Fitting GLMM trees in R

R scripts accompanying the webinar for University of Montreal

Marjolein Fokkema

06-05-2021

Model formulas

The GLMM is given by:

$E[y_{ij}|x_{ij}] = \mu_{ij}$

$g(\mu_{ij}) = x_{ij}^\top \beta_k + z_{ij}^\top b_i$

Where i is an identifier for the level-II unit, and j is an identifier for the level-I unit.

If we have a continuous response variable with normally distributed residuals, then g is the identity function and we have:

$y_{ij} = x_{ij}^\top \beta + z_{ij}^\top b_i + \epsilon_{ij}$

Note: We thus have a fixed-effects part ($x_{ij}^\top \beta$), a random-effects part ($z_{ij}^\top b_i$) and a residual error term (ϵ_{ij}).

The GLMM tree model differs in that the fixed-effect part may differ between subgroups:

$y_{ij} = x_{ij}^\top \beta_k + z_{ij}^\top b_i + \epsilon_{ij}$

Thus, k is an identifier for the subgroup. The GLMM tree algorithm finds these subgroups k, using additional covariates U. These may be measured on the lowest level i (which is commonly encountered in cross-sectional multilevel data), or on the higher level j (which is commonly encountered in longitudinal data).

The subgroups k, and parameters *beta* and b cannot be estimated in a single step. An iterative approach is taken, where the model-base recursive partitioning algorithm of Zeileis, Hothorn & Hornk (2008) is used to estimate the subgroups k, and the usual (restricted) maximum likelihood approach is used to estimate the fixed- and random-effects parameters, see Fokkema et al. (2018).

Example 1: Improving Access to Psychological Therapies data

For this example, we will make use of artificial data modelled after the Stratified Medicine Approaches foR Treatment Selection (SMART) prediction tournament, from the Improving Access to Psychological Therapies (IAPT) project (Lucock et al., 2017). The SMART data contains data from patients receiving mental-health services in the Northern UK. Patients were (non-randomly) assigned to low intensity treatment (e.g., guided self-help, computerized cognitive behavior therapy) or high intensity treatment (e.g., face-to-face psychological therapies). The aim of the SMART tournament was to identify patients who would benefit most from HI vs. LI treatment. I do not own the data, so I generated an artificial dataset which mimics the original data, available in the file "SMART mimic data.txt".

```r
library("glmertree")
SMART <- read.table("SMART mimic data.txt", stringsAsFactors = TRUE)
names(SMART)
```

```
##  [1] "recovered"  "Treatment"  "Age"        "Gender"     "Ethnicity" 
##  [6] "Diagnosis"  "Employment" "Disability" "PHQ9_pre"   "GAD7_pre"  
## [11] "WSAS_pre"   "Medication" "center"
```

```r
SMART_tree <- glmertree(recovered ~ Treatment | center | Age + PHQ9_pre +
                          GAD7_pre + WSAS_pre + Gender + Ethnicity +
                          Diagnosis + Employment + Disability + Medication,
                        data = SMART, family = binomial)
```

