Examples to use TSC in Fixed Effects Models

Moritz Berger

2017-07-13

CTB/McGraw-Hill

The data set contains results of an achievement test that measures different objectives and subskills of subjects in mathematics and science. The students had to respond to 56 multiple-choice items (31 mathematics, 25 science). For a description of the original data, see [1].

1. Load data

```
library("structree")
data(CTB, package="structree")
```

2. Overview of the data

```
dim(CTB)
```

```
## [1] 1500    9
```

```
str(CTB)
```

```
## 'data.frame':    1500 obs. of  9 variables:
##  $ score   : num  39 35 38 32 40 31 38 34 32 34 ...
##  $ school  : Factor w/ 35 levels "1","2","3","4",..: 15 15 15 15 15 15 15 15 15 15 ...
##  $ size    : num  300 300 300 300 300 300 300 300 300 300 ...
##  $ bachelor: num  0.11 0.11 0.11 0.11 0.11 0.11 0.11 0.11 0.11 0.11 ...
##  $ born    : num  -0.5 -0.5 -0.5 -0.5 -0.5 -0.5 -0.5 -0.5 -0.5 -0.5 ...
##  $ mortgage: num  0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 ...
##  $ language: num  0.36 0.36 0.36 0.36 0.36 0.36 0.36 0.36 0.36 0.36 ...
##  $ type    : Factor w/ 3 levels "1","2","3": 3 3 3 3 3 3 3 3 3 3 ...
##  $ gender  : num  0 0 1 1 0 1 0 0 0 1 ...
```

```
nlevels(CTB$school)
```

```
## [1] 35
```

```
table(CTB$school)
```

```
##
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18
##  51  18  14  36  35  57  96  13  25  98  49  18 107  27  23  47  76  18
##  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35
##  43 114  69  12  13  13  99  47  19  51  46  25  17  10   5  91  18
```

There are 1500 grade 8 students from 35 schools. The response variable score is the overall test score, defined as the number of correctly solved items. Several variables characterise the schools and the students. For the analysis we use the covariate gender (male: 0, female: 1).

3. Estimation of the model

```
mod_CTB <- structree(score ~ tr(1 | school) + gender, data = CTB,
    family = gaussian, stop_criterion = "pvalue", splits_max = 34,
    alpha = 0.05, trace = FALSE)
```

```
# print
mod_CTB
```

```
## Tree Structured Clustering of observation units:
##
## Call: structree.default(formula = score ~ tr(1 | school) + gender,
## data = CTB, family = gaussian, stop_criterion = "pvalue",
## splits_max = 34, alpha = 0.05, trace = FALSE)
##
## Second-level unit: school
## Unit specific effects for: Intercept
## Fixed effects for: gender
## Number of Splits: 5
##
## Number of Groups:
## Intercept
##         6
```

For school-specific intercepts one has to enter `tr(1|school)` into the formula.

4. Number of Splits

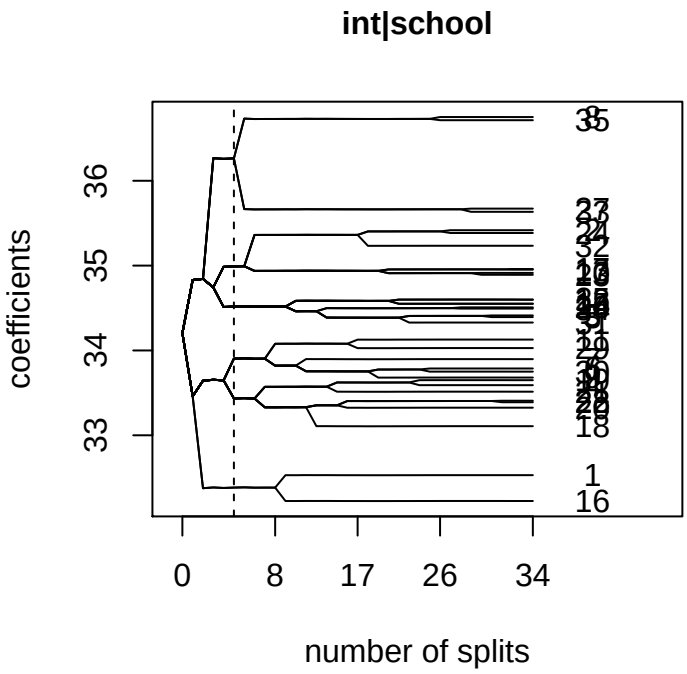
```
mod_CTB$opts
```

```
## [1] 5
```

The algorithm performs five splits, that is, forms six clusters regarding the intercept.

