factorMerger: hierarchical clustering and model visualization

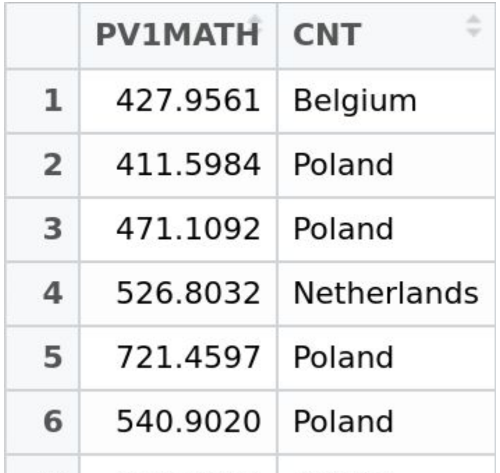
Agnieszka Sitko

University of Warsaw, MI^2 Group

useR!2017 | 06-07-2017

Problem 1

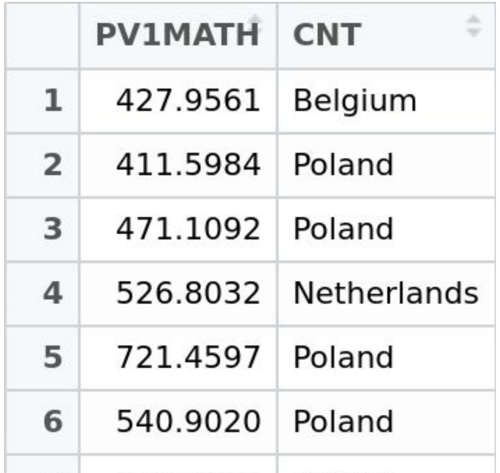
Given a factor C (with k* levels) and a numeric response y* analyze the differences among group means of y.



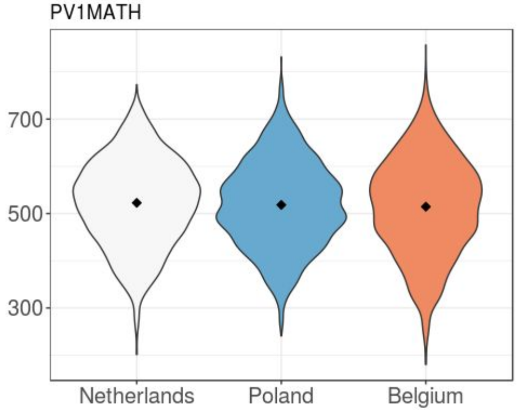
	PV1MATH	CNT
1	427.9561	Belgium
2	411.5984	Poland
3	471.1092	Poland
4	526.8032	Netherlands
5	721.4597	Poland
6	540.9020	Poland

Problem 1

Given a factor C (with k* levels) and a numeric response y* analyze the differences among group means of y.



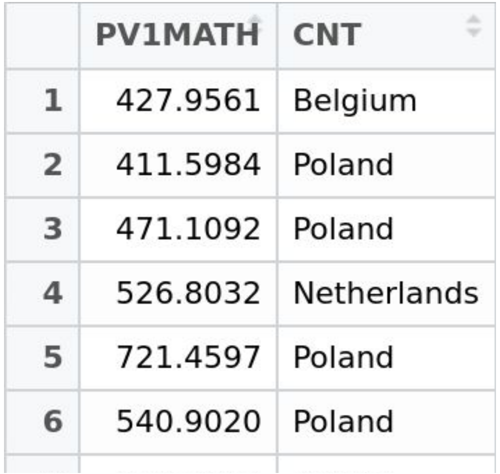
	PV1MATH	CNT
1	427.9561	Belgium
2	411.5984	Poland
3	471.1092	Poland
4	526.8032	Netherlands
5	721.4597	Poland
6	540.9020	Poland



* k is greater than 2,
* y is normally distributed.

Problem 1

Given a factor *C* (with *k** levels) and a numeric response *y** analyze the differences among group means of *y*.

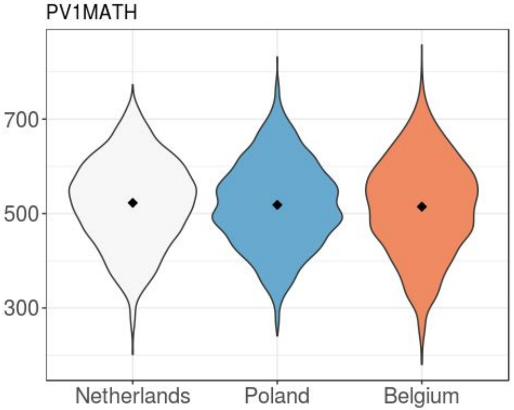


	PV1MATH	CNT
1	427.9561	Belgium
2	411.5984	Poland
3	471.1092	Poland
4	526.8032	Netherlands
5	721.4597	Poland
6	540.9020	Poland

Solution

That’s easy!

Let’s run ANOVA and then post-hoc tests.



* k is greater than 2,
* y is normally distributed.

Solution

Let’s run ANOVA and then post-hoc tests.

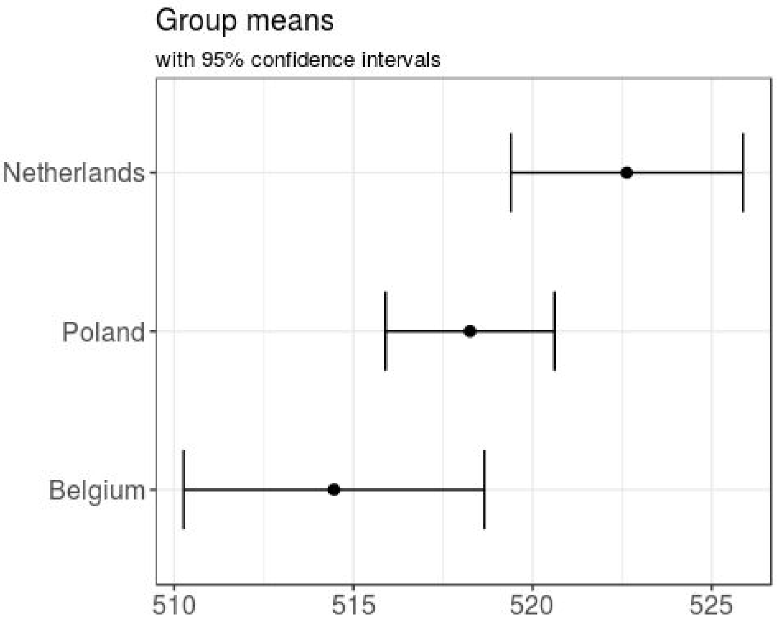
Solution

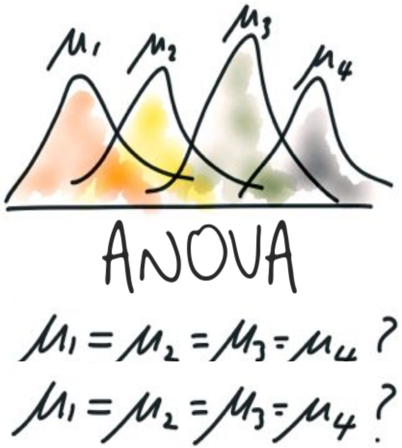
Let’s run ANOVA and then post-hoc tests.

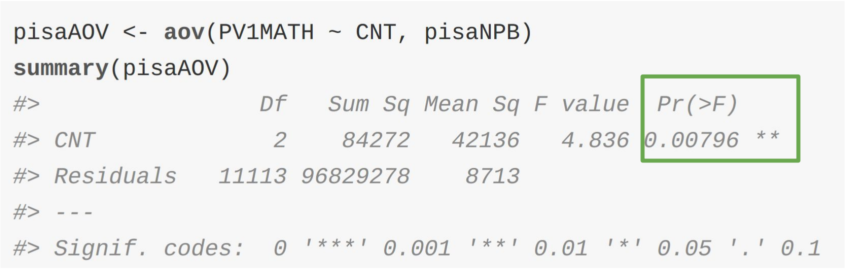
Let’s try this out.

Solution

Let’s run ANOVA and then post-hoc tests.





