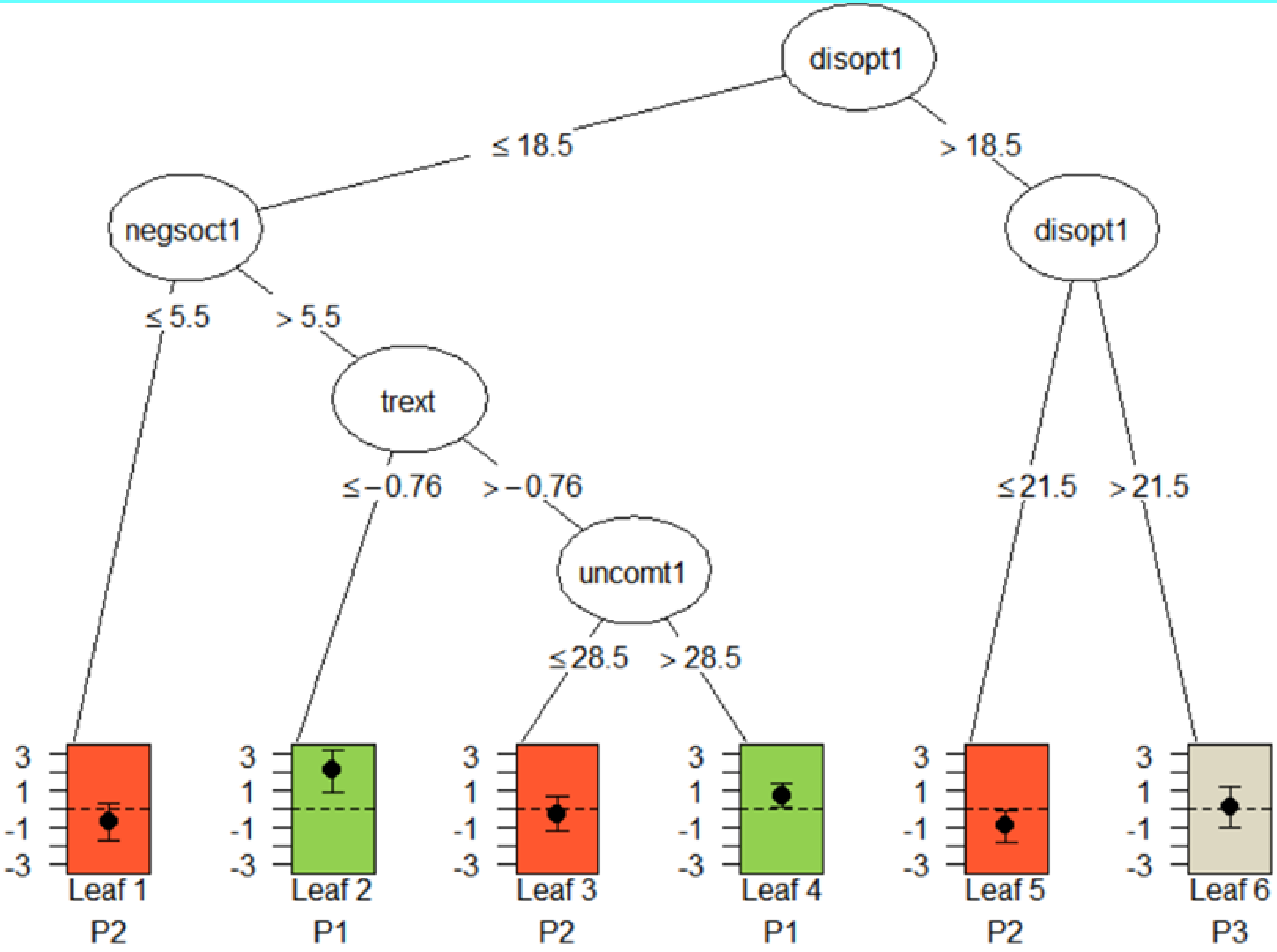
Recursive Partitioning



1. Regression Trees
2. Classification Trees
3. Qualitative Interaction Trees

Introduction to Recursive Partitioning

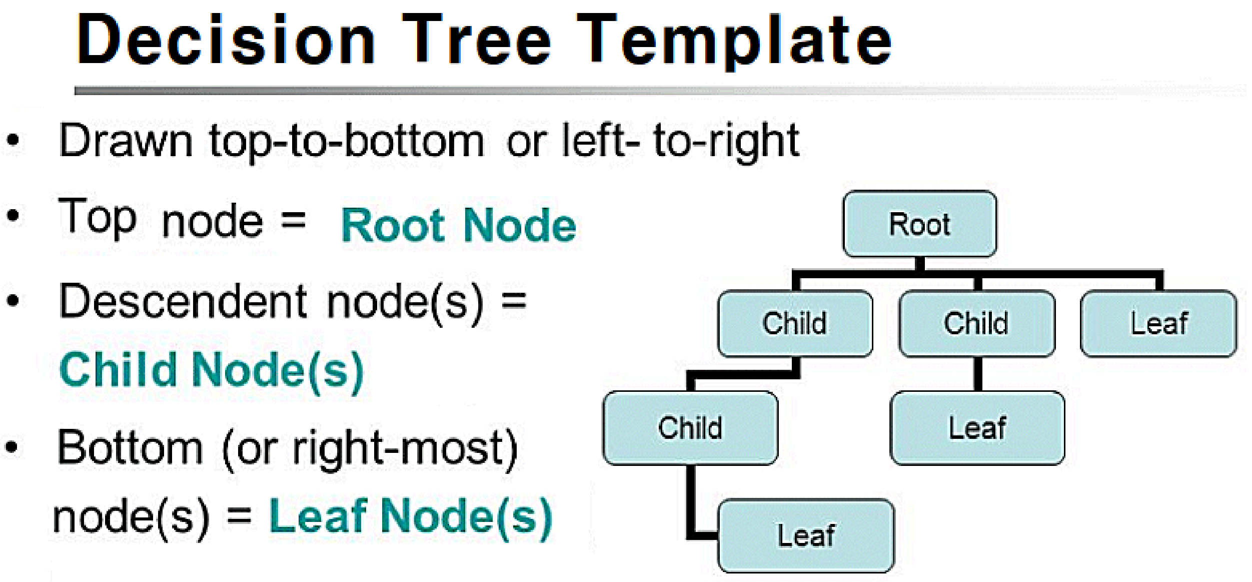
- Recursive partitioning creates a decision tree that strives to correctly classify members of the population by splitting it into sub-populations based on several dichotomous independent variables.
- The process is termed recursive because each sub-population may in turn be split an indefinite number of times until the splitting process terminates after a particular stopping criterion is reached.

[Reference] https://en.wikipedia.org/wiki/Recursive_partitioning

- Decision tree: Graph to represent choices and their results in the form of a tree.
- Graph: Node → event or choice

  Edges → decision rules or conditions.
- Mostly used in Machine Learning and Data Mining applications

https://www.tutorialspoint.com/r/r_decision_tree.htm



Decision Tree Template

- Drawn top-to-bottom or left- to-right
- Top node = **Root Node**
- Descendent node(s) = **Child Node(s)**
- Bottom (or right-most) node(s) = **Leaf Node(s)**

Generation of Decision Tree

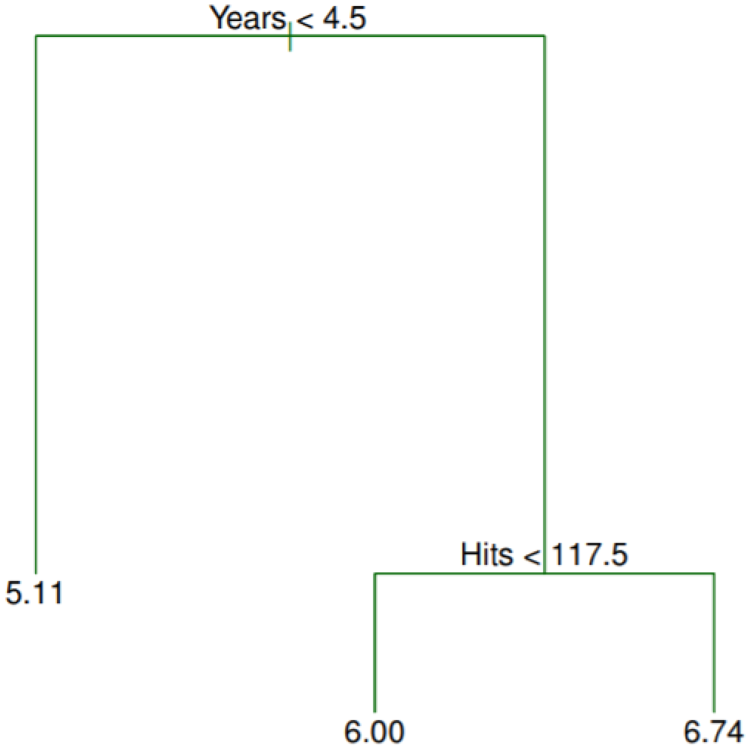
The algorithm to generate a decision tree is recursive.
Every iteration finds the best way to split a current training subset into two parts and get a dividing condition.