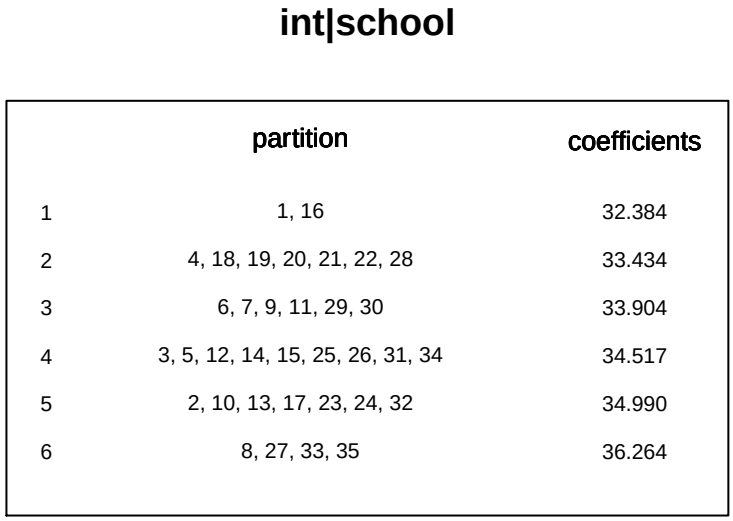
5. Paths of Coefficients

```
plot(mod_CTB, paths=TRUE)
```



6. Estimated Clusters

```
plot(mod_CTB, result=TRUE, cex.txt=0.7, cex.main=1.2)
```



7. Estimated Coefficients

```
coef(mod_CTB)
```

```
## int|school1 int|school2 int|school3 int|school4 int|school5 int|school6
## 32.38370614 33.43440051 33.90350744 34.51706681 34.99024285 36.26388840
##      gender
## -0.08379539
```

National Survey in Guatemala

The data set contains observations of children that were born in the 5-year-period before the National Survey of Maternal and Child Health in Guatemala in 1987. The data was also analysed by [2].

1. Load data

```
library("structree")
data(guPrenat, package="structree")
```

2. Overview of the data

```
dim(guPrenat)
```

```
## [1] 1211    9
```

```
str(guPrenat)
```

```
## 'data.frame':    1211 obs. of  9 variables:
##  $ cluster  : Factor w/ 45 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
##  $ prenat   : num  0 0 1 1 1 1 1 1 1 1 ...
##  $ motherAge: num  0 0 1 0 0 0 1 1 0 1 ...
##  $ indig    : Factor w/ 3 levels "Ladino","NoSpa",..: 3 3 3 3 3 3 1 3 3 3 ...
##  $ momEd    : Factor w/ 3 levels "None","Primary",..: 1 1 1 1 1 1 2 2 2 2 ...
##  $ husEd    : Factor w/ 4 levels "None","Primary",..: 2 2 4 2 2 2 2 2 2 2 ...
##  $ husEmpl  : Factor w/ 5 levels "Unskilled","Professional",..: 5 5 5 5 5 5 5 5 5 1 ...
##  $ toilet   : num  0 0 0 1 1 1 0 1 1 0 ...
##  $ TV       : Factor w/ 3 levels "None","not daily",..: 2 2 2 1 1 1 3 2 2 1 ...
```

```
nlevels(guPrenat$cluster)
```

```
## [1] 45
```

```
table(guPrenat$cluster)
```

```
##
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
## 27 26 21 26 31 24 28 27 22 31 30 28 21 27 29 30 24 21 24 25 29 36 27 29 24
## 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45
## 21 33 22 25 50 33 30 28 21 23 27 25 25 35 22 26 25 23 24 26
```

There are 1211 children living in 45 communities. The response variable prenat is the indicator for modern prenatal care (prenat=1), for example by doctors or nurses, instead of traditional prenatal care (prenat=0). Several variables characterise the children's mothers and their families.