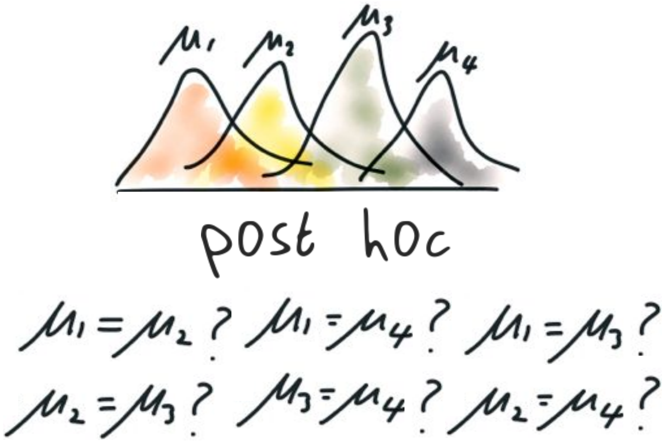


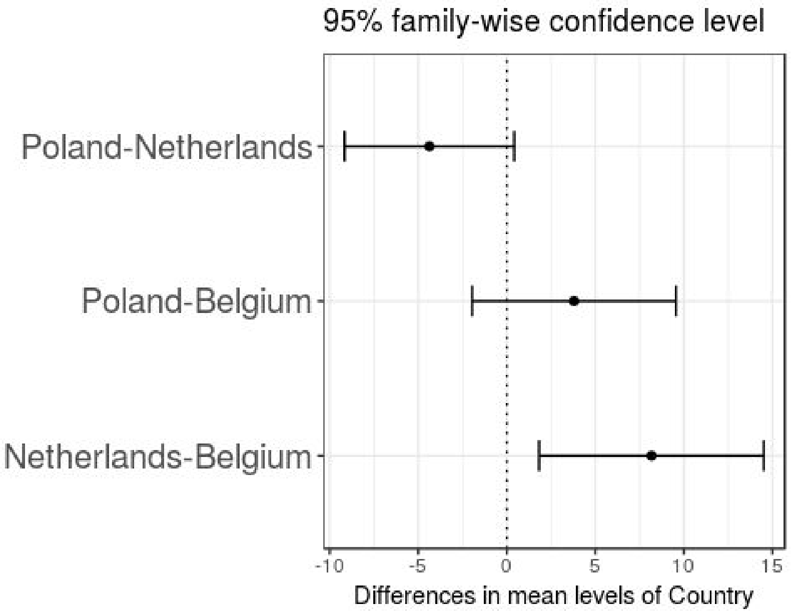
```
pisaAOV <- aov(PV1MATH ~ CNT, pisaNPB)
summary(pisaAOV)
#>                Df   Sum Sq Mean Sq F value  Pr(>F)
#> CNT             2    84272   42136   4.836 0.00796 **
#> Residuals   11113 96829278    8713
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

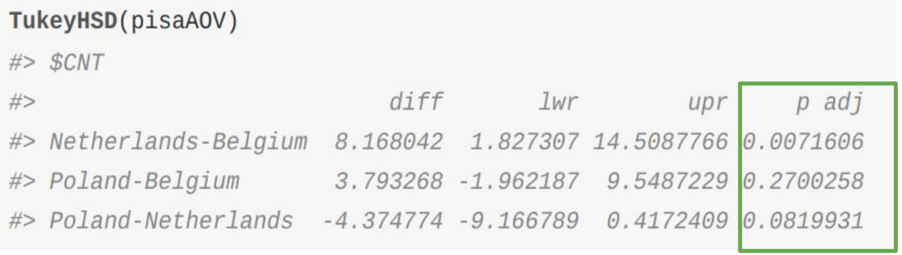
https://www.linkedin.com/pulse/anova-analysis-variance-kumar-p

Solution

Let’s run ANOVA and then post-hoc tests.


μ_1
μ_2
μ_3
μ_4
post hoc
$\mu_1 = \mu_2$? $\mu_1 = \mu_4$? $\mu_1 = \mu_3$?
$\mu_2 = \mu_3$? $\mu_3 = \mu_4$? $\mu_2 = \mu_4$?





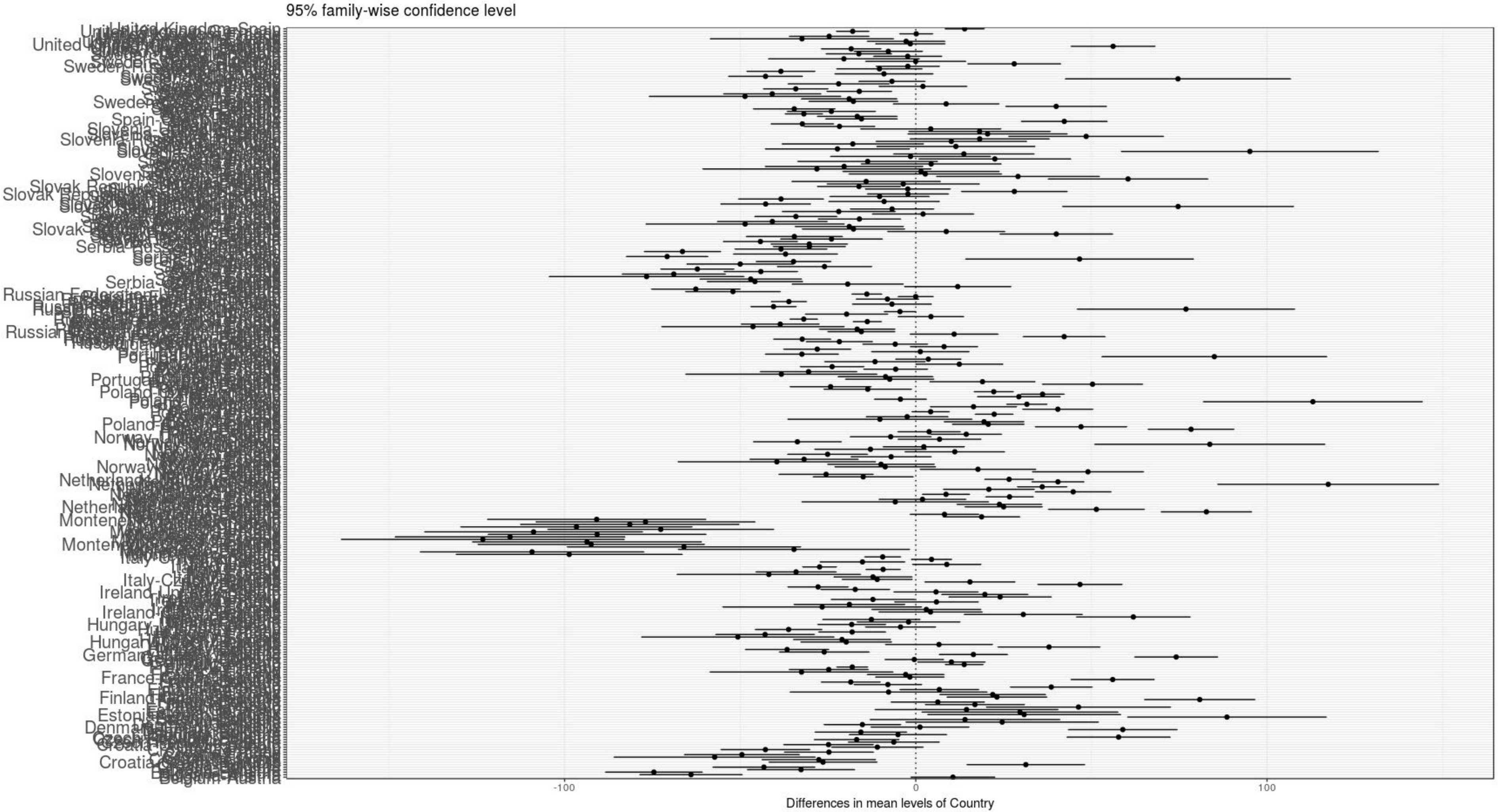
```
TukeyHSD(pisaAOV)
#> $CNT
#>                           diff       lwr        upr     p adj
#> Netherlands-Belgium   8.168042  1.827307 14.5087766 0.0071606
#> Poland-Belgium        3.793268 -1.962187  9.5487229 0.2700258
#> Poland-Netherlands   -4.374774 -9.166789  0.4172409 0.0819931
```

https://www.linkedin.com/pulse/anova-analysis-variance-kumar-p

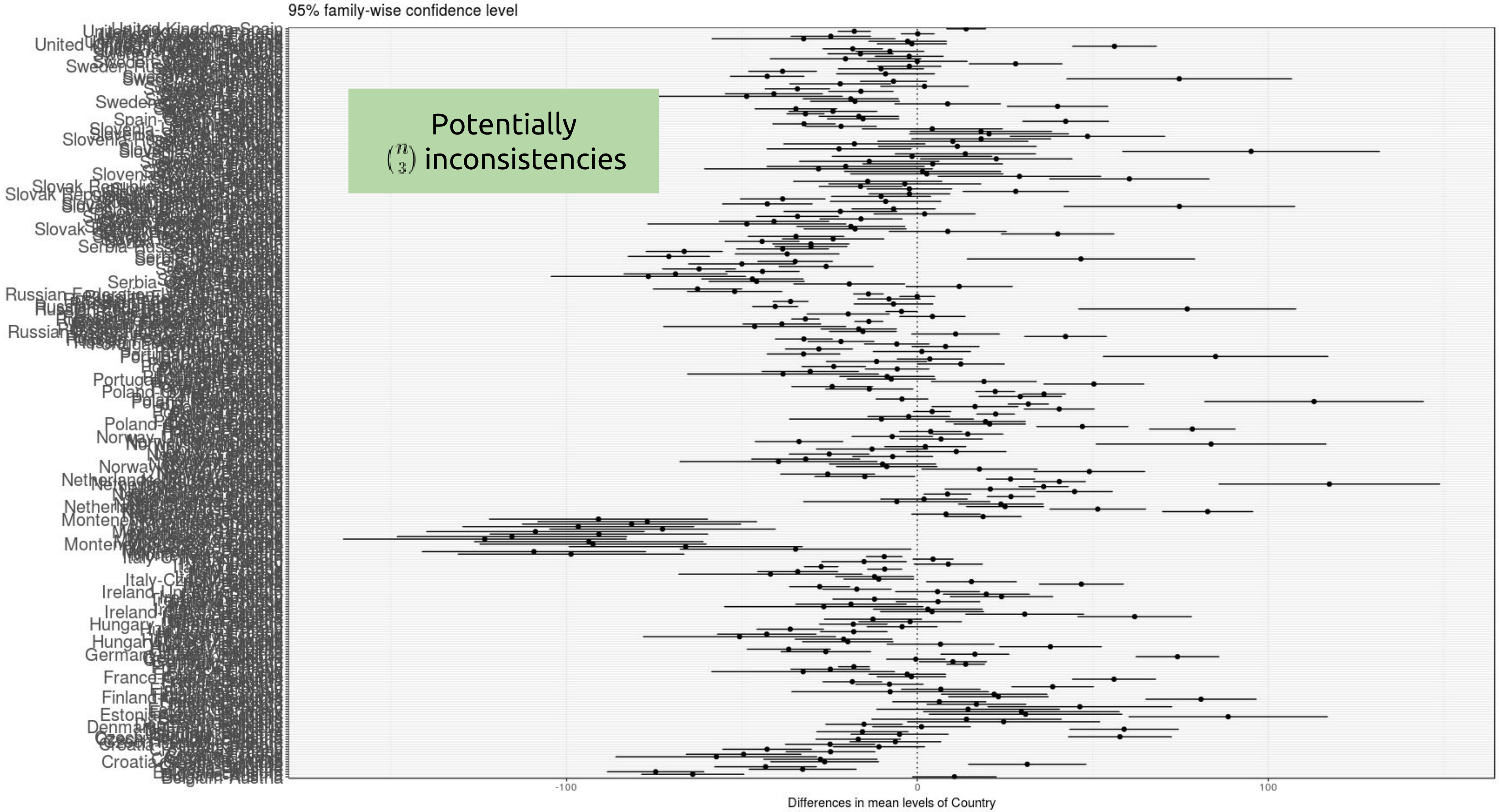
Problem 2

Given a factor *C* (with *k** levels) and
a numeric response *y** analyze the differences