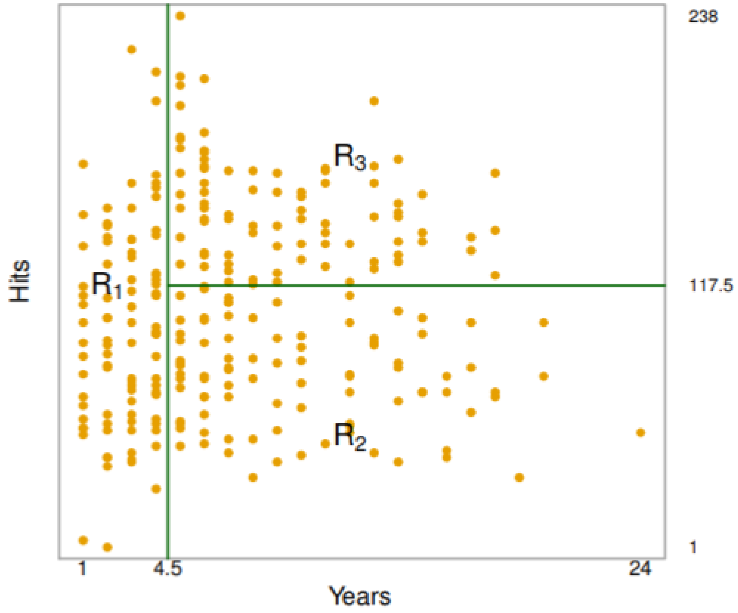
Splitting algorithms
1. Calculate RSS (residual sum of squares) for current subset
2. Sort items by feature value
3. For each possible division into two subsets, calculate RSS for left (L) and right (R) subset;
Find the division with min[RSS(L)+RSS(R)] and return feature name and value.

Criteria to stop recursion
1. RSS for current subset is equal to 0;
2. Items count in current subset is fewer than or equal to the acceptable maximum number of items in a terminal node.

http://www.c-sharpcorner.com/article/random-forest-machine-learning-technique/

Example: Predicting a baseball player's salary





The prediction for a point in region R_i is the average of the training points in R_i.

[Stanford Univ, STATS202 Data Mining and Analysis 2017]

Regression Trees

- The ability to represent the model as a tree is the key to its interpretability and popularity.

- What does it mean to fit a tree?

If we adopt the least squares criterion as our objective, then our estimate for C_m is simply the average of the y_i's in that region:

$$\hat{c}_m = \frac{\sum_i y_i I(\mathbf{x} \in R_m)}{\sum_i I(\mathbf{x} \in R_m)}$$

Our task is to find the optimal splitting variable j and the split point s that yield the largest drop in the residual sum of squares, minimizing the following:

$$\sum_{i: x_j \le s} (y_i - \hat{c}_1)^2 + \sum_{i: x_j > s} (y_i - \hat{c}_2)^2$$

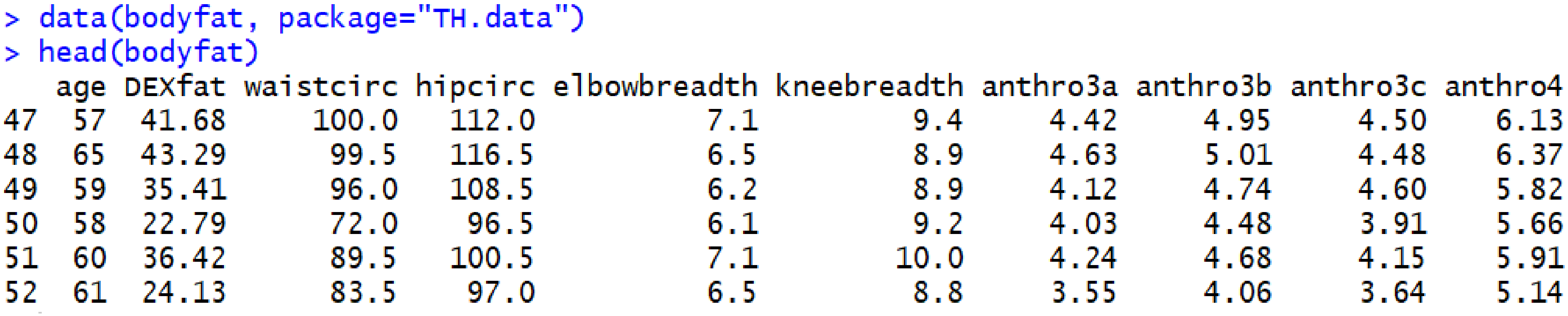
1. Regression Trees

Decision trees works for both **regression** and **classification** by performing binary splits on the recursive predictors.

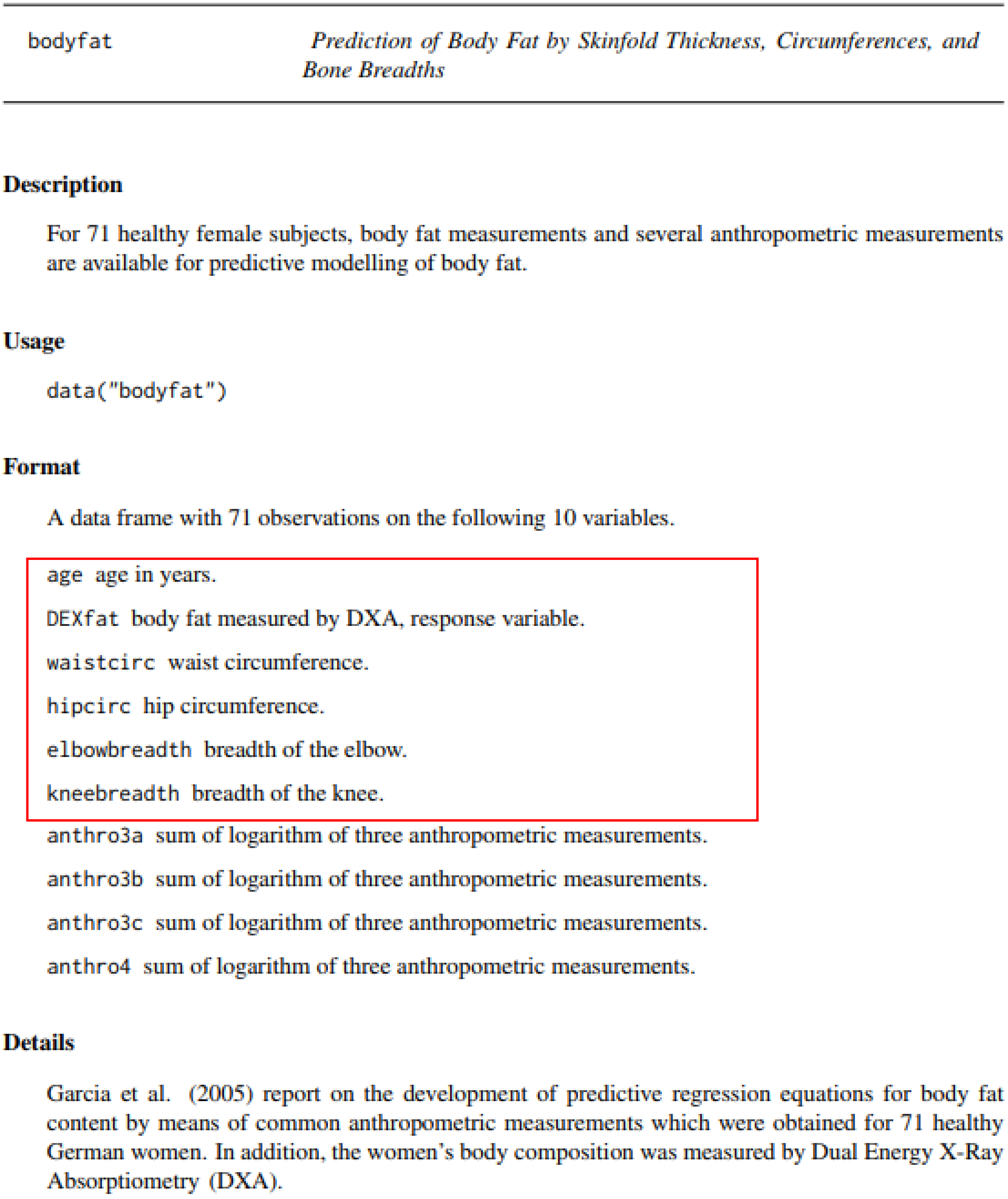
Regression-type trees are generally those where we attempt to predict the values of a continuous dependent variable from one or more continuous or categorical predictor variables.

http://www.statsoft.com/Textbook/Classification-and-Regression-Trees

Sample Dataset: bodyfat



```
> data(bodyfat, package="TH.data")
> head(bodyfat)
   age DEXfat waistcirc hipcirc elbowbreadth kneebreadth anthro3a anthro3b anthro3c anthro4
47  57  41.68     100.0   112.0          7.1         9.4     4.42     4.95     4.50    6.13
48  65  43.29      99.5   116.5          6.5         8.9     4.63     5.01     4.48    6.37
49  59  35.41      96.0   108.5          6.2         8.9     4.12     4.74     4.60    5.82
50  58  22.79      72.0    96.5          6.1         9.2     4.03     4.48     3.91    5.66
51  60  36.42      89.5   100.5          7.1        10.0     4.24     4.68     4.15    5.91
52  61  24.13      83.5    97.0          6.5         8.8     3.55     4.06     3.64    5.14
```



`bodyfat` *Prediction of Body Fat by Skinfold Thickness, Circumferences, and Bone Breadths*

Description

For 71 healthy female subjects, body fat measurements and several anthropometric measurements are available for predictive modelling of body fat.

Usage

```
data("bodyfat")
```

Format

A data frame with 71 observations on the following 10 variables.

`age` age in years.

`DEXfat` body fat measured by DXA, response variable.

`waistcirc` waist circumference.

`hipcirc` hip circumference.

`elbowbreadth` breadth of the elbow.

`kneebreadth` breadth of the knee.

`anthro3a` sum of logarithm of three anthropometric measurements.