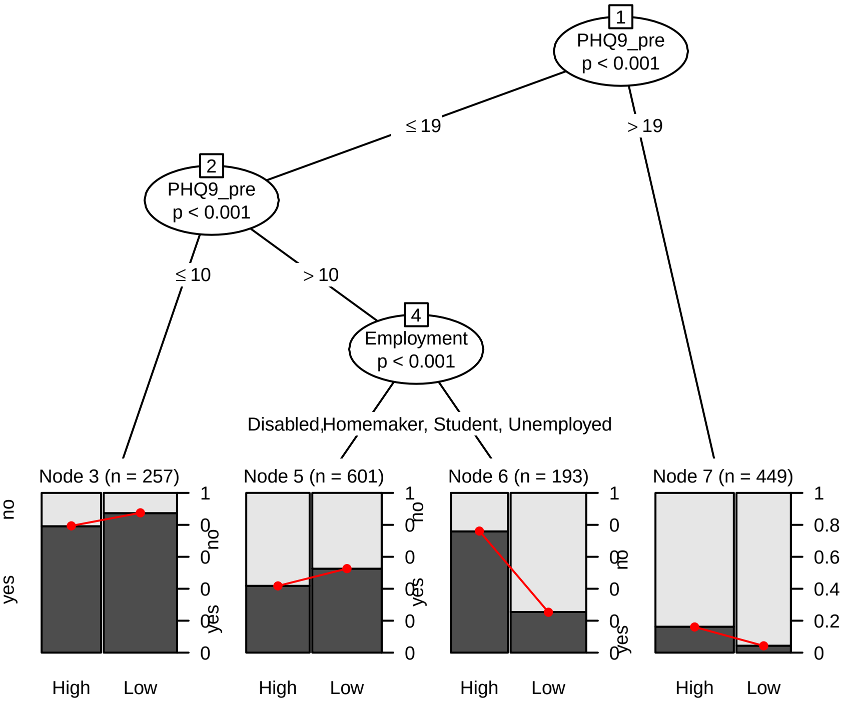
We can print and plot the results as follows:

```r
SMART_tree$tree
```

```
## Generalized linear model tree (family: binomial)
## 
## Model formula:
## recovered ~ Treatment | Age + PHQ9_pre + GAD7_pre + WSAS_pre + 
##     Gender + Ethnicity + Diagnosis + Employment + Disability + 
##     Medication
## 
## Fitted party:
## [1] root
## |   [2] PHQ9_pre <= 19
## |   |   [3] PHQ9_pre <= 10: n = 257
## |   |        (Intercept) TreatmentLow 
## |   |          1.5255853    0.5577717 
## |   |   [4] PHQ9_pre > 10
## |   |   |   [5] Employment in Disabled, Employed, Retired: n = 601
## |   |   |        (Intercept) TreatmentLow 
## |   |   |         -0.2975380    0.4358343 
## |   |   |   [6] Employment in Homemaker, Student, Unemployed: n = 193
## |   |   |        (Intercept) TreatmentLow 
## |   |   |           1.096908    -2.359766 
## |   [7] PHQ9_pre > 19: n = 449
## |        (Intercept) TreatmentLow 
## |          -1.870597    -1.529347 
## 
```

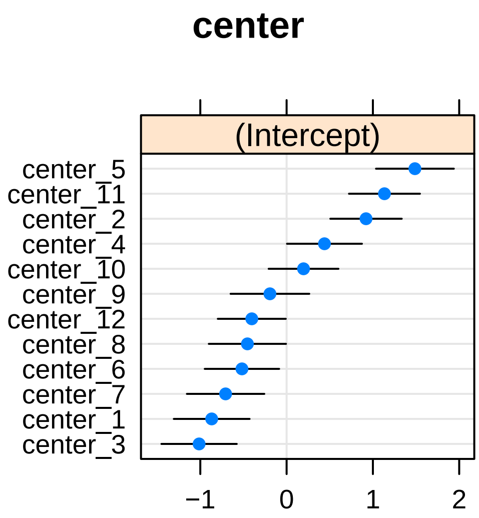
```
## Number of inner nodes:    3
## Number of terminal nodes: 4
## Number of parameters per node: 2
## Objective function (negative log-likelihood): 699.7975
```

```
plot(SMART_tree, which = "tree", gp = gpar(cex = .6))
```



```
plot(SMART_tree, which = "ranef")
```

```
## $center
```



```
VarCorr(SMART_tree)
```

```
##  Groups Name        Std.Dev.
##  center (Intercept) 0.82557
```

```
fixef(SMART_tree)
```

```
##   (Intercept) TreatmentLow
## 3   1.5406327    0.5620038
## 5  -0.2995581    0.4400173
## 6   1.1087452   -2.3810215
## 7  -1.8839567   -1.5372864
```

Example 2: Alcohol data

Curran, Stice, and Chassin (1997) collected data on 82 adolescents at three time points starting at age 14 to assess factors that affect teen drinking behavior. Key variables in the data set “alcohol.csv” (accessed via Singer and Willett, 2003) are as follows:

- `id` = numerical identifier for subject
- `age` = 14, 15, or 16
- `coa` = 1 if the teen is a child of an alcoholic parent; 0 otherwise
- `male` = 1 if male; 0 if female
- `peer` = a measure of peer alcohol use, taken when each subject was 14. This is the square root of the sum of two 6-point items about the proportion of friends who drink occasionally or regularly.
- `alcuse` = the primary response. Four items—(a) drank beer or wine, (b) drank hard liquor, (c) 5 or more drinks in a row, and (d) got drunk—were each scored on an 8-point scale, from 0=“not at all” to 7=“every day”. Then alcuse is the square root of the sum of these four items.