among group means of *y*.

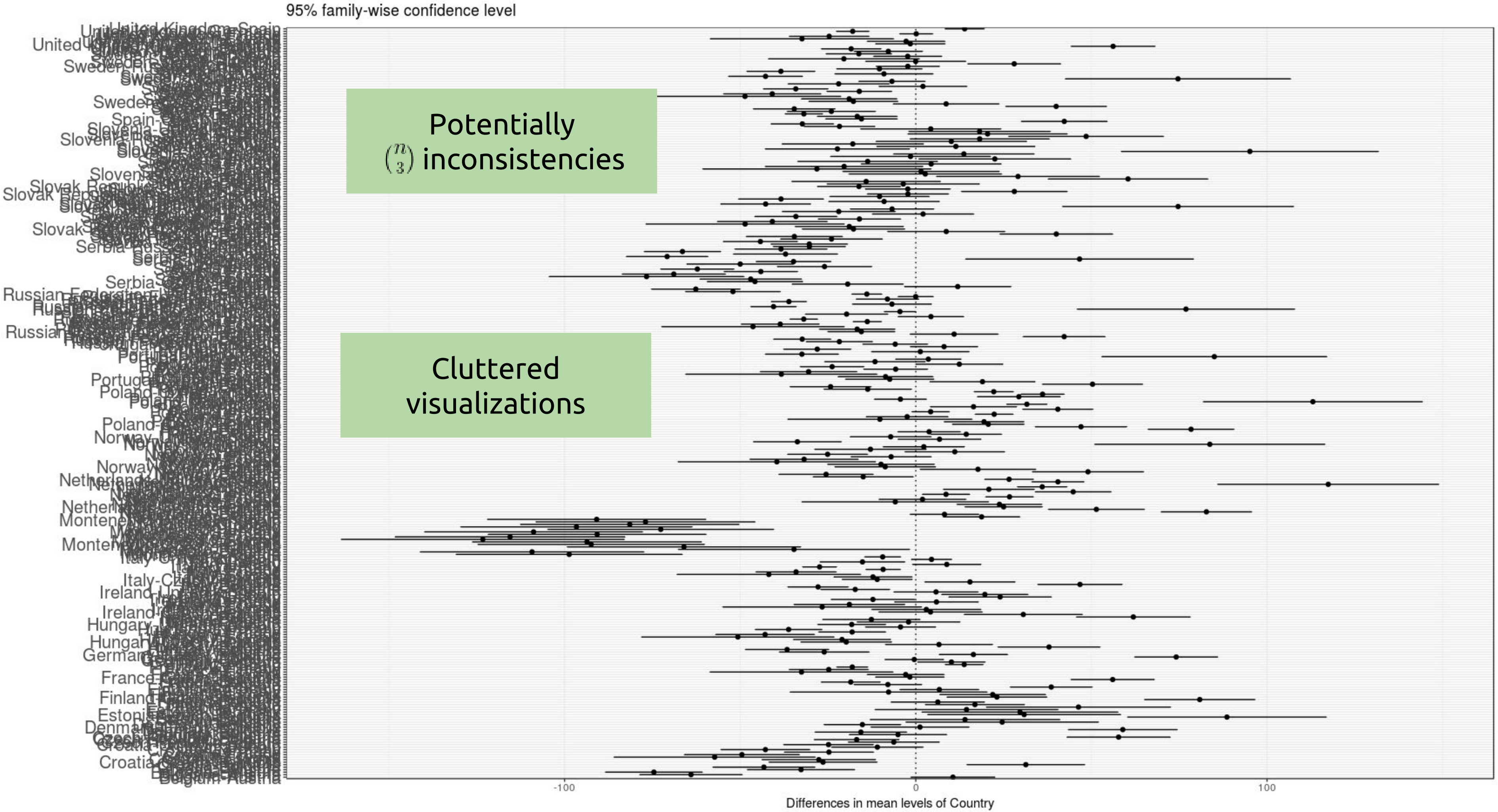
Group levels of *C* into non-overlapping clusters.



95% family-wise confidence level
-100
0
100
Differences in mean levels of Country

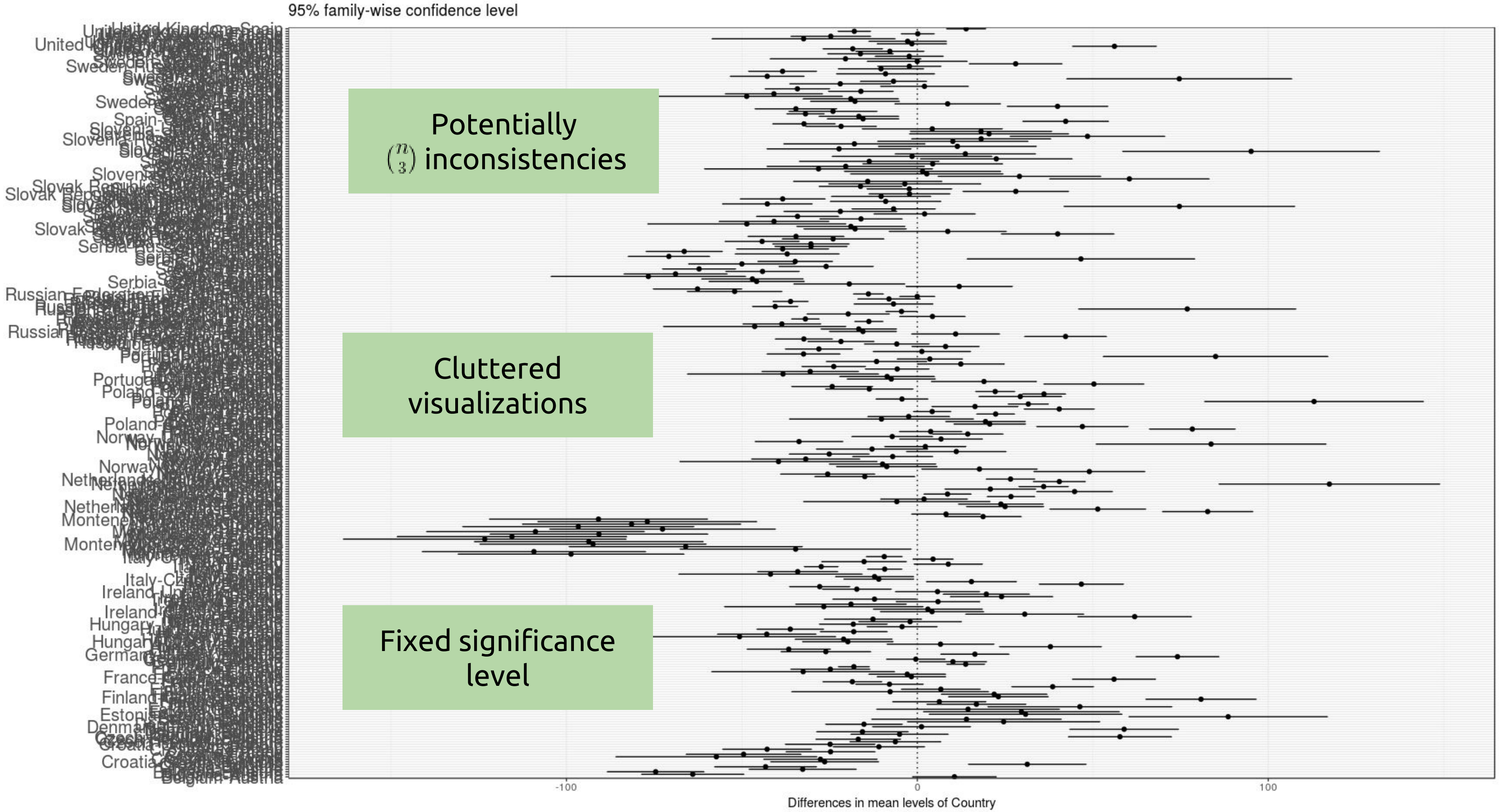


Potentially $\binom{n}{3}$ inconsistencies



Potentially $\binom{n}{3}$ inconsistencies

Cluttered visualizations

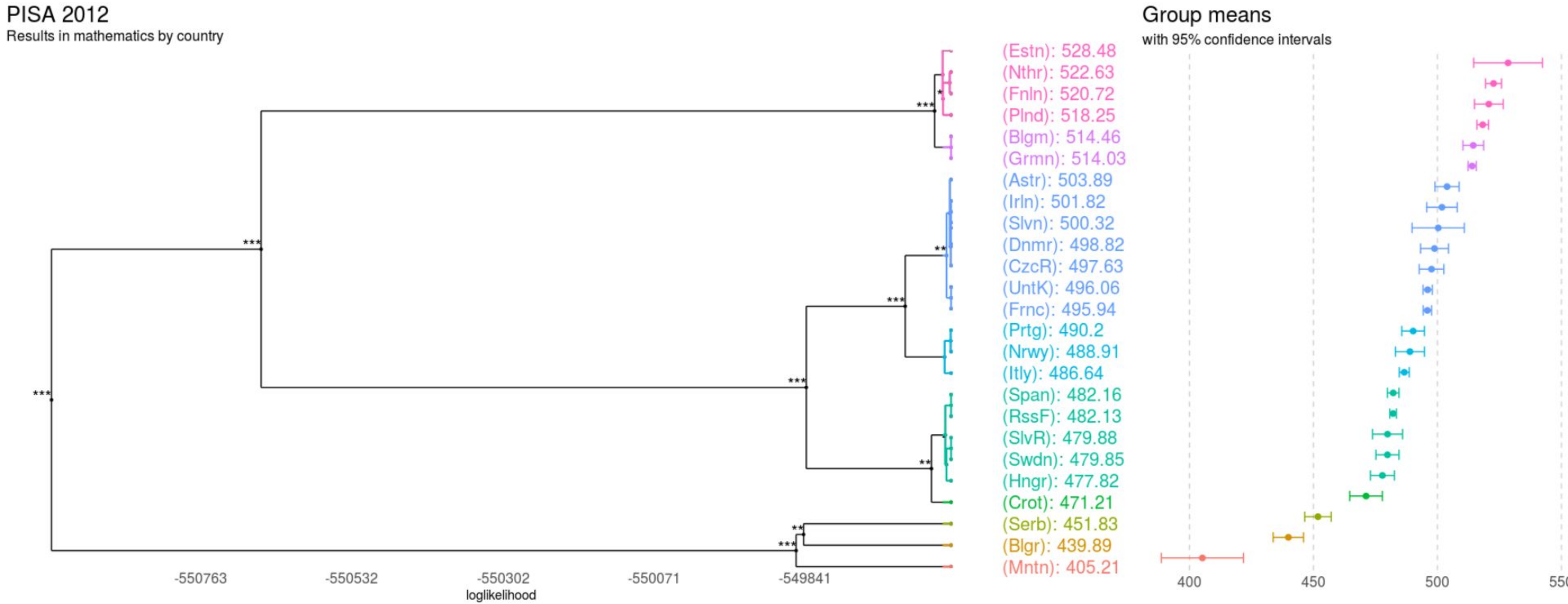


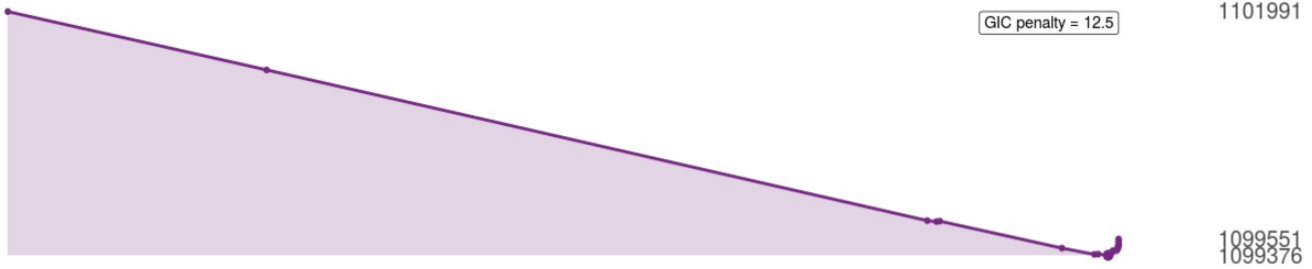
Potentially $\binom{n}{3}$ inconsistencies

Cluttered visualizations

Fixed significance level

Time for factorMerger

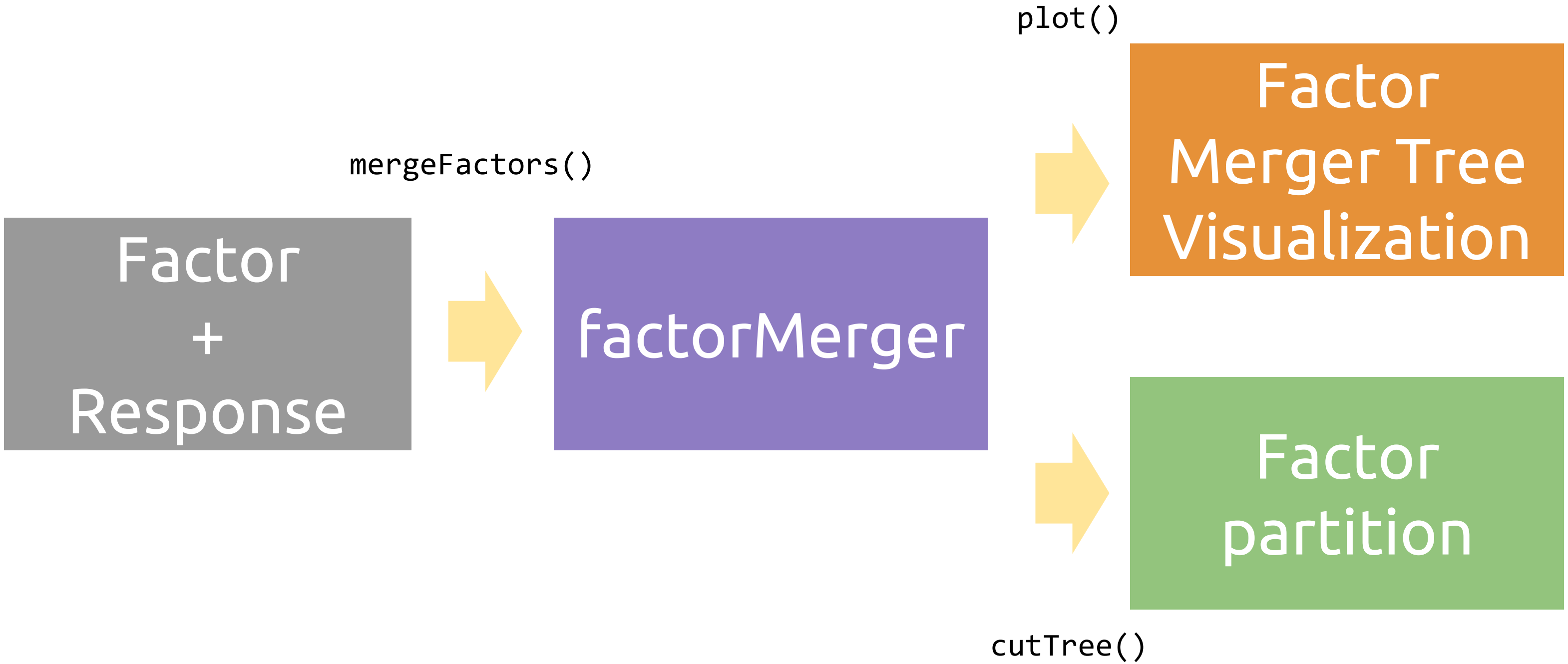




ANOVA table

	Df	F	p-value
factor	25	105939.6	< 2.2e-16
Res	92411		

Working with factorMerger



plot()

Factor Merger Tree Visualization

mergeFactors()

Factor + Response

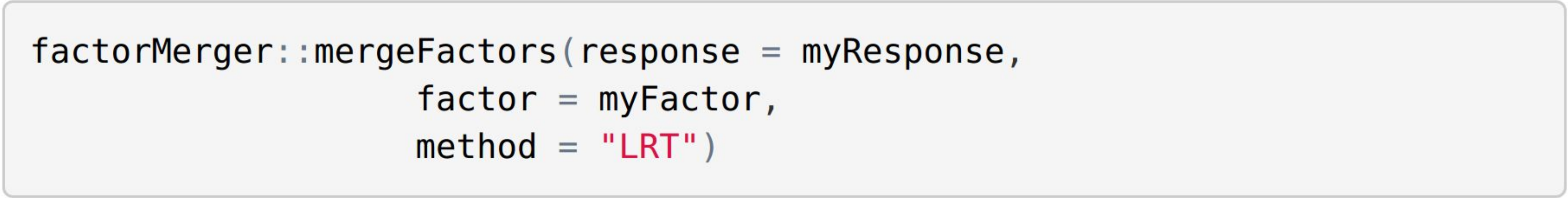
factorMerger

Factor partition

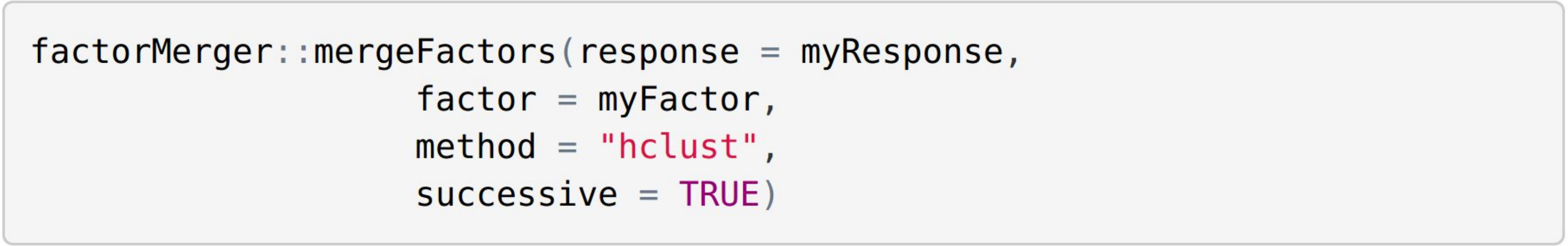
cutTree()

Merge

1. Likelihood Ratio Tests
2. Delete or Merge Regressors



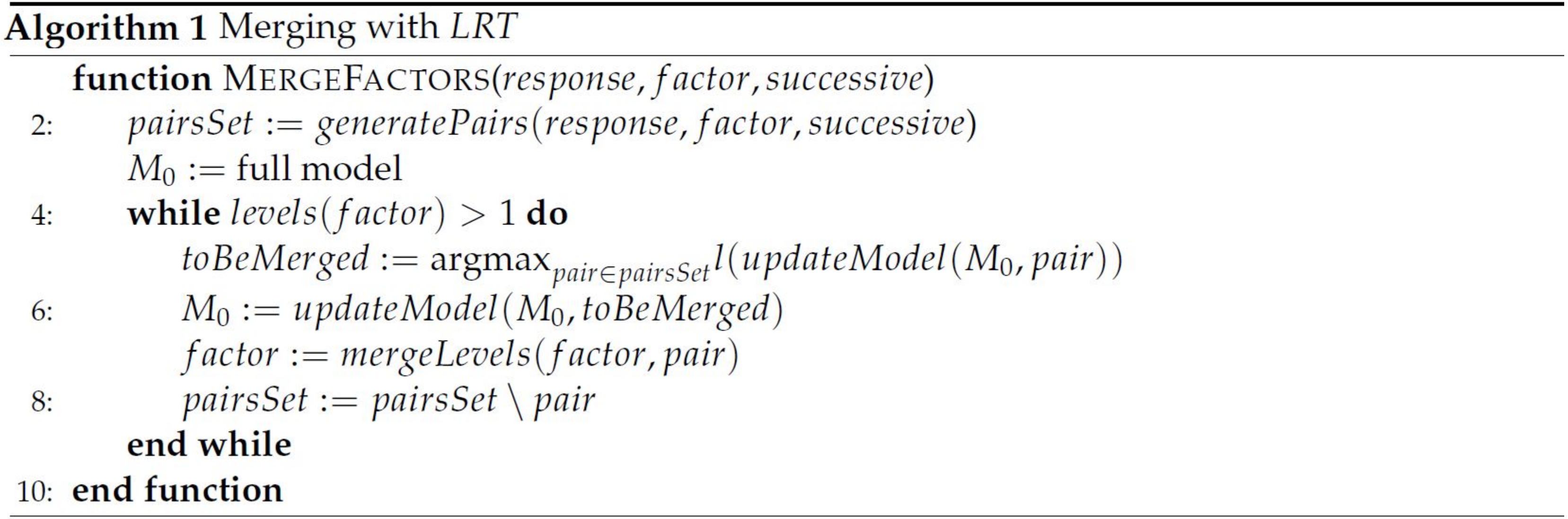
```r
factorMerger::mergeFactors(response = myResponse,
                           factor = myFactor,
                           method = "LRT")
```