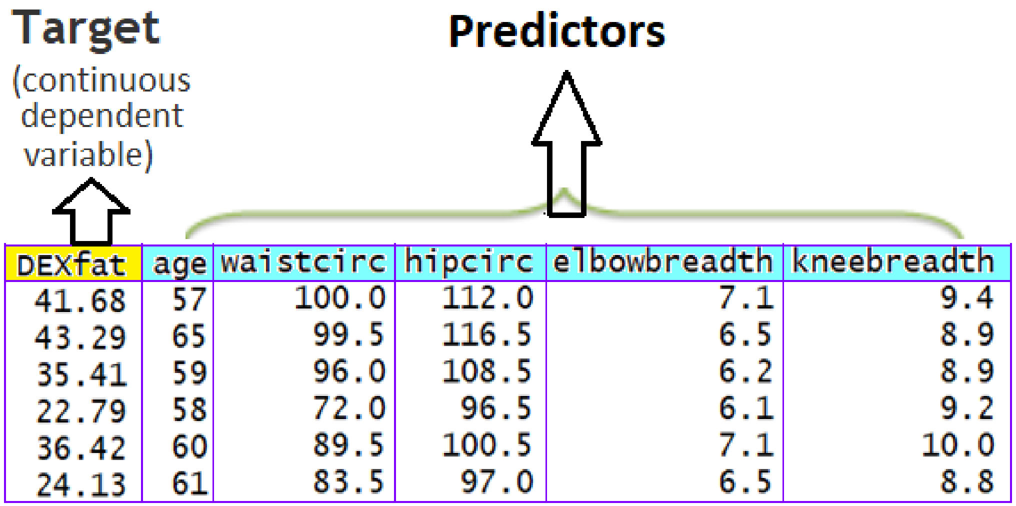
`anthro3b` sum of logarithm of three anthropometric measurements.

`anthro3c` sum of logarithm of three anthropometric measurements.

`anthro4` sum of logarithm of three anthropometric measurements.

Details

Garcia et al. (2005) report on the development of predictive regression equations for body fat content by means of common anthropometric measurements which were obtained for 71 healthy German women. In addition, the women's body composition was measured by Dual Energy X-Ray Absorptiometry (DXA).



Target

(continuous dependent variable)

Predictors

DEXfat	age	waistcirc	hipcirc	elbowbreadth	kneebreadth
41.68	57	100.0	112.0	7.1	9.4
43.29	65	99.5	116.5	6.5	8.9
35.41	59	96.0	108.5	6.2	8.9
22.79	58	72.0	96.5	6.1	9.2
36.42	60	89.5	100.5	7.1	10.0
24.13	61	83.5	97.0	6.5	8.8

rpart(formula, data, method, ...) {rpart}
Recursive Partitioning and Regression Trees

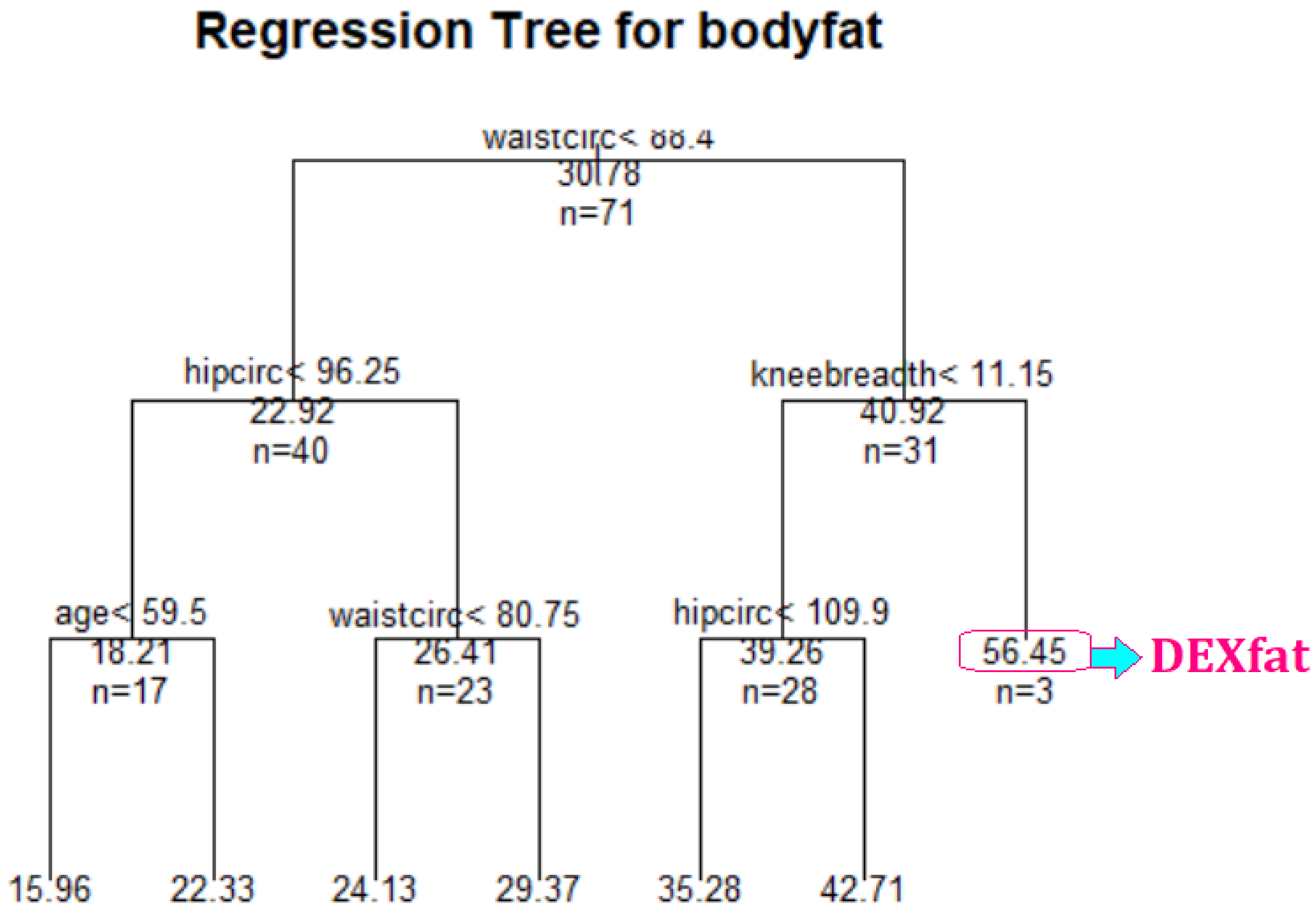
```
> library(rpart)
> rfit <- rpart(DEXfat~age+waistcirc+hipcirc+elbowbreadth+kneebreadth,
+      data=bodyfat,method="anova",control=rpart.control(minsplit=10))
> rfit
n= 71
node), split, n, deviance, yval      * denotes terminal node

    1) root 71 8535.98400 30.78282
      2) waistcirc< 88.4 40 1315.35800 22.92375
        4) hipcirc< 96.25 17  285.91370 18.20765
          8) age< 59.5 11   97.00440 15.96000 *
          9) age>=59.5 6   31.45788 22.32833 *
        5) hipcirc>=96.25 23  371.86530 26.40957
         10) waistcirc< 80.75 13  117.60710 24.13077 *
         11) waistcirc>=80.75 10   98.99016 29.37200 *
      3) waistcirc>=88.4 31 1562.16200 40.92355
        6) kneebreadth< 11.15 28  615.52590 39.26036
         12) hipcirc< 109.9 13  136.29600 35.27846 *
         13) hipcirc>=109.9 15   94.46997 42.71133 *
        7) kneebreadth>=11.15 3  146.28030 56.44667 *
```

yval = the fitted value of the response at the node

1) root : n=71, mean(DEXfat)=30.783
2) waistcirc<88.4: n=40, mean(DEXfat)=22.924
3) waistcirc>=88.4: n=31, mean(DEXfat)=40.924
. . .

[Reference]
T. Hothorn and B. S. Everitt,
A Handbook of Statistical Analyses Using R, 3rd ed. (CRC Press, Boca Raton, FL, 2014) Ch. 9.

```
# Simple plot of regression tree
plot(rfit, uniform=TRUE,
     main="Regression Tree for bodyfat")
text(rfit, use.n=TRUE, all=TRUE, cex=.8)
```

