i, T_i + dt) \mid \mathcal{R}(T_i)\} = \exp(\mathbf{x}\boldsymbol{\beta})\lambda_0(T_i)dt$$

and

$$\begin{aligned}&\boldsymbol{P}\{\text{Individual } i \text{ fails at } T_i \mid \text{one death in } \mathcal{R}(T_i) \text{ at time } T_i\} \\ &= \frac{\exp(\mathbf{x}_i\boldsymbol{\beta})}{\sum_{j\in\mathcal{R}(T_i)}\exp(\mathbf{x}_j\boldsymbol{\beta})}.\end{aligned}$$

Conditional Likelihood

- The entire conditional likelihood is the product of those by failed individual i

$$L(\boldsymbol{\beta}) = \prod_{\text{failure } i} \frac{\exp(\mathbf{x}_i\boldsymbol{\beta})}{\sum_{j\in\mathcal{R}(T_i)} \exp(\mathbf{x}_j\boldsymbol{\beta})}.$$

- Maximizing the conditional likelihood gives rise to the estimates of $\boldsymbol{\beta}$.
- $\hat{\boldsymbol{\beta}}$ has an asymptotic normal distribution.

The Western Collaborative Group Study

- A prospective and long-term study of coronary heart disease.
- In 1960–61, 3154 middle-aged white males from ten large California corporations in the San Francisco Bay Area and Los Angeles entered the WCGS, and they were free of coronary heart disease and cancer.
- After a 33-year follow-up, 417 of 1329 deaths were due to cancer and 43 were lost to follow-up.

The Western Collaborative Group Study

Characteristics	Descriptive Statistics
Age	46.3 ± 5.2 years
Education	High sch. (1424), Col. (431), Grad. (1298)
Systolic blood pressure	128.6 ± 15.1 mmHg
Serum cholesterol	226.2 ± 42.9 (mg/dl)
Behavior pattern	Type A (1589), type B (1565)
Smoking habits	Yes (2439), No (715)
Body mass index	24.7 ± 2.7 (kg/m^2)
Waist-to-calf ratio	2.4 ± 0.2

The Western Collaborative Group Study

- We entered the eight predictors into an initial Cox's model and used a backward stepwise procedure to delete the least significant variable from the model at the threshold of 0.05.
- `coxph(Surv(time, cancer) age + chol + smoke + wcr).`

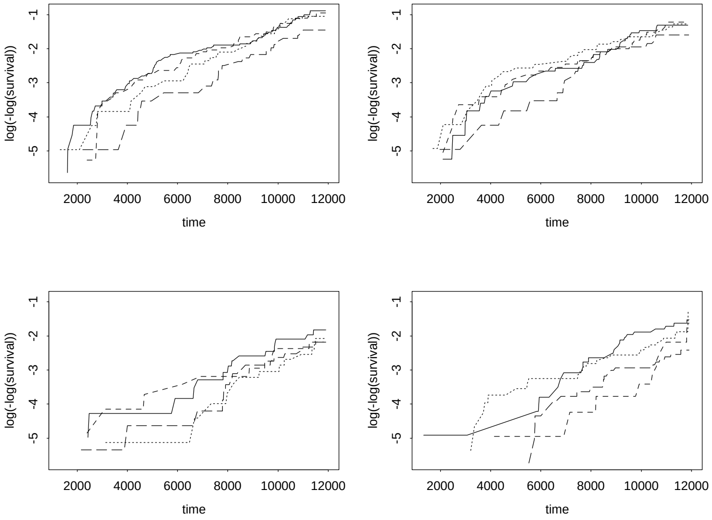
Parameter Estimation for Cox's Model

Variable	Coefficient	S.E.	p-value
Age (`age`)	0.0934	0.009	0.000
Serum cholesterol (`chol`)	0.0026	0.001	0.033
Smoking habits (`smoke`)	0.2263	0.103	0.029
Waist-to-calf ratio (`wcr`)	0.7395	0.271	0.006

Assessing Proportionality

- Dichotomize age, serum cholesterol, and waist-to-calf ratio at their median levels.
- The 2882 $(= 3154 - 272)$ subjects are divided into 16 cohorts
- Within each cohort i, we calculate the Kaplan–Meier survival estimate $\hat{S}_i(t)$.
- Plot $\log(-\log(\hat{S}_i(t)))$ versus time as shown in Figure **??**.