```r
factorMerger::mergeFactors(response = myResponse,
                           factor = myFactor,
                           method = "hclust",
                           successive = TRUE)
```

Merge

1. Likelihood Ratio Tests
2. Delete or Merge Regressors

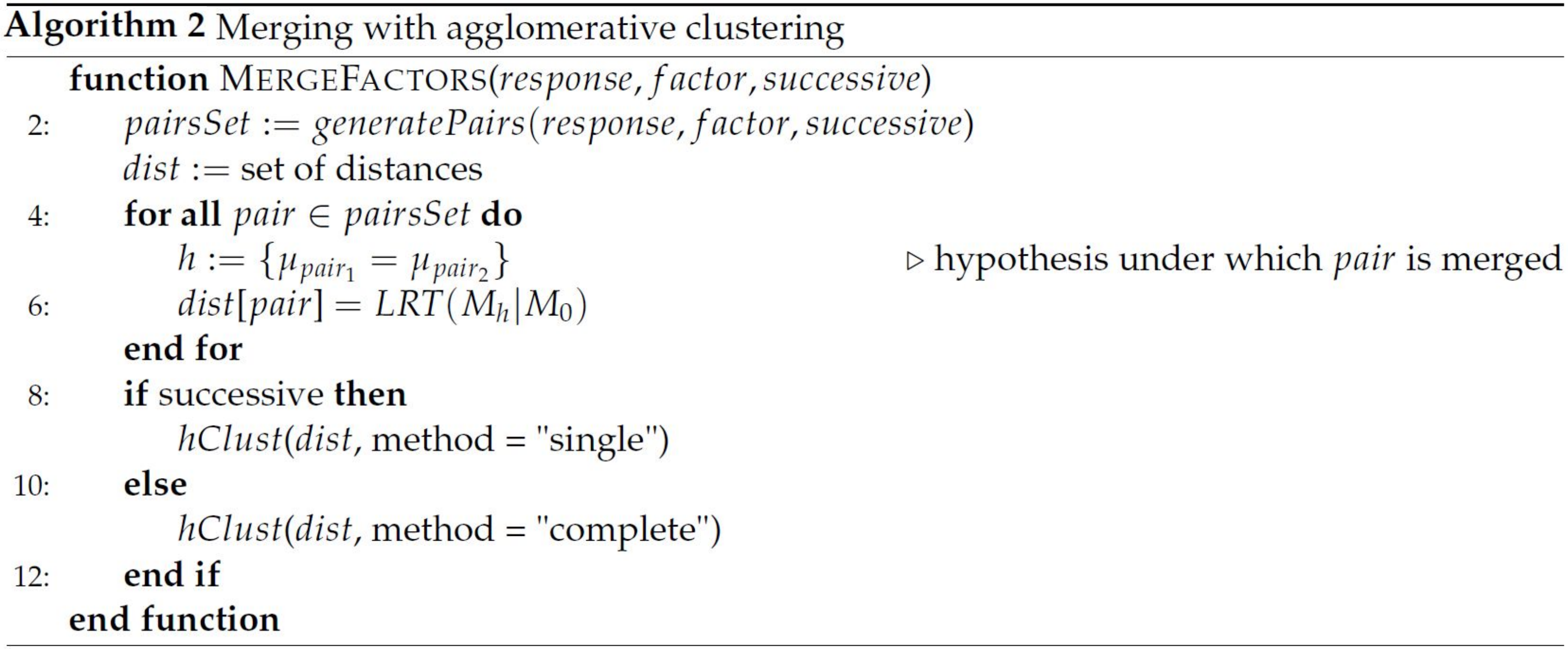


Algorithm 1 Merging with *LRT*