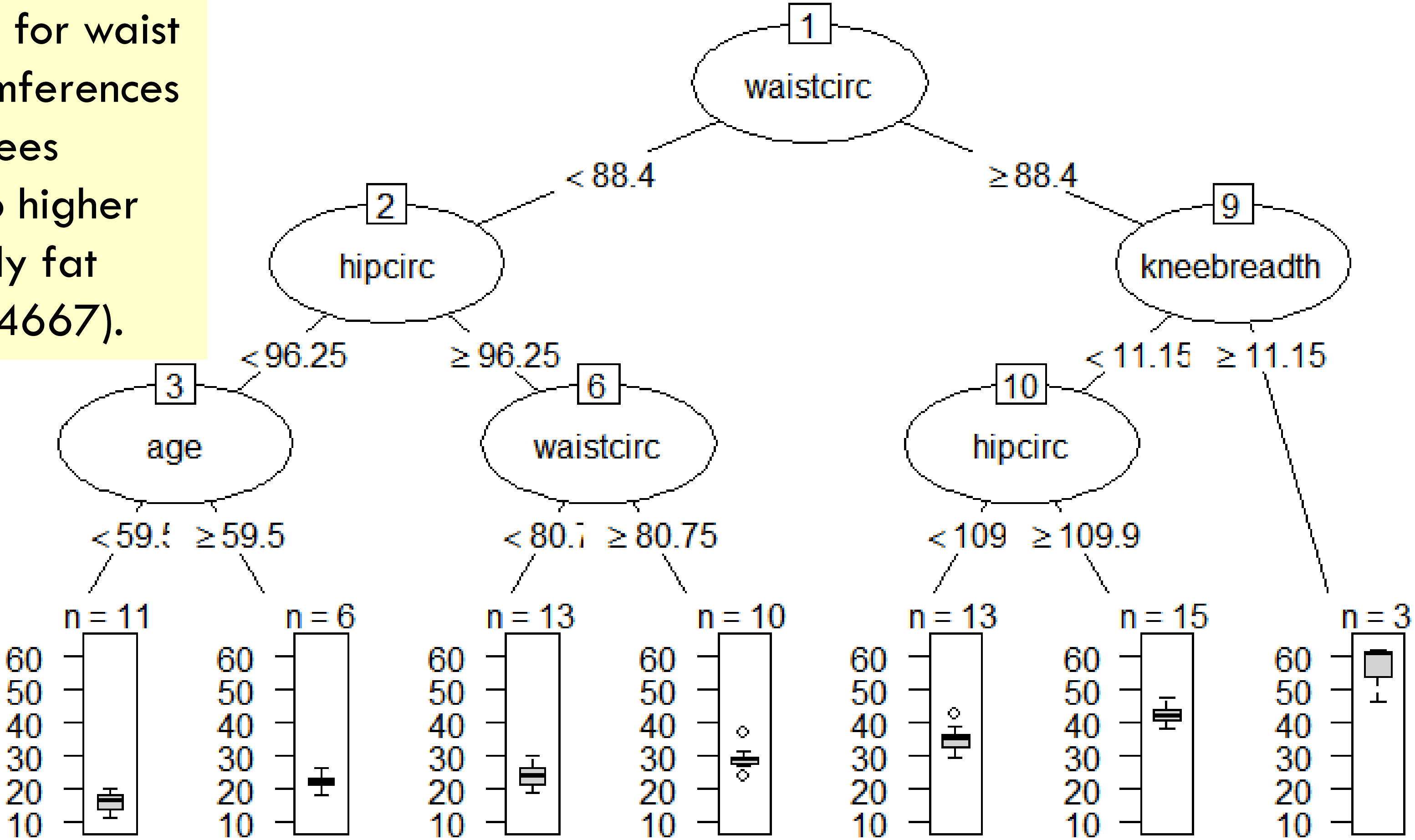


party {partykit}

A class for representing decision trees and corresponding accessor functions.

```
#Pretty plot of regression tree
library(grid); library(libcoin)
library(mvtnorm); library(partykit)
plot(as.party(rfit), tp_args=list(id=FALSE))
```

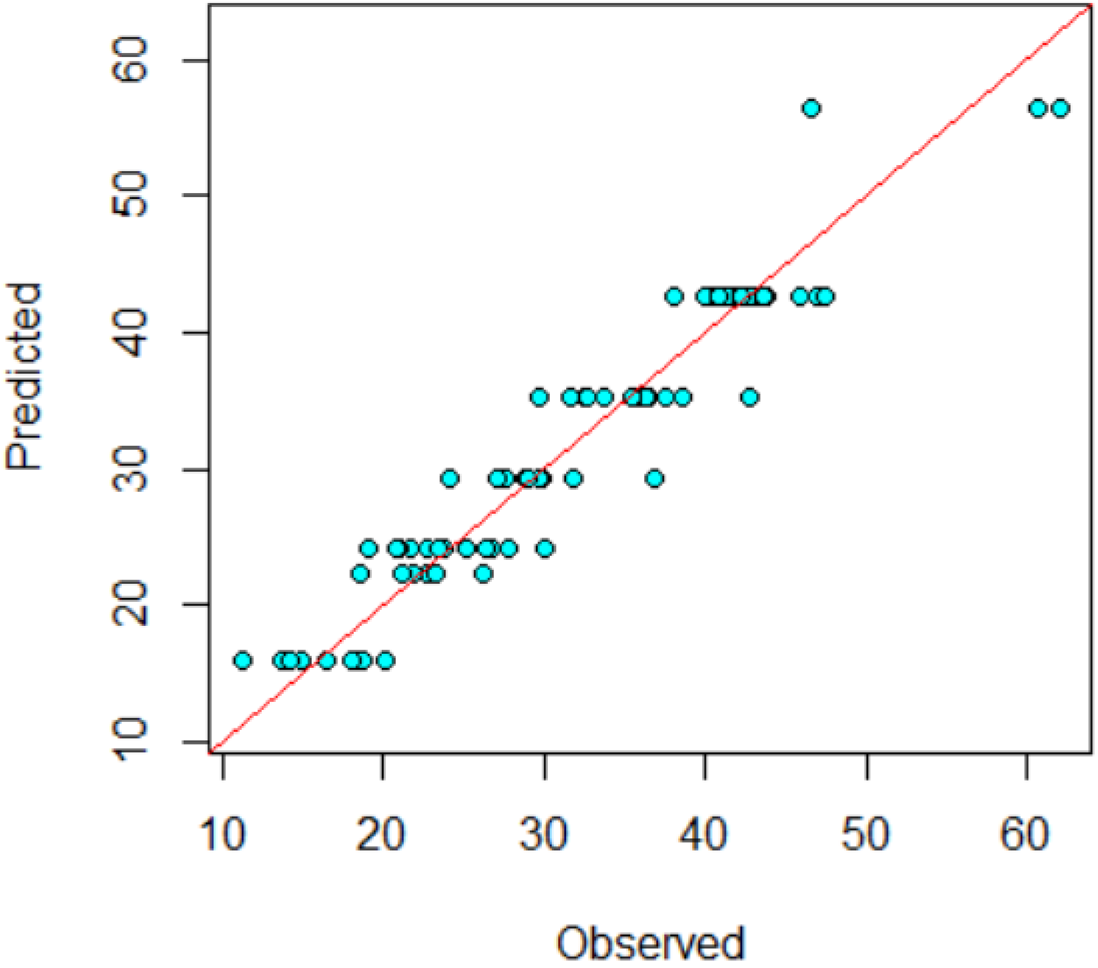
Higher values for waist and hip circumferences and wider knees correspond to higher values of body fat content (56.44667).

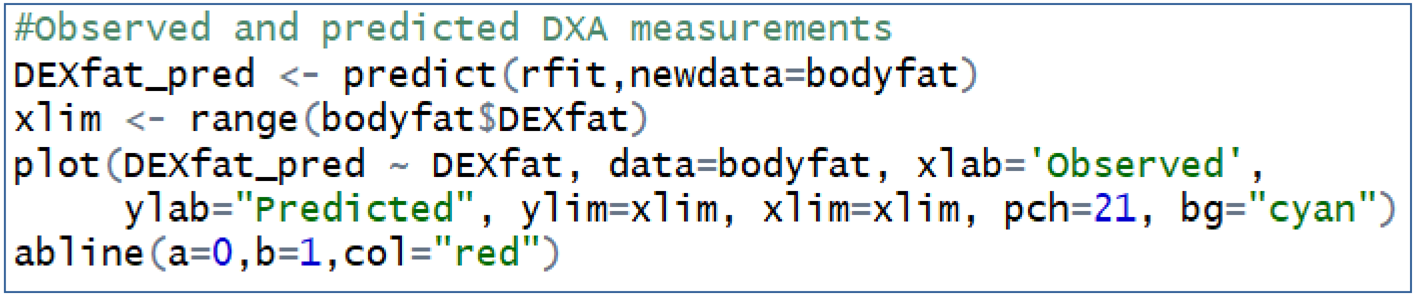


Decisions built based on "ordered" values of the dependent variable.

Regression Tree: Response variable is numeric or continuous.

Given the rpart model, one can compare the model predictions with the actually measured body fat as shown in Figure.





```r
#Observed and predicted DXA measurements
DEXfat_pred <- predict(rfit,newdata=bodyfat)
xlim <- range(bodyfat$DEXfat)
plot(DEXfat_pred ~ DEXfat, data=bodyfat, xlab='Observed',
     ylab="Predicted", ylim=xlim, xlim=xlim, pch=21, bg="cyan")
abline(a=0,b=1,col="red")
```

2. Classification Trees

Classification-type trees are generally those where we attempt to predict values of a **categorical dependent variable** (class, group membership, etc.) from one or more continuous and/or categorical predictor variables.

http://www.statsoft.com/Textbook/Classification-and-Regression-Trees

Sample Dataset: GlaucomaM

data(**GlaucomaM**) {TH.data}

The GlaucomaM data has 196 observations in two classes. 62 variables are derived from a confocal laser scanning image of the optic nerve head, describing its morphology. Observations are from normal and glaucomatous eyes, respectively.