Assessing Proportionality



Assessing Proportionality

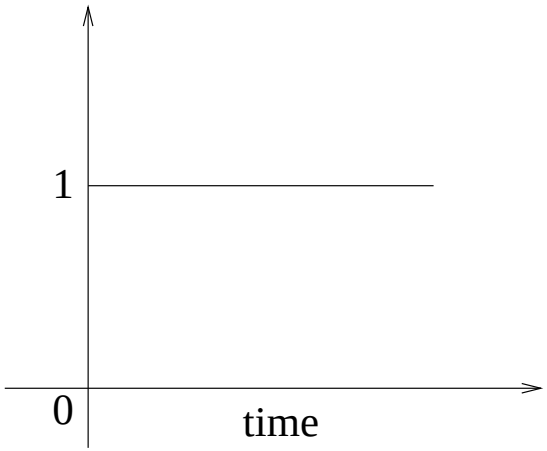
- $h(t) = -\frac{d\log(S(t))}{dt}$, which is equivalent to $S(t) = \exp\left(-\int_0^t h(z)dz\right)$.
- The survival function is

$$\begin{aligned} S(t;\mathbf{x}) &= \exp\left[-\int_0^t \exp(\mathbf{x}\boldsymbol{\beta})\lambda_0(z)\right] \\ &= \exp\left[-\exp(\mathbf{x}\boldsymbol{\beta})\int_0^t \lambda_0(z)\right]. \end{aligned} \tag{1}$$

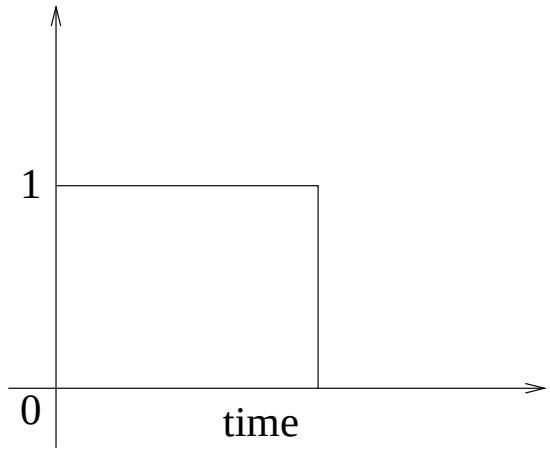
- $\log(-\log[S(t;\mathbf{x})]) = \mathbf{x}\boldsymbol{\beta} + \log\left[\int_0^t \lambda_0(z)\right]$.
- The log-log survival curves in our 16 cohorts are supposed to be parallel if the proportional hazard assumption holds.

Gordon and Olshen's Rule

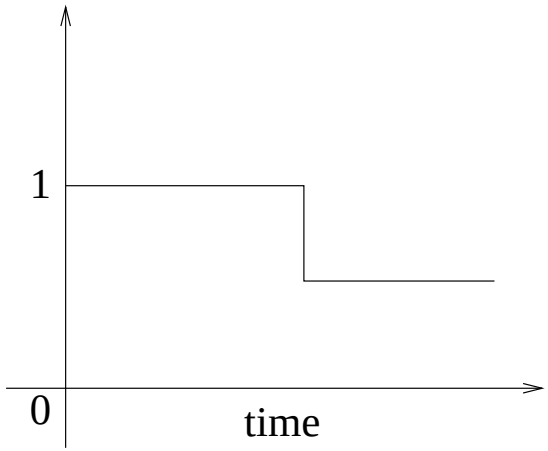
- Gordon and Olshen (1985): What would be an appropriate measure of node impurity in the context of censored data?
- We would regard a node as pure if all failures in the node occurred at the same time.



(a)



(b)



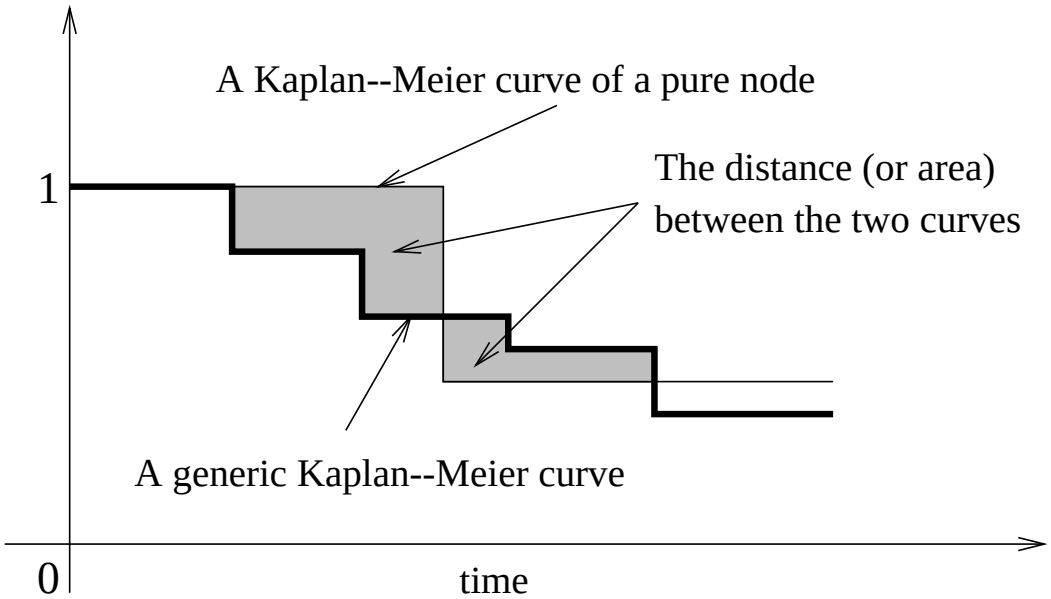
(c)