```
function MergeFactors($response, factor, successive$)
    $pairsSet := generatePairs(response, factor, successive)$
    $M_0 :=$ full model
    while $levels(factor) > 1$ do
        $toBeMerged := \text{argmax}_{pair \in pairsSet} l(updateModel(M_0, pair))$
        $M_0 := updateModel(M_0, toBeMerged)$
        $factor := mergeLevels(factor, pair)$
        $pairsSet := pairsSet \setminus pair$
    end while
end function
```

Merge

1. Likelihood Ratio Tests
2. Delete or Merge Regressors



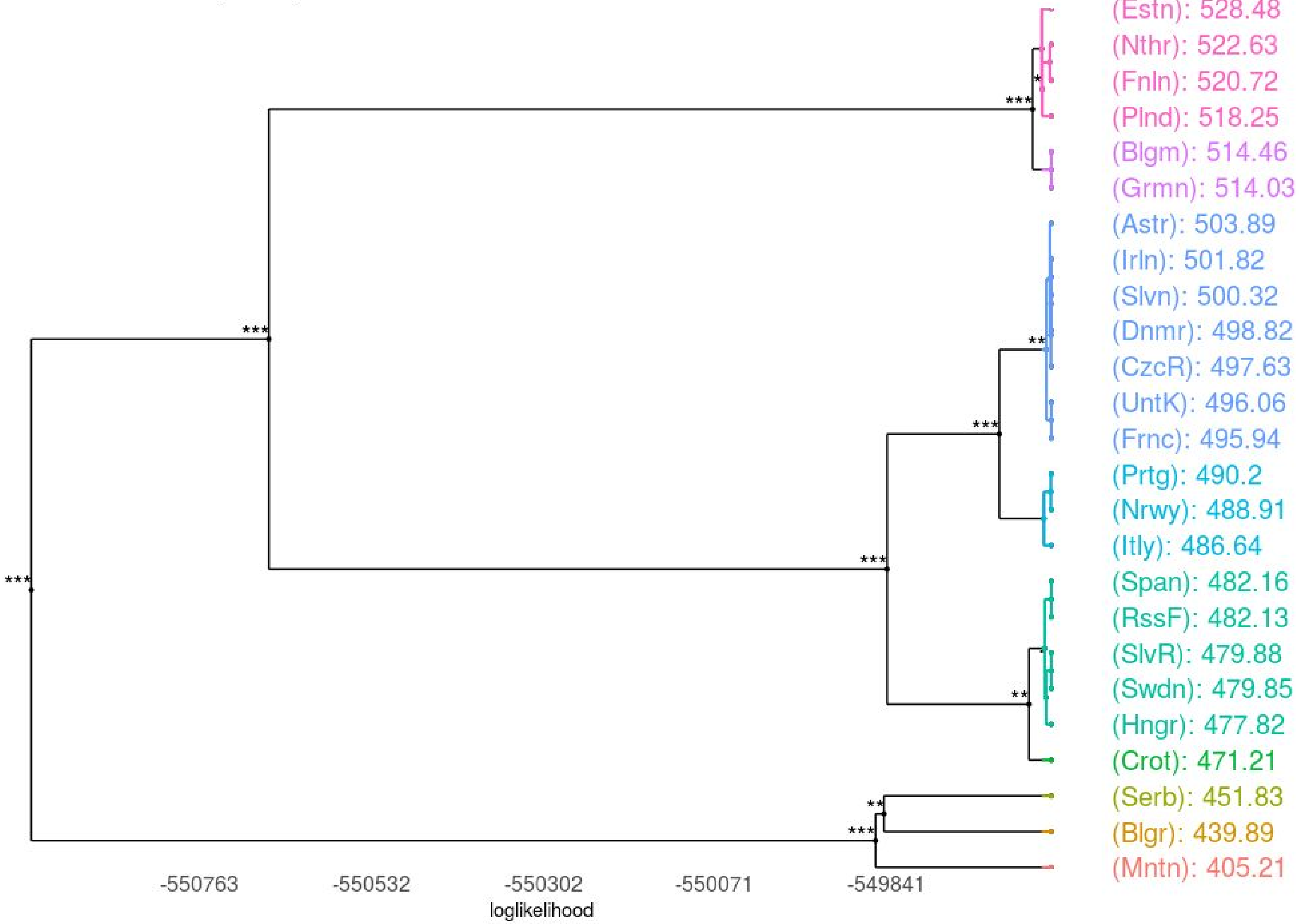
Algorithm 2 Merging with agglomerative clustering

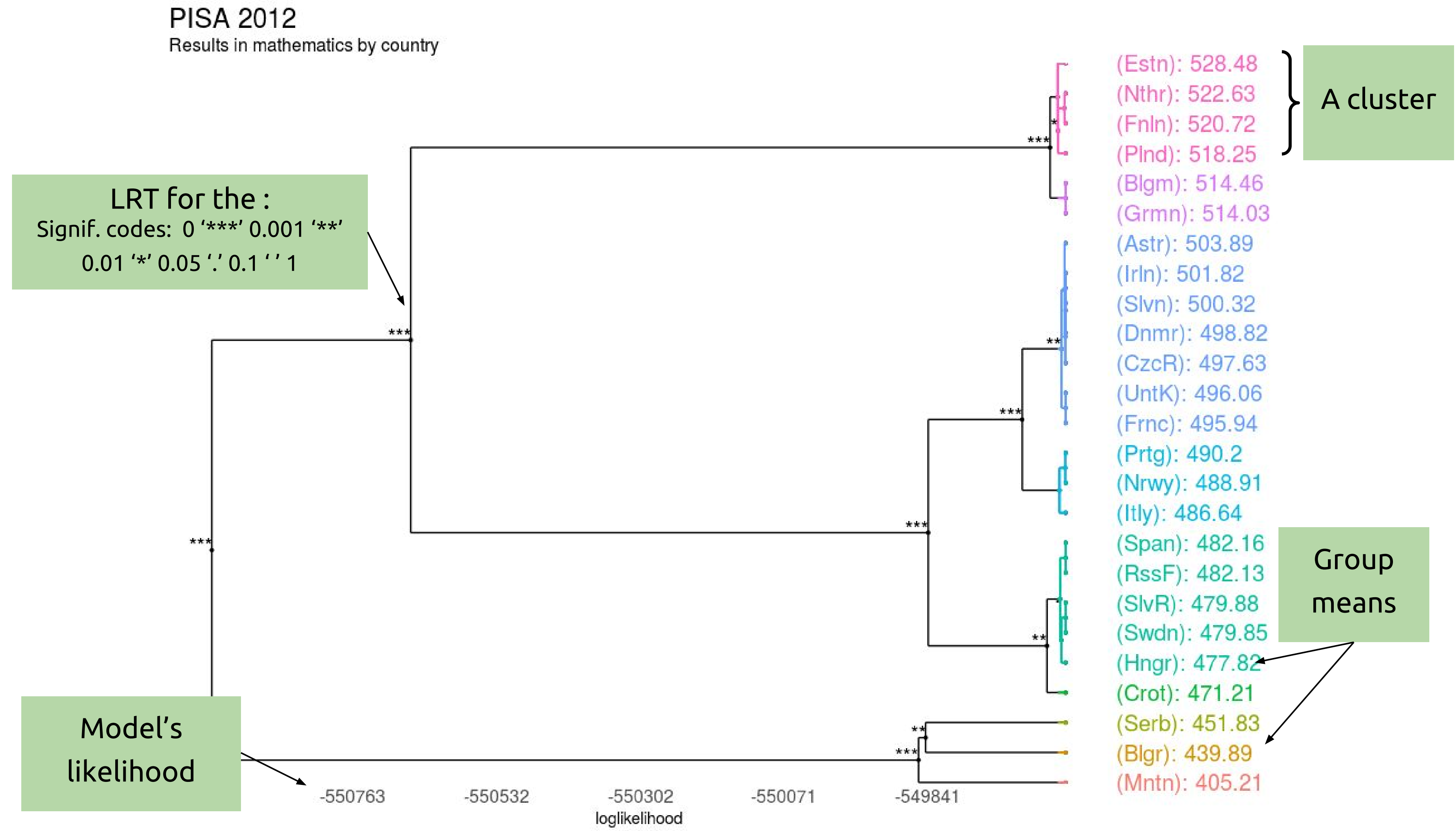
```
function MergeFactors($response, factor, successive$)
    $pairsSet := generatePairs(response, factor, successive)$
    $dist :=$ set of distances
    for all $pair \in pairsSet$ do
        $h := \{\mu_{pair_1} = \mu_{pair_2}\}$        ▷ hypothesis under which $pair$ is merged
        $dist[pair] = LRT(M_h|M_0)$