eas : effective area superior

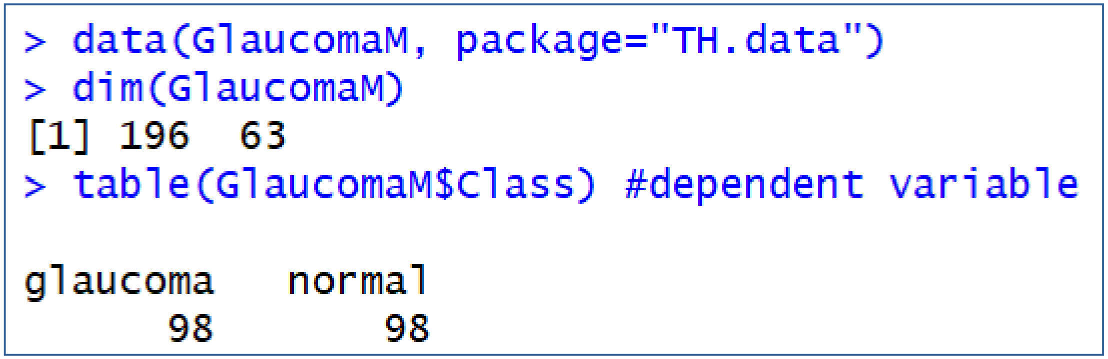
varg : volume above reference global

vars : volume above reference superior

mhcg : mean height contour global

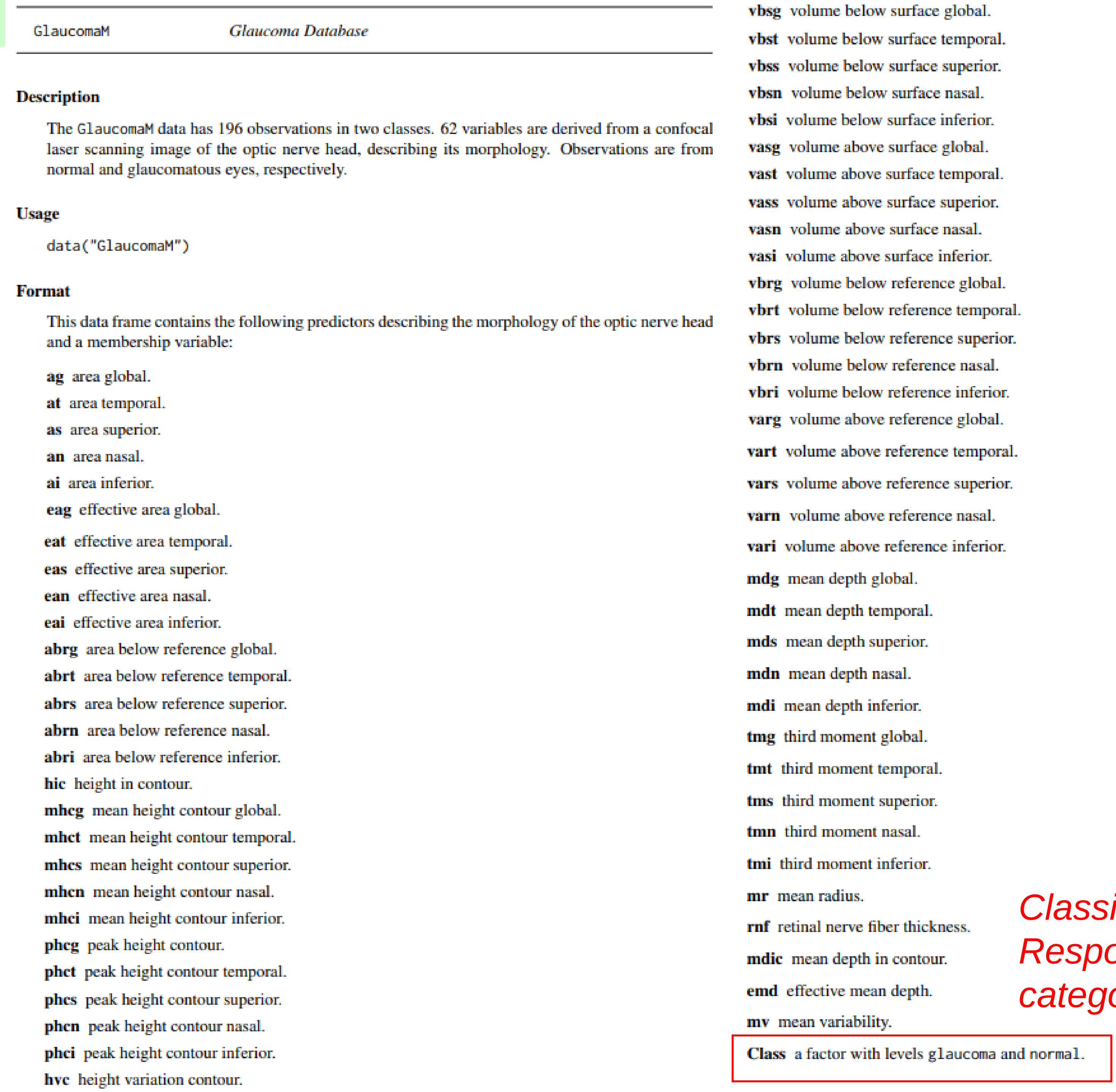
tms : third moment superior

Class : a factor with levels glaucoma and normal



```
> data(GlaucomaM, package="TH.data")
> dim(GlaucomaM)
[1] 196  63
> table(GlaucomaM$Class) #dependent variable

glaucoma   normal 
      98       98 
```

ification Trees



GlaucomaM   *Glaucoma Database*

Description

The GlaucomaM data has 196 observations in two classes. 62 variables are derived from a confocal laser scanning image of the optic nerve head, describing its morphology. Observations are from normal and glaucomatous eyes, respectively.

Usage

```
data("GlaucomaM")
```

Format

This data frame contains the following predictors describing the morphology of the optic nerve head and a membership variable:

ag area global.

at area temporal.

as area superior.

an area nasal.

ai area inferior.

eag effective area global.

eat effective area temporal.

eas effective area superior.

ean effective area nasal.

eai effective area inferior.

abrg area below reference global.

abrt area below reference temporal.

abrs area below reference superior.

abrn area below reference nasal.

abri area below reference inferior.

hic height in contour.

mhcg mean height contour global.

mhct mean height contour temporal.

mhcs mean height contour superior.

mhcn mean height contour nasal.

mhci mean height contour inferior.

phcg peak height contour.

phct peak height contour temporal.

phcs peak height contour superior.

phcn peak height contour nasal.

phci peak height contour inferior.

hvc height variation contour.

vbsg volume below surface global.

vbst volume below surface temporal.

vbss volume below surface superior.

vbsn volume below surface nasal.

vbsi volume below surface inferior.

vasg volume above surface global.

vast volume above surface temporal.

vass volume above surface superior.

vasn volume above surface nasal.

vasi volume above surface inferior.

vbrg volume below reference global.

vbrt volume below reference temporal.

vbrs volume below reference superior.

vbrn volume below reference nasal.

vbri volume below reference inferior.

varg volume above reference global.

vart volume above reference temporal.

vars volume above reference superior.

varn volume above reference nasal.

vari volume above reference inferior.

mdg mean depth global.

mdt mean depth temporal.

mds mean depth superior.

mdn mean depth nasal.

mdi mean depth inferior.

tmg third moment global.

tmt third moment temporal.

tms third moment superior.

tmn third moment nasal.

tmi third moment inferior.

mr mean radius.

rnf retinal nerve fiber thickness.

mdic mean depth in contour.

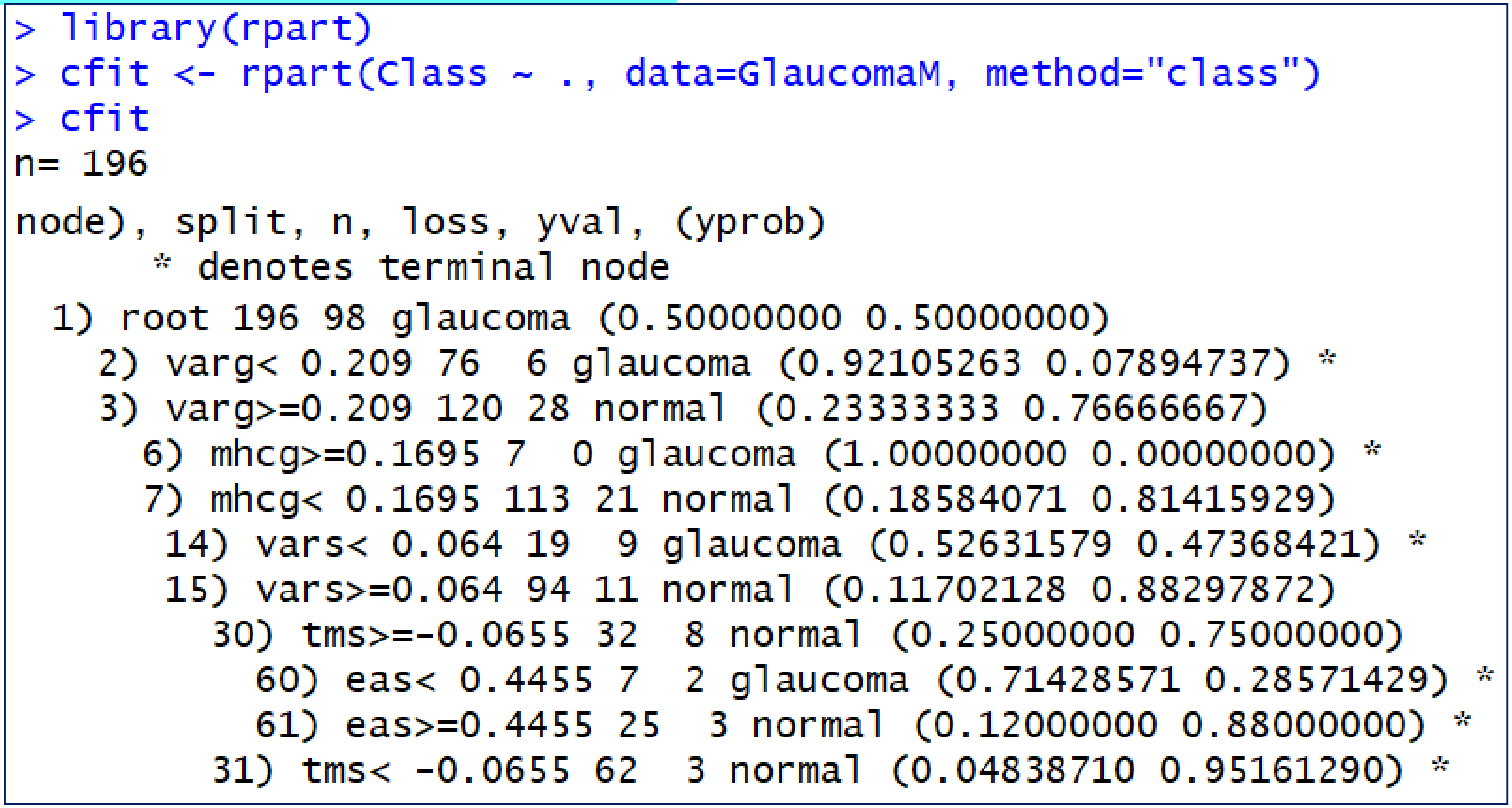
emd effective mean depth.

mv mean variability.