Primary research questions included: Do trajectories of alcohol use differ by parental alcoholism? Do trajectories of alcohol use differ by peer alcohol use?

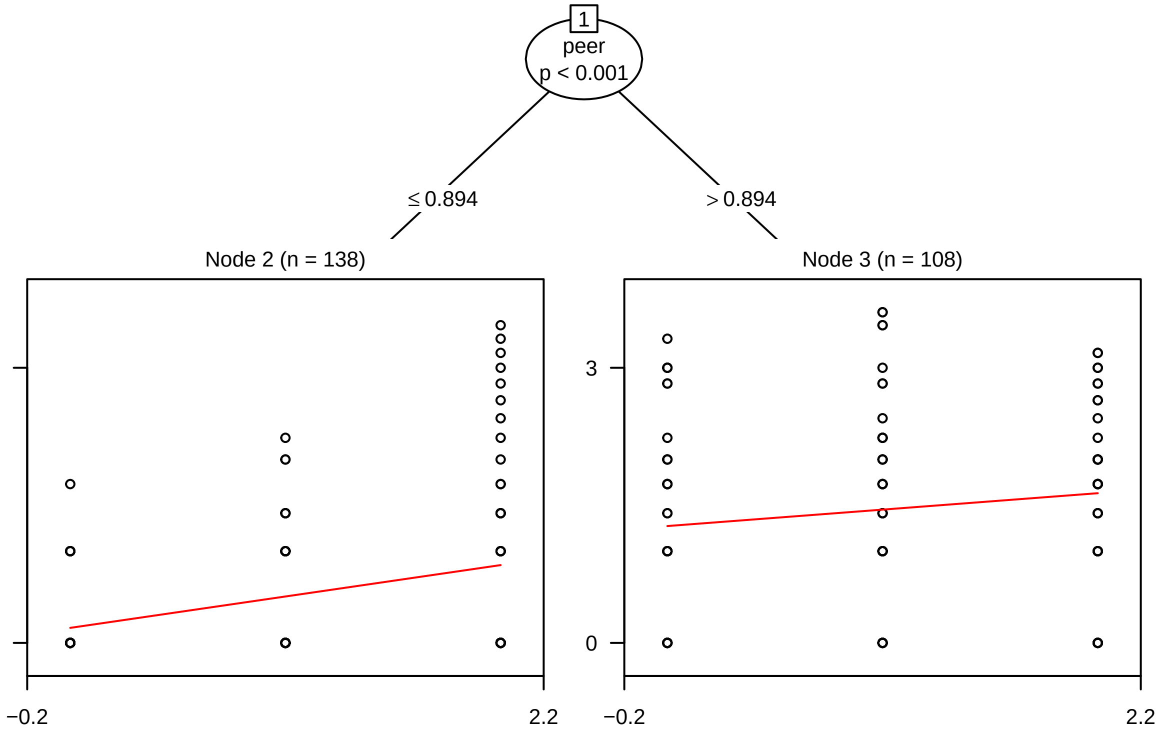
```
alco <- read.table("alcohol.csv", header= TRUE, sep = ",",
                   stringsAsFactors = TRUE)[ , -1]
summary(alco)
```

```
##        id             age          coa              male             peer       
##  Min.   : 1.0   Min.   :14   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:21.0   1st Qu.:14   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :41.5   Median :15   Median :0.0000   Median :1.0000   Median :0.8944  
##  Mean   :41.5   Mean   :15   Mean   :0.4512   Mean   :0.5122   Mean   :1.0176  
##  3rd Qu.:62.0   3rd Qu.:16   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.5492  
##  Max.   :82.0   Max.   :16   Max.   :1.0000   Max.   :1.0000   Max.   :2.5298  
##      alcuse     
##  Min.   :0.000  
##  1st Qu.:0.000  
##  Median :1.000  
##  Mean   :0.922  
##  3rd Qu.:1.732  
##  Max.   :3.606  
```