    end for
    if successive then
        $hClust(dist,$ method = "single")
    else
        $hClust(dist,$ method = "complete")
    end if
end function
```

More about the DMR algorithm: https://arxiv.org/abs/1505.04008

PISA 2012

Results in mathematics by country





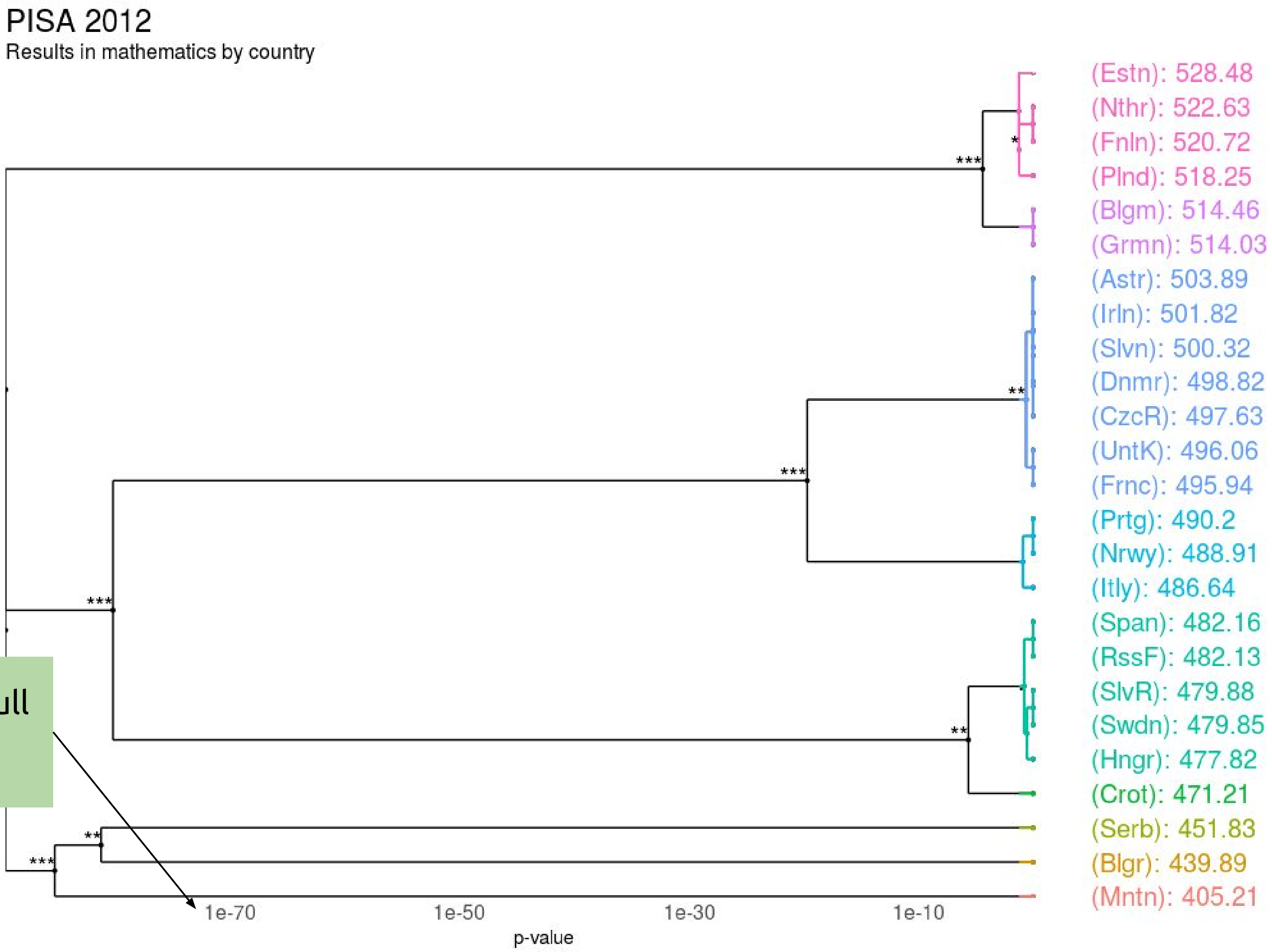
A cluster

LRT for the :

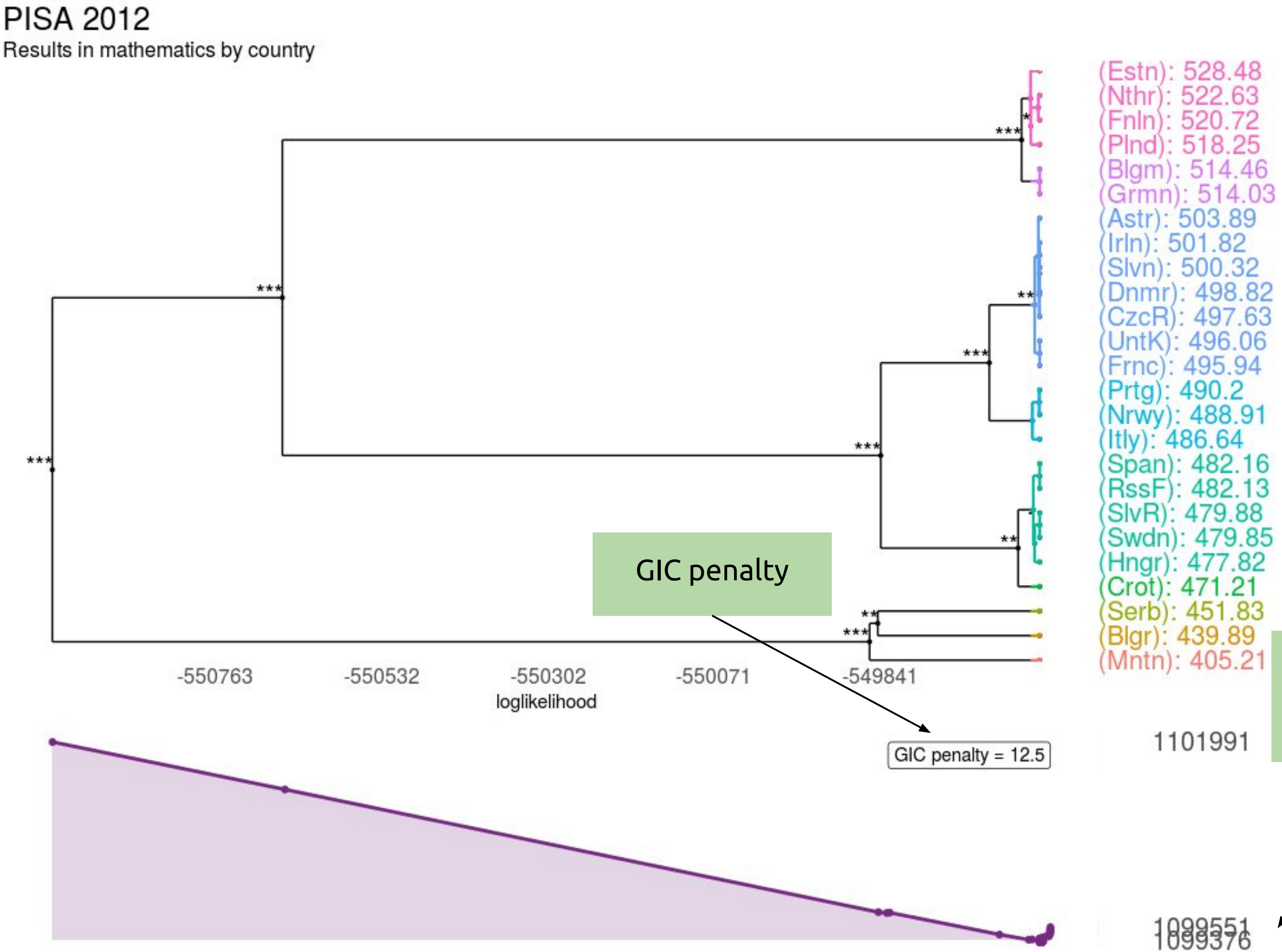
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Group means

Model's likelihood



LRT vs. full model

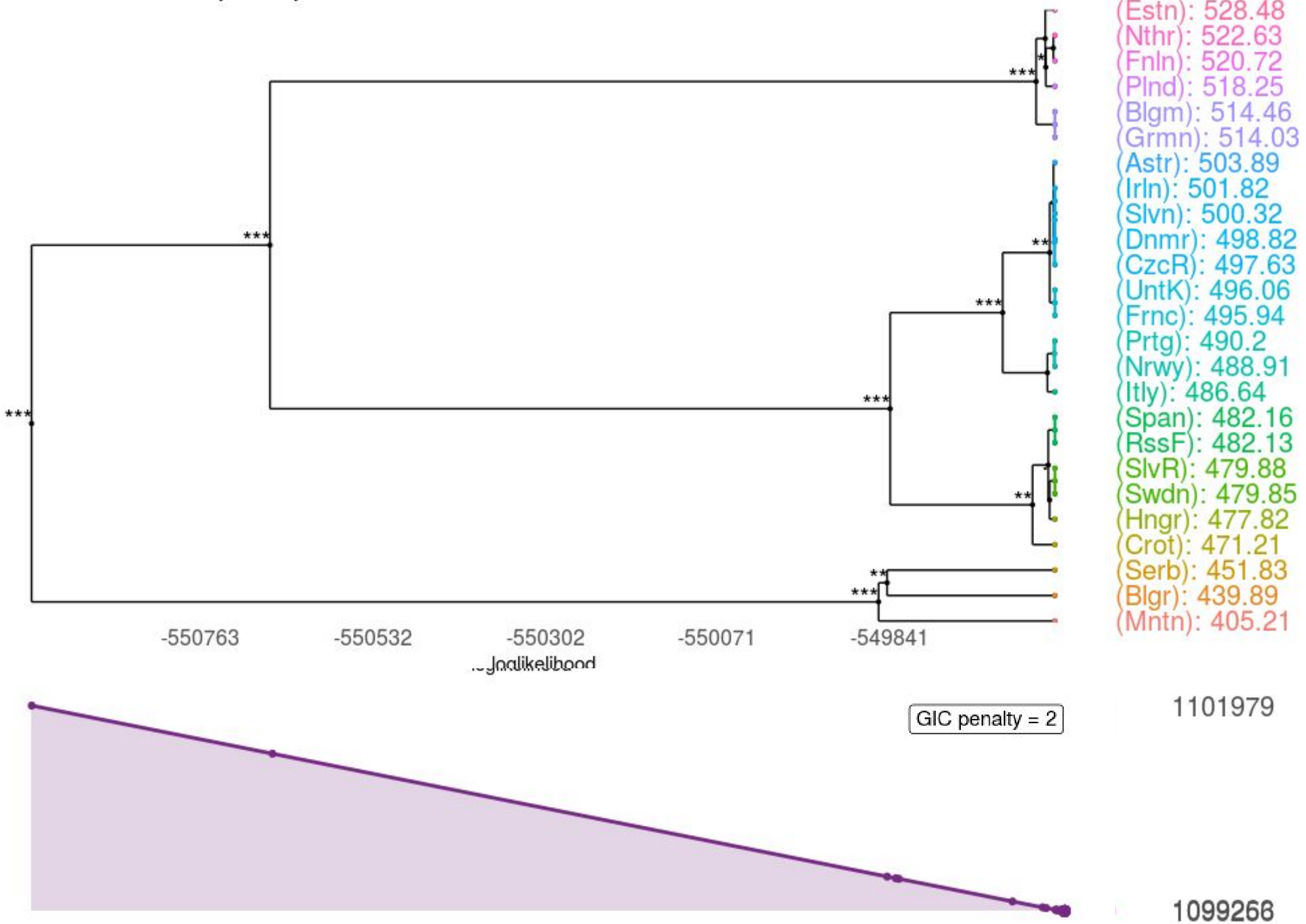


GIC penalty

Models: constant, full best

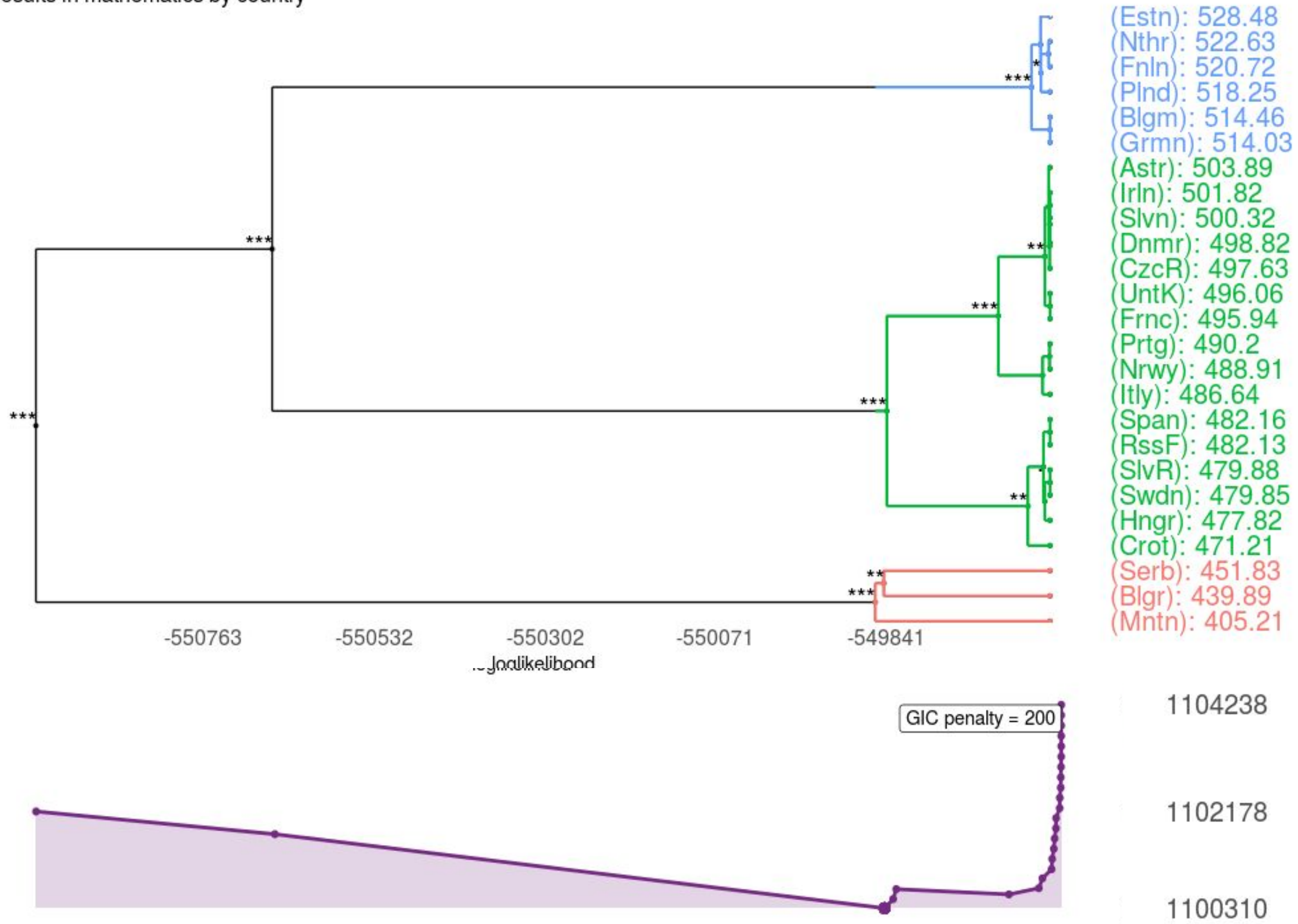
PISA 2012

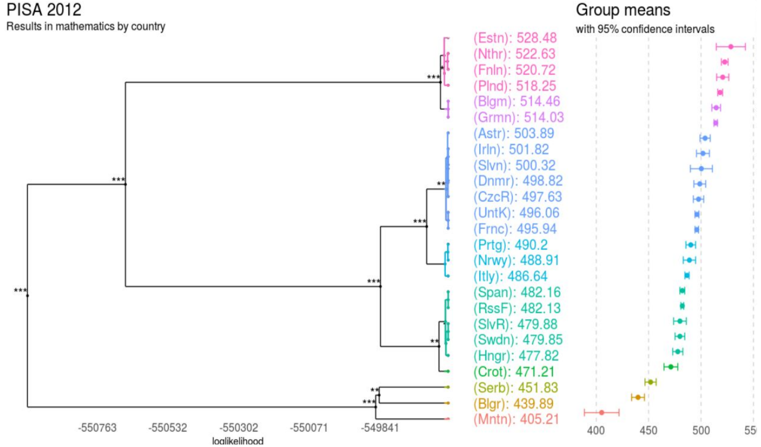
Results in mathematics by country



PISA 2012

Results in mathematics by country



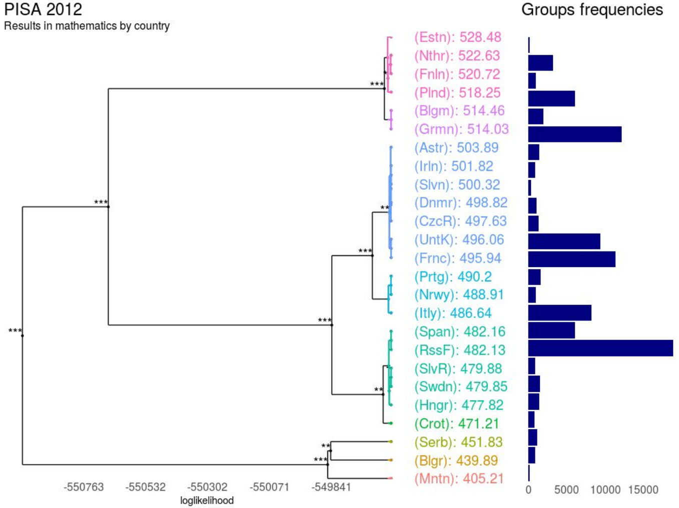


PISA 2012

Results in mathematics by country

Group means

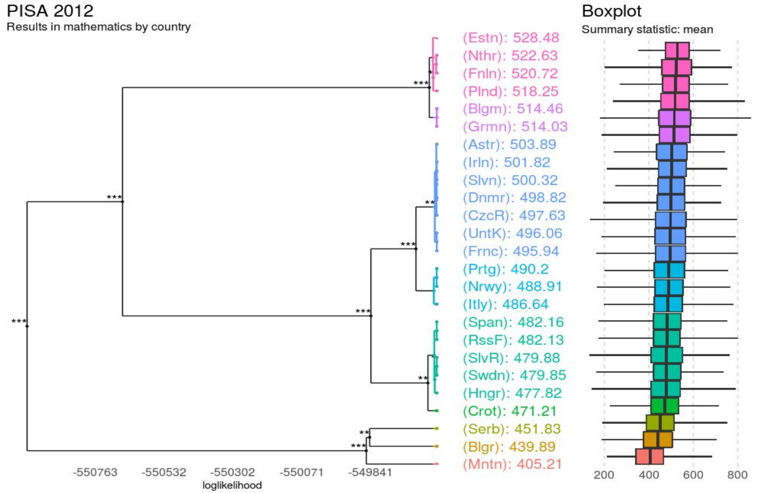
with 95% confidence intervals



PISA 2012