Class a factor with levels glaucoma and normal.

Classification Tree: Response variable is categorical.

• Classification using rpart



```
> library(rpart)
> cfit <- rpart(Class ~ ., data=GlaucomaM, method="class")
> cfit
n= 196

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 196 98 glaucoma (0.50000000 0.50000000)
   2) varg< 0.209 76  6 glaucoma (0.92105263 0.07894737) *
   3) varg>=0.209 120 28 normal (0.23333333 0.76666667)
     6) mhcg>=0.1695 7  0 glaucoma (1.00000000 0.00000000) *
     7) mhcg< 0.1695 113 21 normal (0.18584071 0.81415929)
      14) vars< 0.064 19  9 glaucoma (0.52631579 0.47368421) *
      15) vars>=0.064 94 11 normal (0.11702128 0.88297872)
        30) tms>=-0.0655 32  8 normal (0.25000000 0.75000000)
          60) eas< 0.4455 7  2 glaucoma (0.71428571 0.28571429) *
          61) eas>=0.4455 25  3 normal (0.12000000 0.88000000) *
        31) tms< -0.0655 62  3 normal (0.04838710 0.95161290) *
```

yval is the mean variable

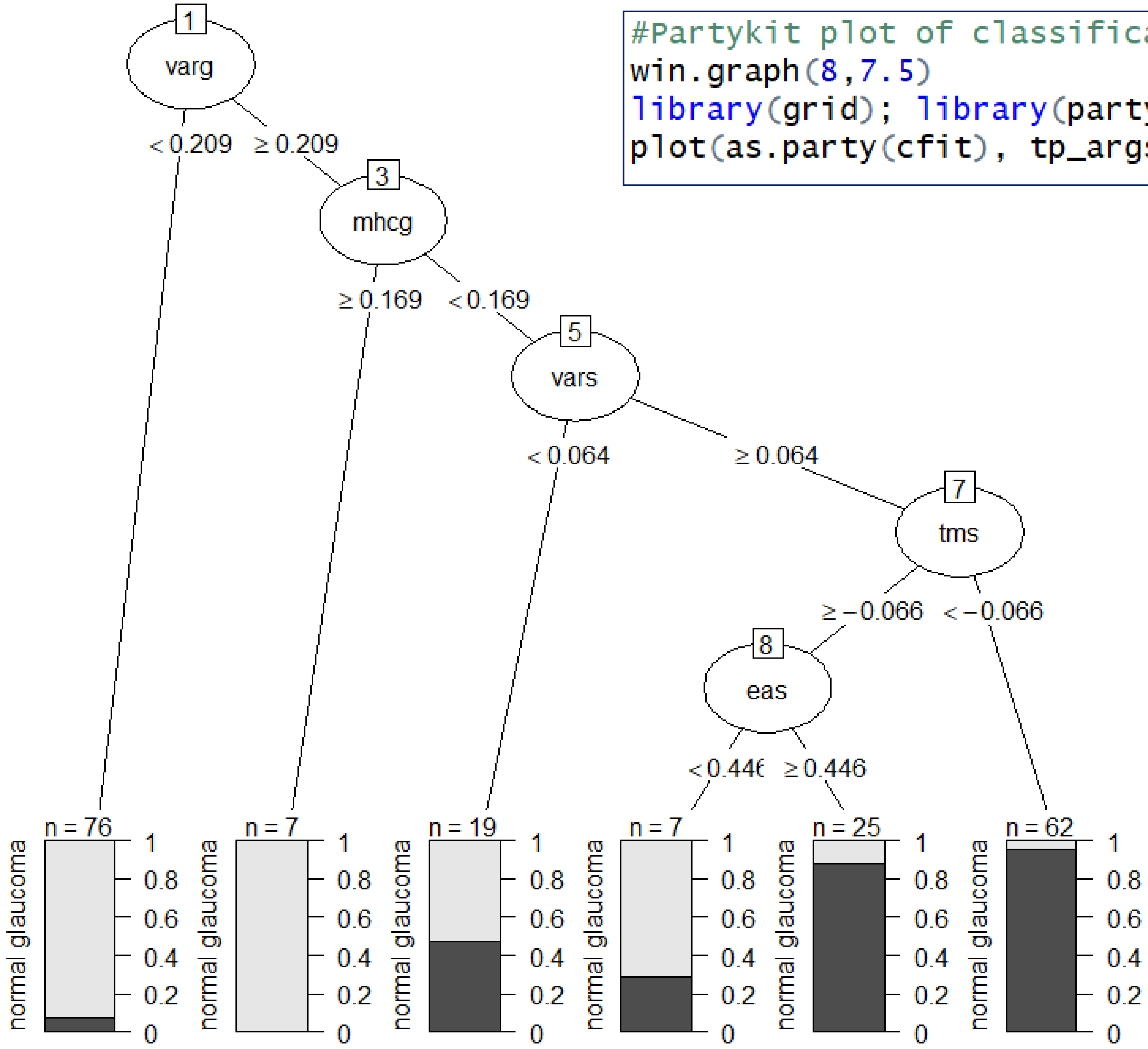
loss means how many do not have the majority trait

1) Root : n=196, 98(50%) are “glaucoma”, 98(50%) are “normal”
2) Varg<0.209 : n=76, 70(92.11%) are “glaucoma”, 6(7.89%) are “normal”
3) Varg>=0.209: n=120, 28(23.33%) are “glaucoma”, 92(76.67%) are “normal”

. . .

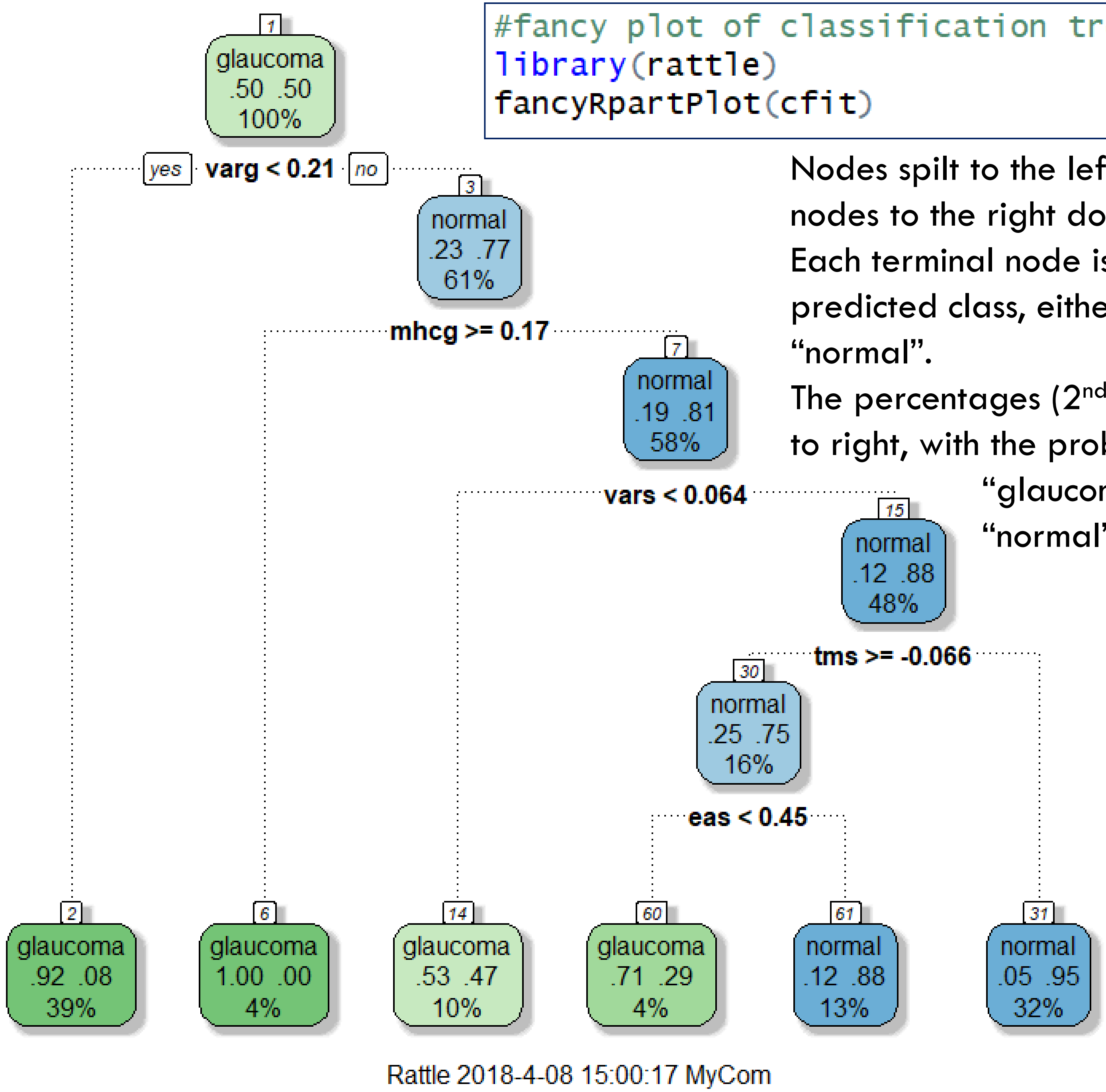
• Classification using rpart

```
#Partykit plot of classification tree
win.graph(8,7.5)
library(grid); library(partykit)
plot(as.party(cfit), tp_args=list(id=FALSE))
```



• Classification using rpart

```
#fancy plot of classification tree
library(rattle)
fancyRpartPlot(cfit)
```



Nodes spilt to the left meet the criteria while nodes to the right do not.