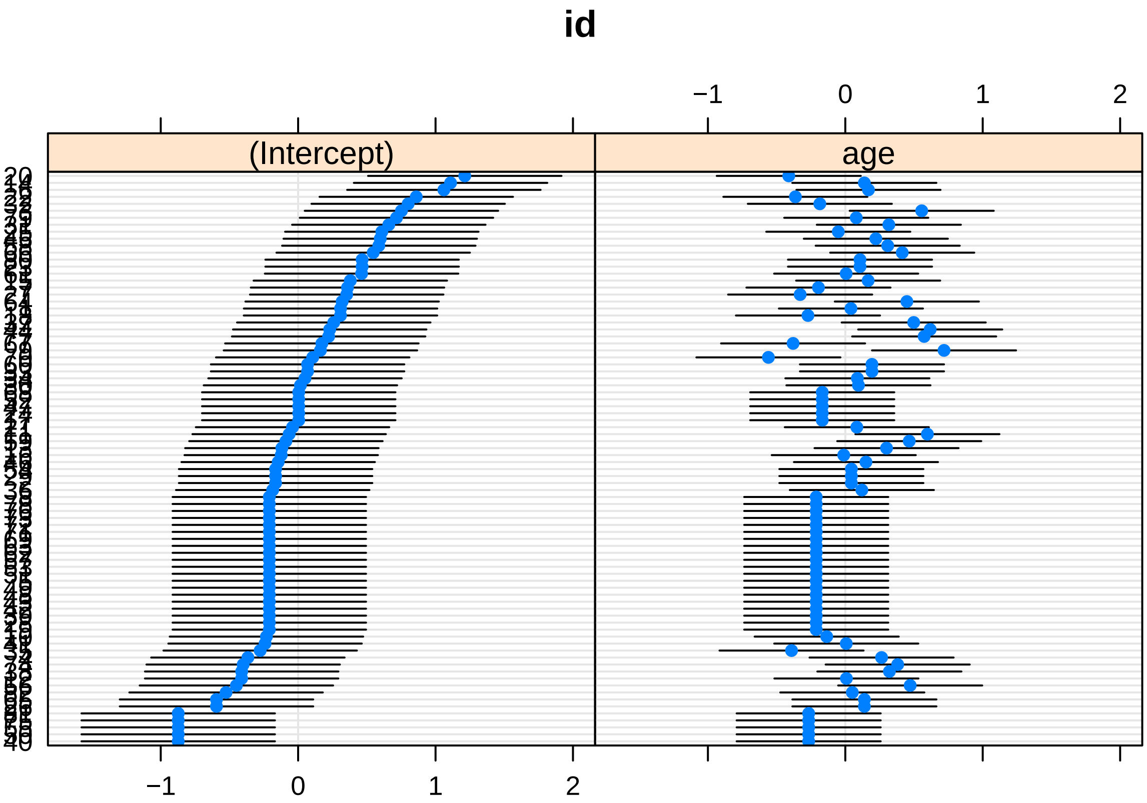
```
alco$age <- alco$age - 14L
```

Because the possible partitioning variables are measured on level II, we use the `cluster` argument to specify this to the algorithm (otherwise, the parameter stability tests will likely be overpowered):

```
lt <- lmertree(alcuse ~ age | (age|id) | coa + male + peer, data = alco,
               cluster = id)
plot(lt, fitted = "marginal", gp = gpar(cex = 0.7))
```



```
## $id
```



The plots of the random effects look quite messy, because of the large number of study participants. We can request the (co)variances of the random effects as follows:

```
fixef(lt)
```

```
##   (Intercept)       age
## 2   0.1638493 0.3423005
## 3   1.2741616 0.1790998
```

```
VarCorr(lt)
```

```
##  Groups   Name        Std.Dev. Corr
##  id       (Intercept) 0.57947
##           age         0.39048  -0.127
##  Residual             0.58077
```

Example 3: Musical performances

Stage fright can be a serious problem for performers, and understanding the personality underpinnings of performance anxiety is an important step in determining how to minimize its impact. Sadler and Miller (2010) studied the emotional state of musicians before performances and factors which may affect their emotional state. Data was collected by having 37 undergraduate music majors from a competitive undergraduate music program fill out diaries prior to performances over the course of an academic year. In particular, study participants completed a Positive Affect Negative Affect Schedule (PANAS) before each performance. The PANAS instrument provided two key outcome measures: negative affect (a state measure of anxiety) and positive affect (a state measure of happiness). We will focus on negative affect as our primary response measuring performance anxiety.