Results in mathematics by country

Groups frequencies

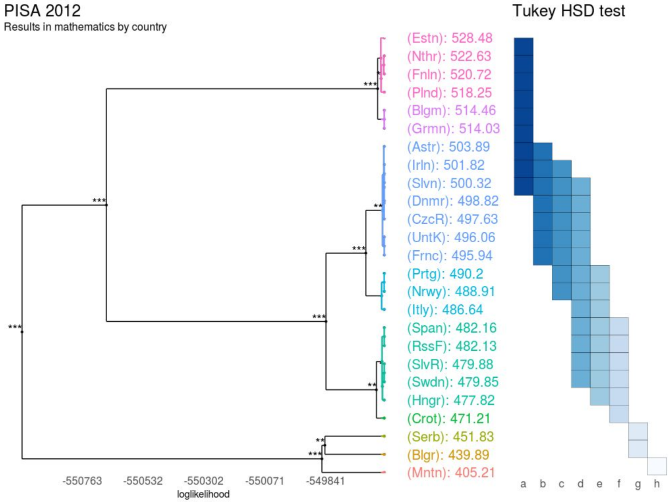


PISA 2012

Results in mathematics by country

Boxplot

Summary statistic: mean



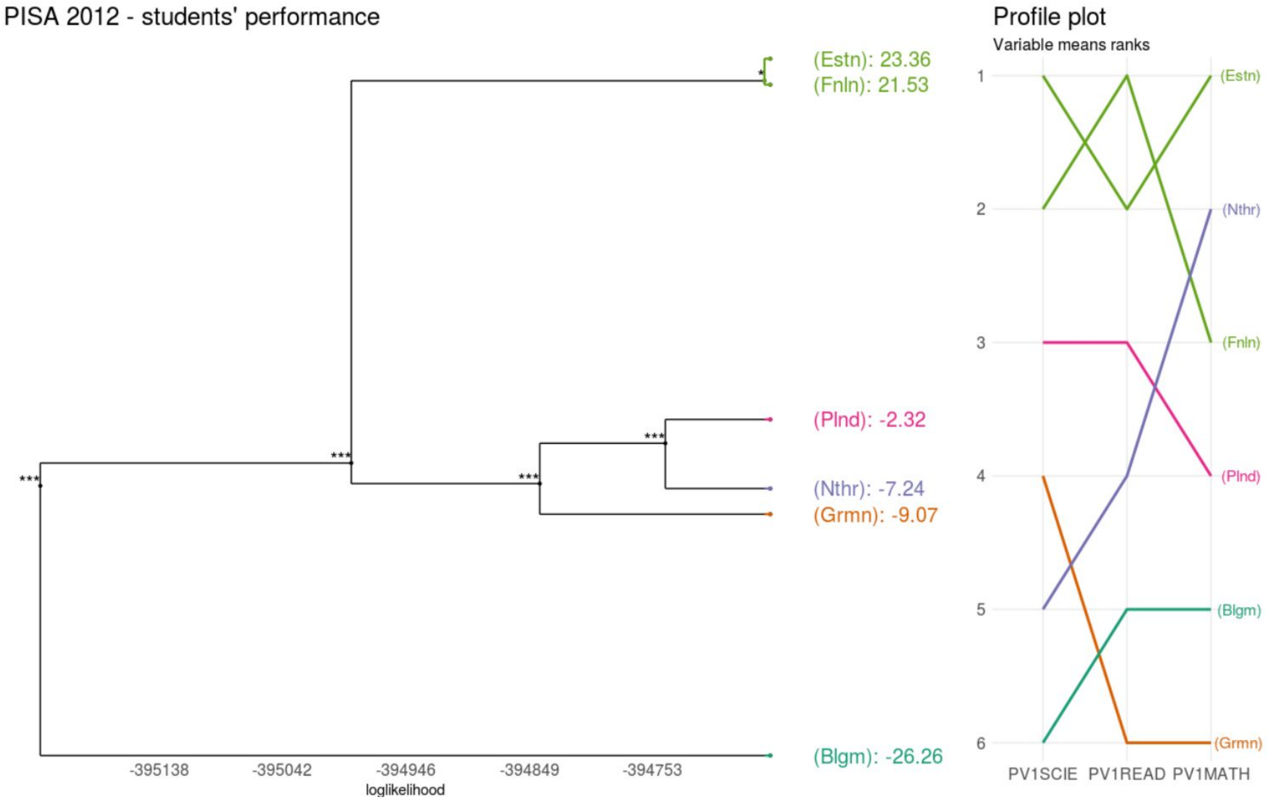
PISA 2012

Results in mathematics by country

Tukey HSD test

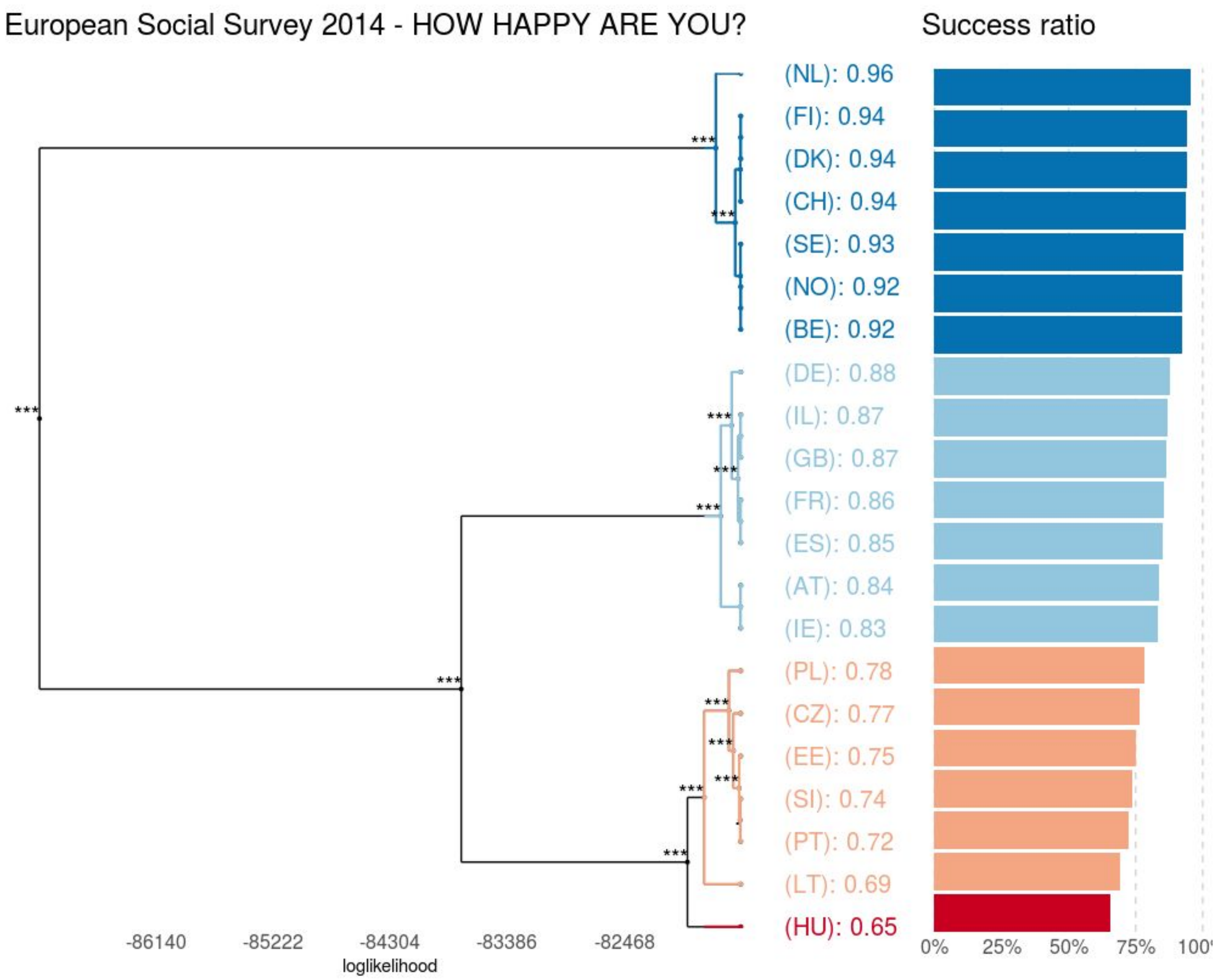
Other parametric models

1. multi dimensional Gaussian model,
2. binomial model,
3. survival model.



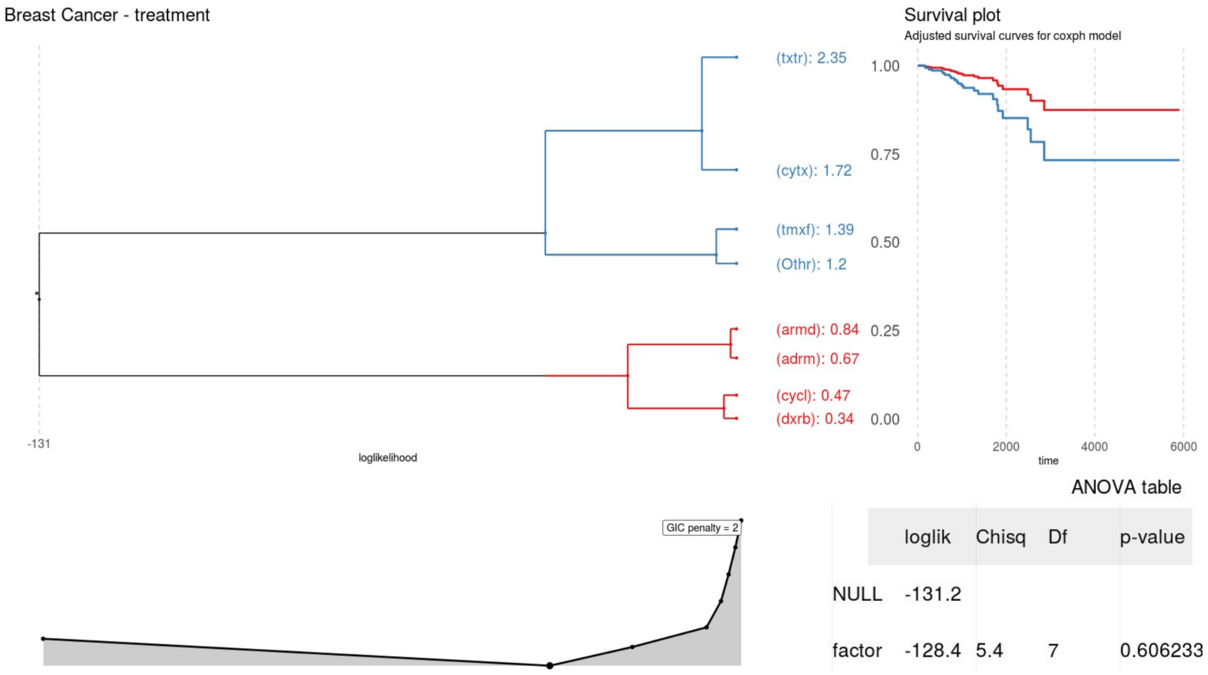
Other parametric models

1. multi dimensional Gaussian model,
2. binomial model,
3. survival model.



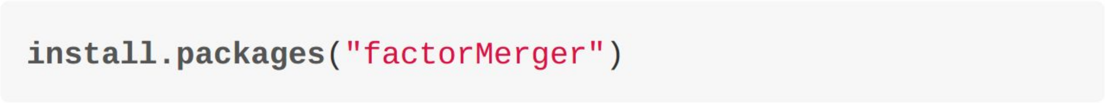
Other parametric models

1. multi dimensional Gaussian model,
2. binomial model,
3. survival model.



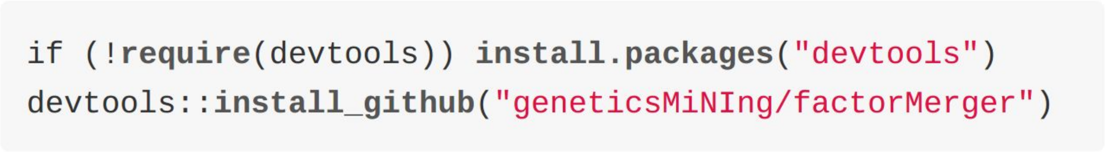
	loglik	Chisq	Df	p-value
NULL	-131.2			
factor	-128.4	5.4	7	0.606233

Install and use the package



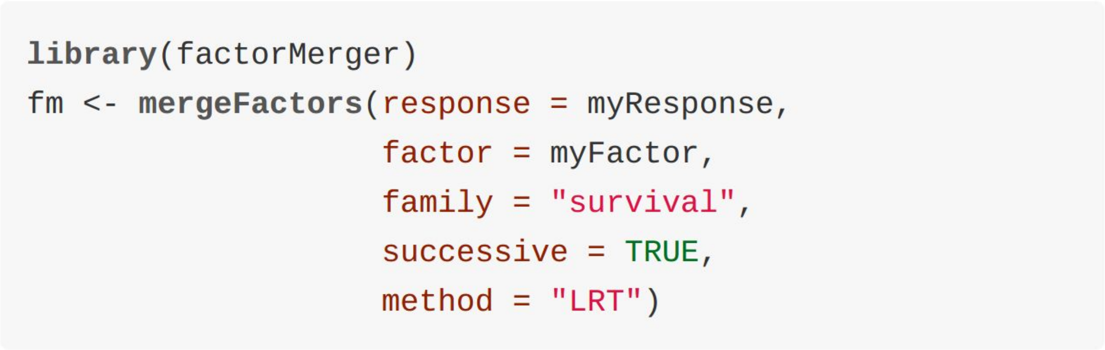
```r
install.packages("factorMerger")
```

CRAN