Each terminal node is labelled by the predicted class, either “glaucoma” or “normal”.

The percentages (2^{nd} line) are read from left to right, with the probability of being “glaucoma” on the left and “normal” on the right.

3. Qualitative Interaction Trees

When two treatment alternatives (say A and B) are available for some problem, one may be interested in qualitative treatment-subgroup interactions. Such interactions imply the existence of subgroups of persons (patients) which are such that in one subgroup Treatment A outperforms Treatment B, whereas the reverse holds in another subgroup.

The interaction tree is to identify subgroups that are involved in meaningful qualitative treatment-subgroup interactions.

quint(formula, data, control) {quint} Qualitative Interaction Trees

It performs a subgroup analysis by QUalitative INteraction Trees (QUINT) and is suitable for data from a two-arm randomized controlled trial.

quint.control(maxl=10, B=25, ...) {quint}

Various parameters that control aspects of the “quint” algorithm.

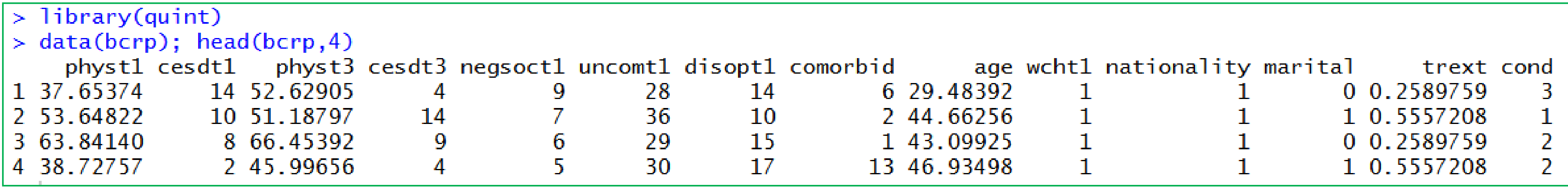
maxl	maximum number of leaves (L) of the tree.
B	the number of bootstrap samples to be drawn.

(1) Example: bcrp

bcrp {quint} Breast Cancer Recovery Project

Data from a three-arm randomized controlled trial. Women with early-stage breast cancer were randomly assigned to a nutrition intervention (n = 85), an education intervention (n = 83) or standard care (n = 84). They were measured before and after treatment. These data contain the baseline measurement and the 9-month follow-up.

physt1	physical functioning (from SF-36) at baseline.
cesdt1	depression score (CESD) at baseline.
physt3	physical functioning (from SF-36) at 9 months follow-up.
cesdt3	depression score (CESD) at 9 months follow-up.
negsoct1	negative social interaction at baseline.
uncomt1	unmitigated communion at baseline.
disopt1	dispositional optimism at baseline.
comorbid	number of comorbidities (e.g. diabetes, migraines, arthritis, or angina).
age	age at baseline.
wcht1	weight change since diagnosis: yes [1] or no [0].
nationality	Caucasian [1] or not [0].
marital	married [1] or not [0].
trext	treatment extensiveness index
cond	experimental condition: nutrition [1], education [2] or standard care [3].

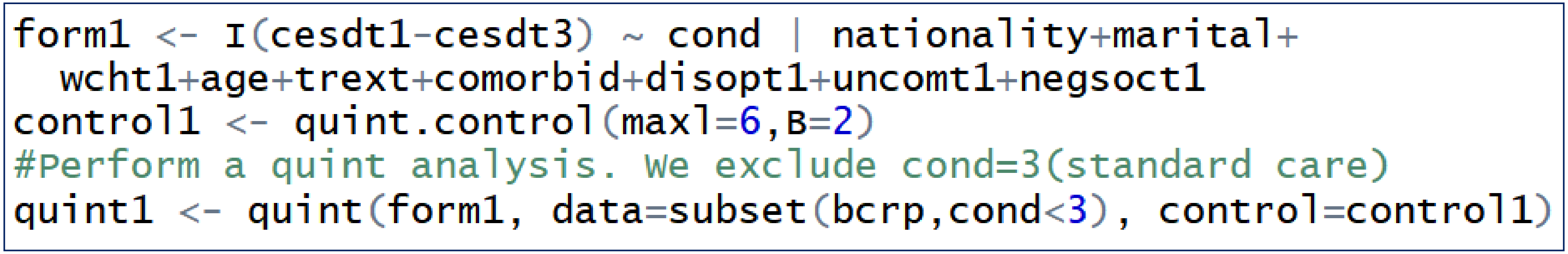


```
> library(quint)
> data(bcrp); head(bcrp,4)
    physt1 cesdt1   physt3 cesdt3 negsoct1 uncomt1 disopt1 comorbid      age wcht1 nationality marital     trext cond
1 37.65374     14 52.62905      4        9      28      14        6 29.48392     1           1       0 0.2589759    3
2 53.64822     10 51.18797     14        7      36      10        2 44.66256     1           1       1 0.5557208    1
3 63.84140      8 66.45392      9        6      29      15        1 43.09925     1           1       0 0.2589759    2
4 38.72757      2 45.99656      4        5      30      17       13 46.93498     1           1       1 0.5557208    2
```

(2) Qualitative Interaction Tree for the Outcome improvement in Depression

Ref: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4891398/

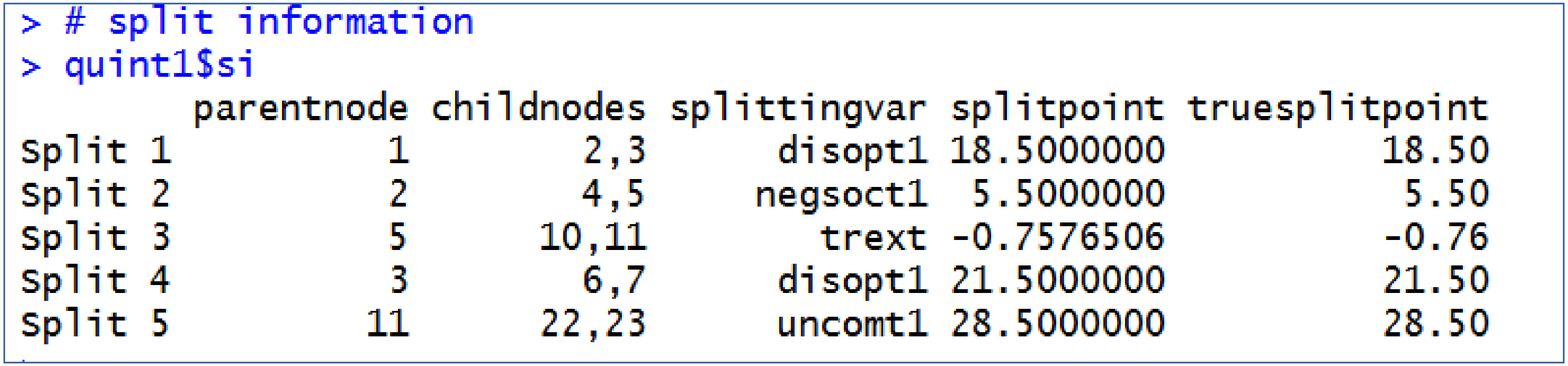
For which subgroup of women is a nutrition intervention more effective than an education intervention, for which subgroup does the reverse hold true, and for which subgroup do the two interventions not lead to clearly different outcomes?



```
form1 <- I(cesdt1-cesdt3) ~ cond | nationality+marital+
  wcht1+age+trext+comorbid+disopt1+uncomt1+negsoct1
control1 <- quint.control(maxl=6,B=2)
#Perform a quint analysis. We exclude cond=3(standard care)
quint1 <- quint(form1, data=subset(bcrp,cond<3), control=control1)
```

- Outcome Y=cesdt1-cesdt3 is a change score between timepoint 3 and timepoint 1.
- Positive Y value indicates an improvement in depression.
- As we focus on the comparison between the nutrition and the education condition, we exclude the standard care (cond=3).

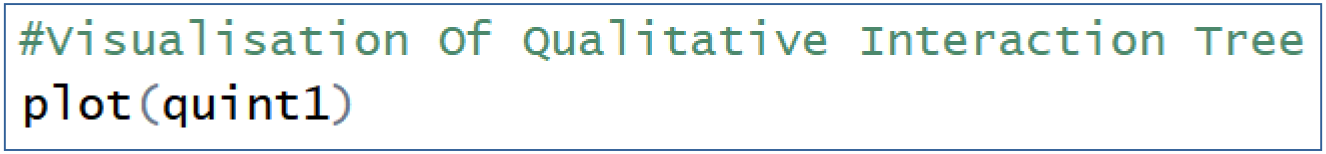
(2) Qualitative Interaction Tree for the Outcome improvement in Depression



```
> # split information
> quint1$si
        parentnode childnodes splittingvar splitpoint truesplitpoint
Split 1          1        2,3      disopt1 18.5000000          18.50
Split 2          2        4,5     negsoct1  5.5000000           5.50
Split 3          5      10,11        trext -0.7576506          -0.76
Split 4          3        6,7      disopt1 21.5000000          21.50
Split 5         11      22,23      uncomt1 28.5000000          28.50
```

The split information shows in the first two columns the node numbers of the parent nodes that were split and those of the resulting child nodes.