Gordon and Olshen's Rule

- How far the within-node Kaplan–Meier curve deviates from any of the curves in $\mathcal{P}$.
- Need first to define a distance between the two Kaplan–Meier curves.
- L^p Wasserstein metrics $\left[\int_0^1 |F_1^{-1}(u) - F_2^{-1}(u)|^p du\right]^{1/p}$, where $F_i^{-1}(u) = \min\{t : F_i(t) \geq u\}, i = 1, 2.$

Wasserstein Metrics for Survival Functions

$\left[\int_0^m |F_1^{-1}(u) - F_2^{-1}(u)|^p du\right]^{1/p}$, where the upper limit of the integral m is the minimum of m_1 and m_2.



Node Impurity

- $i(\tau) = \min_{\delta_S \in \mathcal{P}} d_p(S_\tau, \delta_S)$, where S_τ is the Kaplan–Meier curve within node τ, and the minimization $\min_{\delta_S \in \mathcal{P}}$ means that S_τ is compared with its best match among the three curves.
- When $p = 1$, this can be viewed as the deviation of survival times toward their median.
- When $p = 2$, it corresponds to the variance of the Kaplan–Meier distribution estimate of survival.
- With this node impurity, we can grow a survival tree.

Maximizing the Difference

- When two daughter nodes are relatively pure, they tend to differ from each other.
- Finding two different daughters is a means to increase the between variation and consequently to reduce the within variation.
- Select a split that maximizes the “difference” between the two daughter nodes, or, equivalently, minimizes their similarity.
- Ciampi et al. (1986) and Segal (1988): The log-rank test is a commonly used approach for testing the significance of the difference between the survival times of two groups.

Likelihood Functions

- Davis and Anderson (1989): assume that the survival function within any given node is an exponential function with a constant hazard.
- LeBlanc and Crowley (1992) and Ciampi et al. (1995): the hazard functions in two daughter nodes are proportional (the full or partial likelihood function can be used).
- All individuals in node τ have the hazard $\lambda_\tau(t) = \theta_\tau \lambda_0(t)$, where $\lambda_0(t)$ is the baseline hazard independent of the node and θ_τ is a nonnegative parameter corresponding to $\exp(\mathbf{x}\boldsymbol{\beta})$.
- Treat the value of the “covariate" as the same inside each daughter node, hence $\exp(\mathbf{x}\boldsymbol{\beta})$ becomes a single parameter θ_τ.

Likelihood Functions

- The survival function of individuals in node τ is $S(t;\tau) = \exp[-\theta_\tau \Lambda_0(t)]$, where $\Lambda_0(t)$ is the baseline cumulative hazard function integrated from $\lambda_0(t)$.
- The full likelihood function within node τ as $L(\theta_\tau, \lambda_0) = \prod_{\{i \in \text{ node } \tau\}} [\lambda_0(T_i)\theta_\tau]^{\delta_i} \exp[-\Lambda_0(U_i)\theta_\tau]$.
- The full likelihood of the entire learning sample for a tree $\mathcal{T}$ can be expressed as $L(\boldsymbol{\theta}, \lambda_0; \mathcal{T}) = \prod_{\tau \in \tilde{\mathcal{T}}} L(\theta_\tau, \lambda_0)$, which is the product of the full likelihoods contributed by all terminal nodes of $\mathcal{T}$.