```r
if (!require(devtools)) install.packages("devtools")
devtools::install_github("geneticsMiNIng/factorMerger")
```

Github



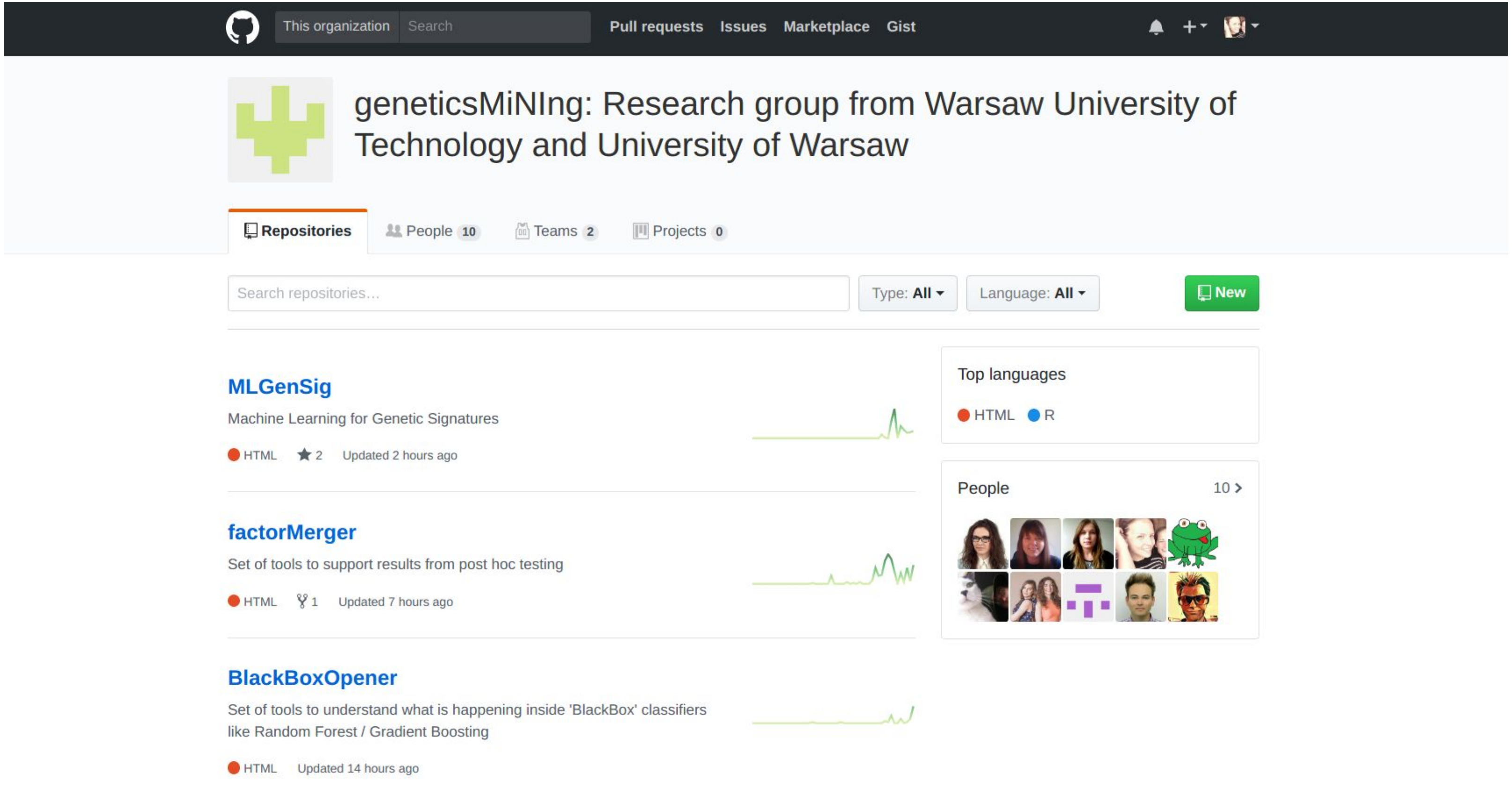
```r
library(factorMerger)
fm <- mergeFactors(response = myResponse,
                   factor = myFactor,
                   family = "survival",
                   successive = TRUE,
                   method = "LRT")
```

Find more: *https://github.com/geneticsMiNIng/factorMerger*

The aim of the factorMerger package

1. Create an algorithm which outputs an unequivocal data partition.
2. Improve visualizations.
3. Include other parametric models:
    a. multi dimensional Gaussian model,
    b. binomial model,
    c. survival model.

geneticsMiNIng



geneticsMiNIng: Research group from Warsaw University of Technology and University of Warsaw

MLGenSig

Machine Learning for Genetic Signatures

● HTML ★ 2 Updated 2 hours ago

factorMerger

Set of tools to support results from post hoc testing

● HTML 1 Updated 7 hours ago

BlackBoxOpener

Set of tools to understand what is happening inside 'BlackBox' classifiers like Random Forest / Gradient Boosting

● HTML Updated 14 hours ago

Top languages

● HTML ● R

People 10 >

Any questions?

Agnieszka Sitko

ag.agnieszka.sitko@gmail.com

06-07-2017