Factors which were examined for their potential relationships with performance anxiety included: performance type (solo, large ensemble, or small ensemble); audience (instructor, public, students, or juried); if the piece was played from memory; age; gender; instrument (voice, orchestral, or keyboard); and, years studying the instrument. In addition, the personalities of study participants were assessed at baseline through the Multidimensional Personality Questionnaire (MPQ). The MPQ provided scores for one lower-order factor (absorption) and three higher-order factors: positive emotionality (PEM—a composite of well-being, social potency, achievement, and social closeness); negative emotionality (NEM—a composite of stress reaction, alienation, and aggression); and constraint (a composite of control, harm avoidance, and traditionalism).

Here, we look at trajectories of negative affect scores over the course of repeated assessments at solo performances.

```r
music <- read.table("musicdata.csv", header=T, sep=",",
                    stringsAsFactors = TRUE)[ , -1]
summary(music)
```

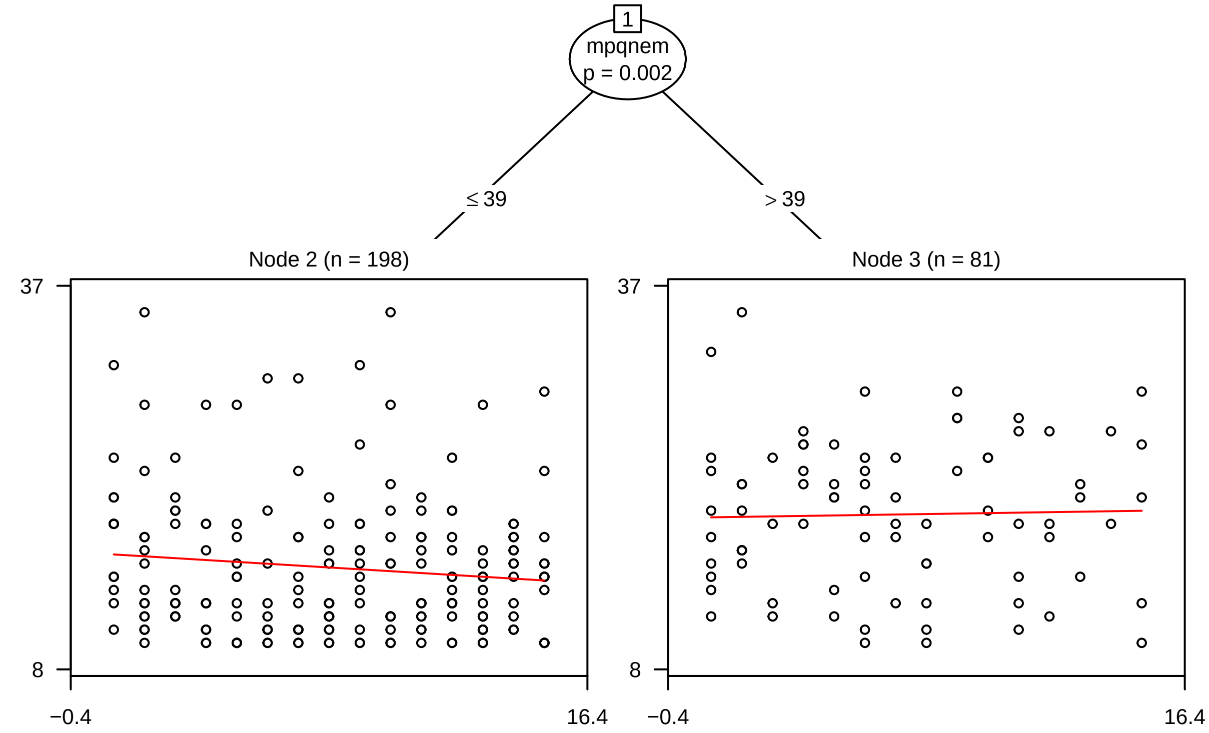
```
##        id            diary           previous              perform_type
##  Min.   : 1.00   Min.   : 1.000   Min.   : 0.000   Large Ensemble:136
##  1st Qu.:10.00   1st Qu.: 4.000   1st Qu.: 3.000   Small Ensemble: 82
##  Median :22.00   Median : 8.000   Median : 7.000   Solo          :279
##  Mean   :22.24   Mean   : 7.781   Mean   : 6.781
##  3rd Qu.:34.00   3rd Qu.:11.000   3rd Qu.:10.000
##  Max.   :43.00   Max.   :15.000   Max.   :14.000
##          memory                  audience           pa              na
##  Memory     :149   Instructor        :149   Min.   :11.00   Min.   :10.00
##  Score      :274   Juried Recital    : 44   1st Qu.:26.00   1st Qu.:12.00
##  Unspecified: 74   Public Performance:204   Median :32.00   Median :15.00
##                    Student(s)        :100   Mean   :32.27   Mean   :16.21
##                                             3rd Qu.:38.00   3rd Qu.:19.00
##                                             Max.   :50.00   Max.   :35.00
##       age           gender                        instrument   years_study
##  Min.   :18.00   Female:339   keyboard (piano or organ): 75   Min.   : 2.000
##  1st Qu.:19.00   Male  :158   orchestral instrument    :235   1st Qu.: 5.000
##  Median :20.00                voice                    :187   Median : 7.000
##  Mean   :19.75                                                Mean   : 8.143
##  3rd Qu.:21.00                                                3rd Qu.:11.000
##  Max.   :22.00                                                Max.   :17.000
##      mpqab           mpqsr           mpqpem          mpqnem
##  Min.   : 7.00   Min.   : 1.00   Min.   :23.00   Min.   :11.00
##  1st Qu.:17.00   1st Qu.: 7.00   1st Qu.:41.00   1st Qu.:24.00
##  Median :23.00   Median :11.00   Median :52.00   Median :28.00
##  Mean   :21.99   Mean   :11.84   Mean   :48.14   Mean   :31.63
##  3rd Qu.:27.00   3rd Qu.:16.00   3rd Qu.:57.00   3rd Qu.:40.00  
##  Max.   :31.00   Max.   :22.00   Max.   :68.00   Max.   :50.00  
##      mpqcon     
##  Min.   :22.00  
##  1st Qu.:39.00  
##  Median :52.00  
##  Mean   :51.83  
##  3rd Qu.:64.00  
##  Max.   :88.00
```

```
music <- music[music$perform_type == "Solo", ]
```