Likelihood Functions

- Every time we partition a node into two, we need to maximize the full tree likelihood.
- Too ambitious for computation; e.g., the cumulative hazard Λ_0 is unknown and must be estimated over and over again.
- As a remedy, LeBlanc and Crowley propose to use a one-step Breslow's (1972) estimate $\hat{\Lambda}_0(t) = \frac{\sum_{i:Y_i \le t} \delta_i}{|\mathcal{R}(t)|}$, where the denominator $|\mathcal{R}(t)|$ is the number of subjects at risk at time t.
- The one-step estimate of θ_τ is then $\hat{\theta}_\tau = \frac{\sum_{\{i \in \text{ node } \tau\}} \delta_i}{\sum_{\{i \in \text{ node } \tau\}} \hat{\Lambda}_0(Y_i)}$, which can be interpreted as the number of failures divided by the expected number of failures in node τ under the assumption of no structure in survival times.

A Straightforward Extension

- Zhang (1995): we observe a binary death indicator, δ, and the observed time.
- Treat them as two outcomes, we can compute the within-node impurity, i_δ, of the death indicator and the within-node quadratic loss function, i_y, of the time.
- Combine $w_\delta i_\delta + w_y i_y$.

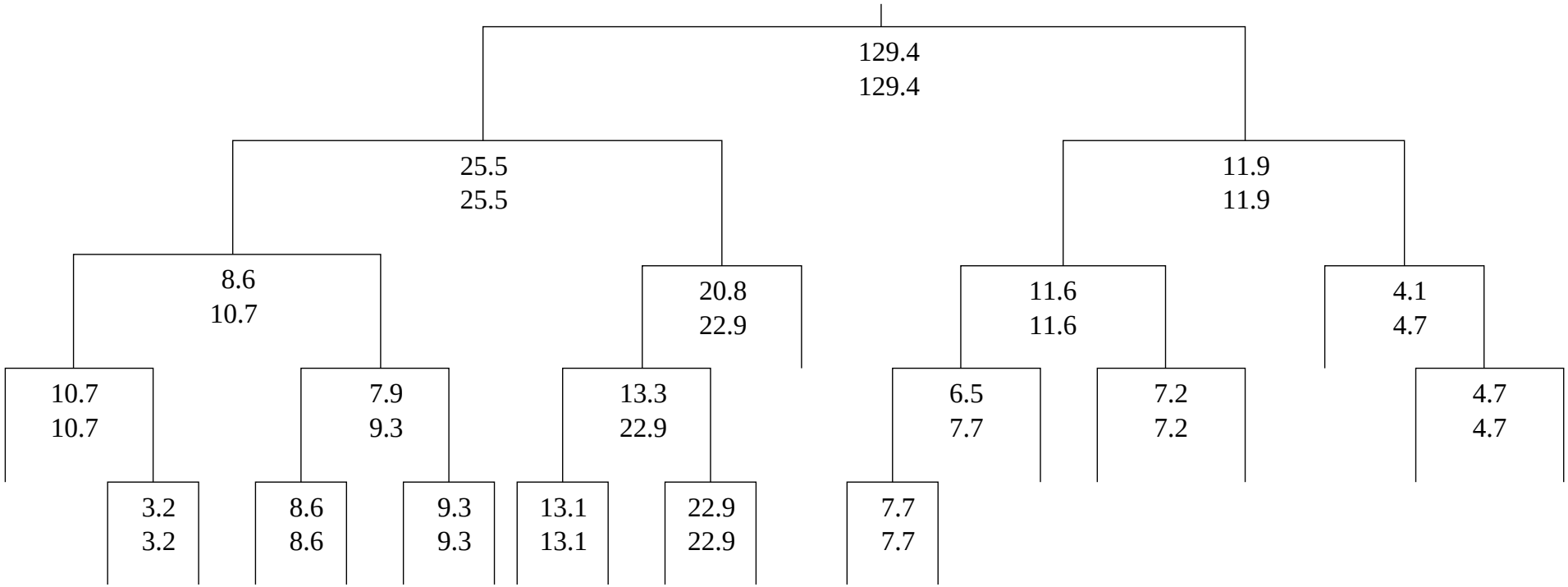
Pruning a Survival Tree

- Using any of the splitting criteria above, we can produce an initial tree.
- How do we prune the initial survival tree, $\mathcal{T}$?
- Cost-complexity $R_\alpha(\mathcal{T}) = R(\mathcal{T}) + \alpha|\tilde{\mathcal{T}}|$, where $R(\mathcal{T})$ is the sum of the costs over all terminal nodes of $\mathcal{T}$.

Random Survival Forests

- Once we are able to construct a survival tree, we can use the same method described above to construct a random survival forest (Ishwaran et al. 2008).

Survival Trees for the Western Collaborative Group Study



An initial large tree obtained by the log-rank testing statistic. The top and bottom numbers under the node are respectively the original and maximized values of the statistic.

Maximum Log-rank Statistic vs. Tree Size