3. Estimation of the model

```
mod_gua <- structree(prenat ~ tr(1 | cluster) + indig + momEd + husEd + husEmpl +
    TV + motherAge + toilet, data = guPrenat, family = binomial(link = "logit"),
    stop_criterion = "pvalue", splits_max = 10, alpha = 0.05, trace = FALSE)

# print
mod_gua
```

```
## Tree Structured Clustering of observation units:
##
## Call: structree.default(formula = prenat ~ tr(1 | cluster) + indig +
## momEd + husEd + husEmpl + TV + motherAge + toilet, data = guPrenat,
## family = binomial(link = "logit"), stop_criterion = "pvalue",
## splits_max = 10, alpha = 0.05, trace = FALSE)
##
## Second-level unit: cluster
## Unit specific effects for: Intercept
## Fixed effects for: indig, momEd, husEd, husEmpl, TV, motherAge, toilet
## Number of Splits: 2
##
## Number of Groups:
## Intercept
##         3
```

For community-specific intercepts one has to enter `tr(1|cluster)` into the formula.

4. Number of Splits

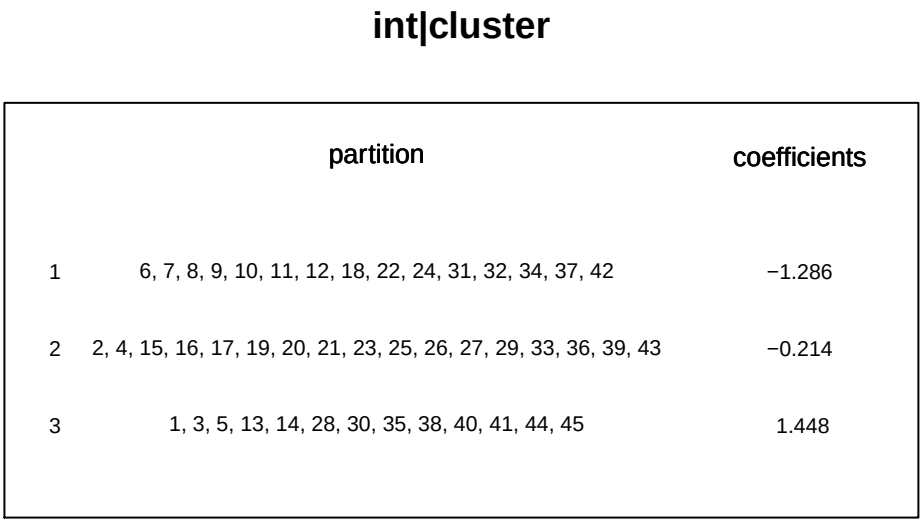
```
mod_gua$opts
```

```
## [1] 2
```

The algorithm performs two splits, that is, forms two clusters regarding the intercept.

5. Estimated Clusters

```
plot(mod_gua, result=TRUE, cex.txt=0.7, cex.main=1.2)
```



6. Estimated Coefficients

```
coef(mod_gua)
```

```
##        int|cluster1        int|cluster2        int|cluster3
##          -1.28550852         -0.21405291          1.44821948
##          indigNoSpa        indigSpanish        momEdPrimary
##          -1.09025050         -0.43352670          0.67346718
##     momEdSecondary+        husEdPrimary     husEdSecondary+
##           1.40487412          0.81742073          0.04867560
##        husEdUnknown husEmplProfessional  husEmplAgri (self)
##           0.52045664         -0.09467386         -0.06521850
##  husEmplAgri (empl)      husEmplSkilled         TVnot daily
##          -0.10047073         -0.12537638          0.22553671
##              TVdaily           motherAge              toilet
##           0.92808865          0.06144982         -1.00835578
```

References

[1] De Boeck, P. and M. Wilson (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer Verlag.

[2] Rodriguez, G. and N. Goldman (2001). Improved estimation procedures for multilevel models with binary response: A case-study. *Journal of the Royal Statistical Society. Series A (Statistics in Society) 164(2)*, 339-355.