- `id` = unique musician identification number
- `diary` = cumulative total of diaries filled out by musician (level I; timing metric)
- `audience` = who attended performance (Instructor, Public, Students, or Juried) (level I)
- `na` = negative affect score from PANAS (level I)
- `gender` = musician gender (level II)
- `instrument` = Voice, Orchestral, or Piano (level II)
- `mpqab` = absorption subscale from MPQ (level II)
- `mpqpem` = positive emotionality (PEM) composite scale from MPQ (level II)
- `mpqnem` = negative emotionality (NEM) composite scale from MPQ (level II)

```
levels(music$instrument) <- c("keyboard", "orch_instr", "voice")

lmmt1 <- lmertree(na ~ diary | (1|id) | gender + instrument +
                    mpqab + mpqpem + mpqnem, data = music, cluster = id)
plot(lmmt1, fitted = "marginal", which = "tree", gp = gpar(cex = .7))
```



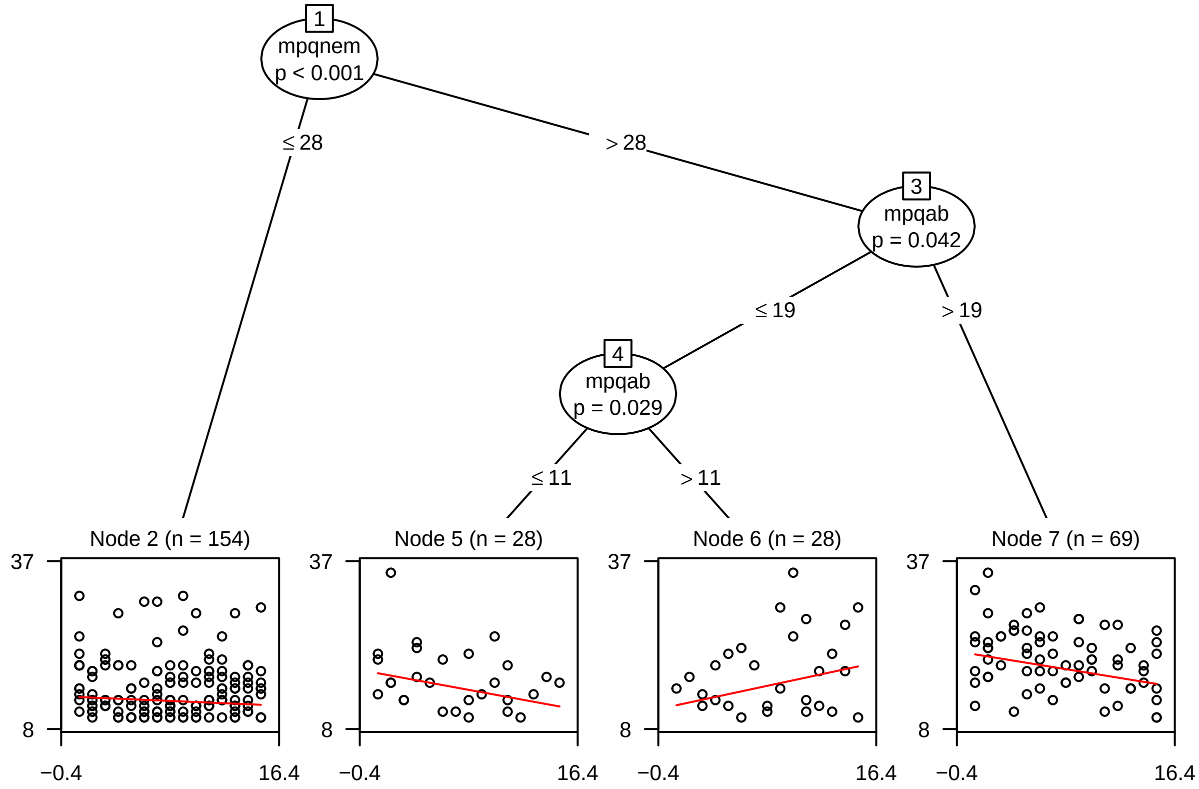
```
fixef(lmmt1, which = "tree")
```

```
##   (Intercept)       diary
## 2    16.82887 -0.14059527
## 3    19.45945  0.03539297
```

```
fixef(lmmt1, which = "global")
```

```
## named numeric(0)
```

```
lmmt2 <- lmertree(na ~ diary | audience + (1|id) | gender + instrument +
                       mpqab + mpqpem + mpqnem, data = music, cluster = id)
plot(lmmt2, fitted = "marginal", which = "tree", gp = gpar(cex = .7))
```


1
mpqnem
p < 0.001
≤ 28
> 28
3
mpqab
p = 0.042
≤ 19
> 19
4
mpqab
p = 0.029
≤ 11
> 11
Node 2 (n = 154)
Node 5 (n = 28)
Node 6 (n = 28)
Node 7 (n = 69)
37
8
−0.4
16.4

Further reading

The vignette of package **glmertree** provides further info on how the GLMM tree models can be further customized, and checks on model fit can be performed. You can access it on https://cran.r-project.org/web/packages/glmertree/vignettes/glmertree.pdf or in **R** by typing:

```
vignette("glmertree")
```

Furthermore, Fokkema, Smits, Zeileis, Hothorn & Kelderman (2018) provides an in-depth technical discussion of GLMM trees, while Fokkema, Edbrooke-Childs & Wolpert (2020) provides a less technical introduction.

References

Curran, P. J., Stice, E., & Chassin, L. (1997). The relation between adolescent alcohol use and peer alcohol use: a longitudinal random coefficients model. *Journal of Consulting and Clinical Psychology, 65*(1), 130.

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., & Kelderman, H. (2018). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behavior Research Methods, 50*(5), 2016-2034. https://doi.org/10.3758/s13428-017-0971-x

Fokkema, M., Edbrooke-Childs, J., & Wolpert, M. (2020). Generalized linear mixed-model (GLMM) trees: A flexible decision-tree method for multilevel and longitudinal data. *Psychotherapy Research,31*(3), 329-341. https://doi.org/10.1080/10503307.2020.1785037

Lucock, M., Barkham, M., Donohoe, G., Kellett, S., McMillan, D., Mullaney, S., … & Delgadillo, J. (2017). The role of Practice Research Networks (PRN) in the development and implementation of evidence: The Northern improving access to psychological therapies PRN case study. *Administration and Policy in Mental Health and Mental Health Services Research, 44*(6), 919-931.

Sadler, M. E., & Miller, C. J. (2010). Performance anxiety: A longitudinal study of the roles of personality and experience in musicians. *Social Psychological and Personality Science, 1*(3), 280-287.

Singer, J.D. & Willett, J.B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* Oxford University Press.

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics, 17*(2), 492-514. https://doi.org/10.1198/106186008X319331