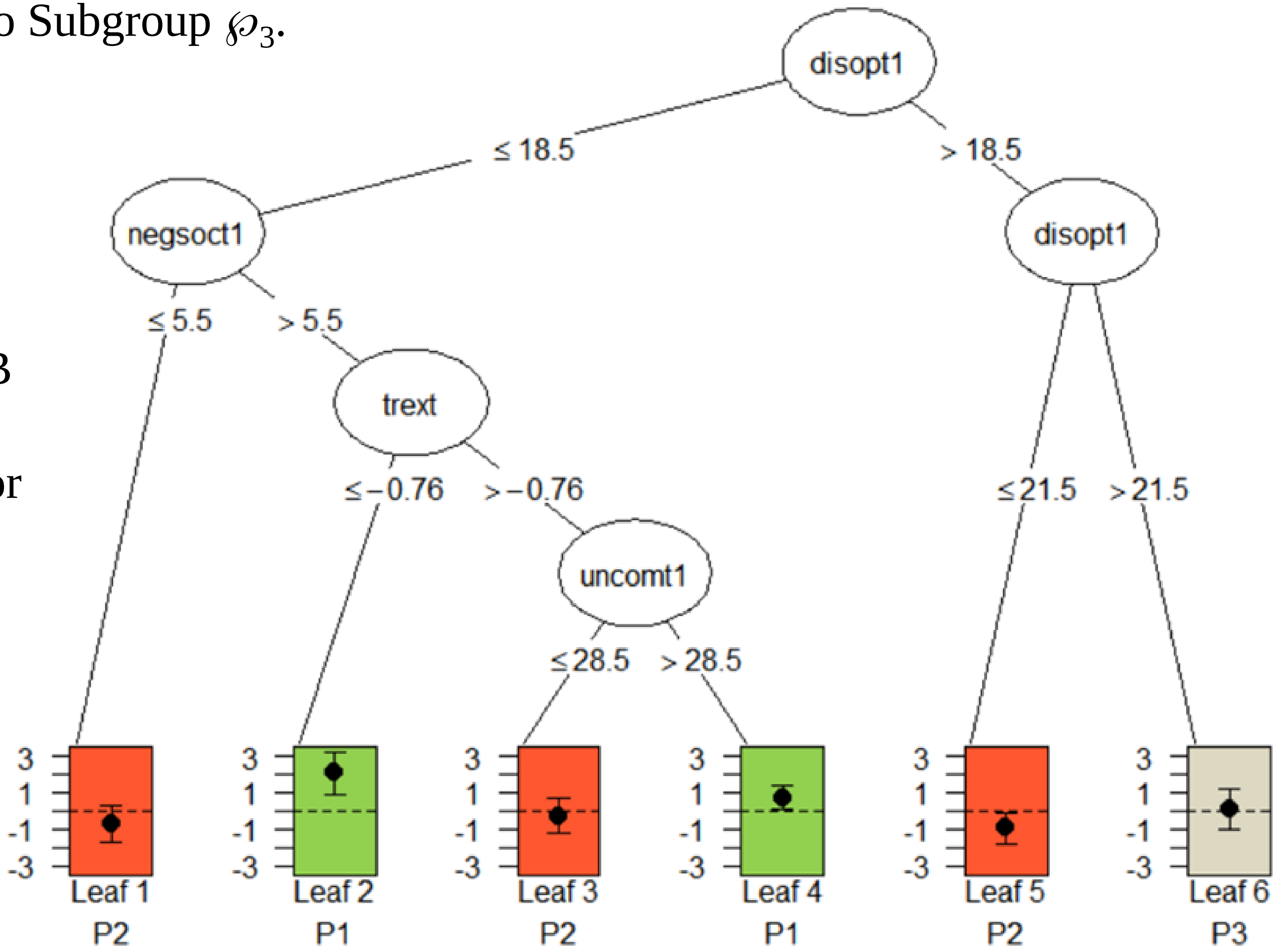
(2) Qualitative Interaction Tree for the Outcome improvement in Depression



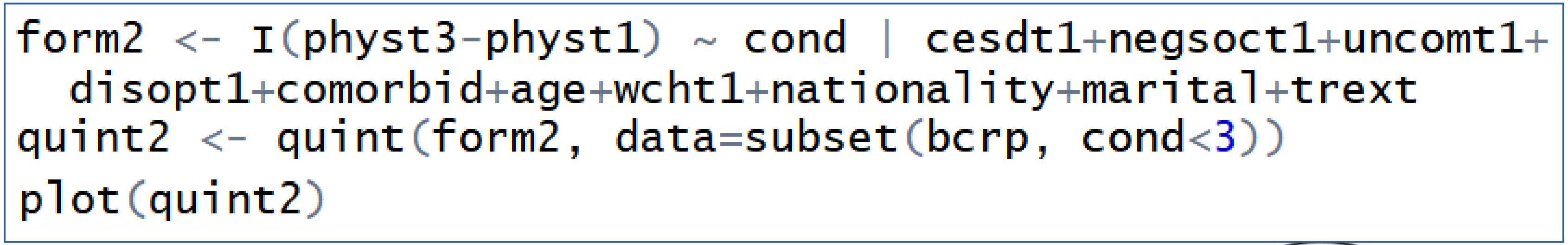
```
#Visualisation Of Qualitative Interaction Tree
plot(quint1)
```

The root node is split into two internal nodes on the basis of the value 18.5 on the variable “disopt1”. Clients who score 18.5 or lower fall into the left child node and the others fall into the right child node. Each leaf of the tree represents a client type and is assigned to one of the three subgroups, colored in green, red, or grey. A green leaf belongs to Subgroup $\wp_1$, a red leaf to Subgroup $\wp_2$, and a grey leaf to Subgroup $\wp_3$.

Subgroup $\wp_1$ contains those clients for whom Treatment A (nutrition) is better than Treatment B (education), Subgroup $\wp_2$ those for whom B is better than A, and (the optional) Subgroup $\wp_3$ those for whom it does not make any difference.

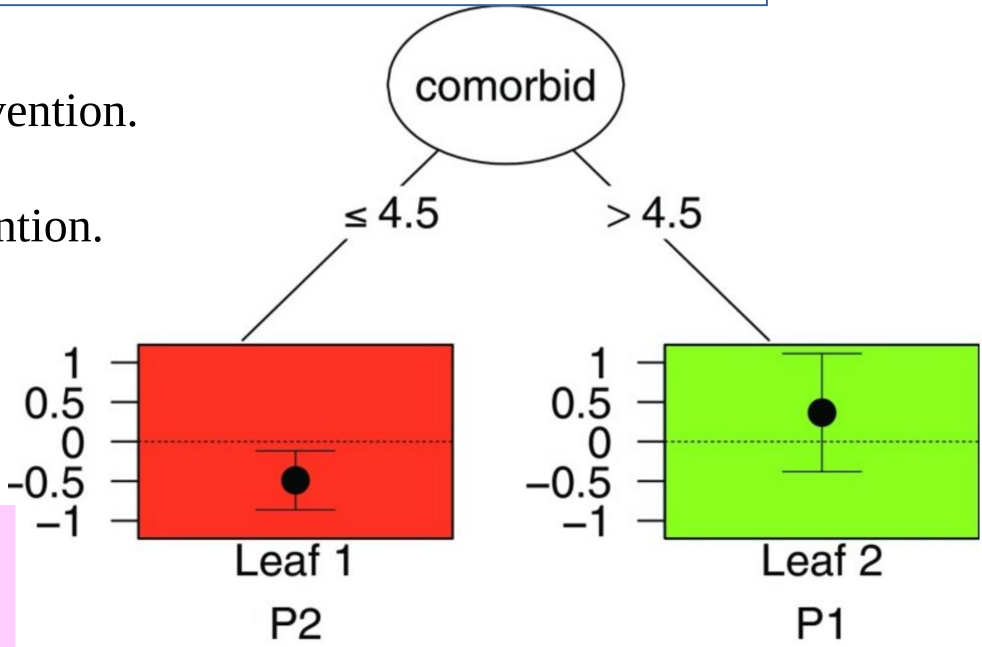


(3) Qualitative Interaction Tree for the Outcome Improvement in Physical Functioning



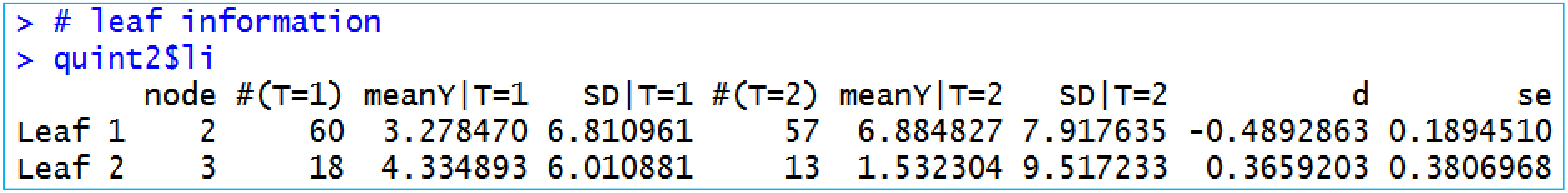
```
form2 <- I(physt3-physt1) ~ cond | cesdt1+negsoct1+uncomt1+
  disopt1+comorbid+age+wcht1+nationality+marital+trext
quint2 <- quint(form2, data=subset(bcrp, cond<3))
plot(quint2)
```

Subgroup $\wp_2$ of women (red) : the education intervention was better than the nutrition intervention. Subgroup $\wp_1$ of women (green): the nutrition intervention outperforms the education intervention.



Leaf 1 : the mean improvement was 3.28 for the nutrition intervention and 6.88 for the education intervention.

Leaf 2 : the mean improvement was 4.33 for the nutrition intervention and 1.53 for the education intervention.



```
> # leaf information
> quint2$li
       node #(T=1) meanY|T=1   SD|T=1 #(T=2) meanY|T=2   SD|T=2          d        se
Leaf 1    2     60  3.278470 6.810961     57  6.884827 7.917635 -0.4892863 0.1894510
Leaf 2    3     18  4.334893 6.010881     13  1.532304 9.517233  0.3659203 0.3806968
```