factorMerger: hierarchical clustering and model visualization

Agnieszka Sitko
Warsaw University of Technology, MI^2 Group
useR!2017 | 06-07-2017
Problem

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

Data taken from Programme for International Student Assessment (PISA 2012, OECD)
Problem

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

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

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

**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.

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

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

95% family-wise confidence level

TukeyHSD(pisaAOV)

<table>
<thead>
<tr>
<th></th>
<th>diff</th>
<th>lwr</th>
<th>upr</th>
<th>p adj</th>
</tr>
</thead>
<tbody>
<tr>
<td>Netherlands-Belgium</td>
<td>8.168042</td>
<td>1.827307</td>
<td>14.5087766</td>
<td>0.0071606</td>
</tr>
<tr>
<td>Poland-Belgium</td>
<td>3.793268</td>
<td>-1.962187</td>
<td>9.5487229</td>
<td>0.2700258</td>
</tr>
<tr>
<td>Poland-Netherlands</td>
<td>-4.374774</td>
<td>-9.166789</td>
<td>0.4172409</td>
<td>0.0819931</td>
</tr>
</tbody>
</table>

https://www.linkedin.com/pulse/anova-analysis-variance-kumar-p
Potentially \( (n/3) \) inconsistencies
95% family-wise confidence level

Potentially $(n^3)$ inconsistencies

Cluttered visualizations
Potentially inconsistent
Cluttered visualizations
Fixed significance level
Time for factorMerger
Install and use the package

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

devtools::install_github("geneticsMiNIng/factorMerger")

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

plot(fm)
```

Find more: [https://github.com/geneticsMiNIng/factorMerger](https://github.com/geneticsMiNIng/factorMerger)
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)
```
mergeFactors()

1. Likelihood Ratio Tests
2. Delete or Merge Regressors

Algorithm 1 Merging with LRT

function MERGEFACTORS(response, factor, successive)

2: pairsSet := generatePairs(response, factor, successive)
M₀ := full model

4: while levels(factor) > 1 do

4.1: toBeMerged := argmax_{pair \in pairsSet} l(updateModel(M₀, pair))

6: M₀ := updateModel(M₀, toBeMerged)
factor := mergeLevels(factor, pair)

8: pairsSet := pairsSet \ pair

end while

10: end function
mergeFactors()
1. Likelihood Ratio Tests
2. Delete or Merge Regressors

Algorithm 2 Merging with agglomerative clustering

function MERGEFACTORS(response, factor, successive)
2: pairsSet := generatePairs(response, factor, successive)
dist := set of distances
4: for all pair ∈ pairsSet do
5:    h := \{μ_{pair_1} = μ_{pair_2}\} \quad \triangleright \text{hypothesis under which pair is merged}
6:    dist[pair] = LRT(M_h|M_0)
end for
8: if successive then
9:    hClust(dist, method = "single")
10: else
11:    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

LRT for the:
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Model’s likelihood

Group means

A cluster

(Estn): 528.48
(Nthr): 522.63
(Flnh): 520.72
(Pnd): 518.25
(Bglm): 514.46
(Grmn): 514.03
(Astr): 503.89
(Iln): 501.82
(Slnv): 500.32
(Dnmr): 498.82
(CzcR): 497.63
(UntK): 496.06
(Frcn): 495.94
(Prtg): 490.2
(Nrwy): 488.91
(Itly): 486.64
(Span): 482.16
(RssF): 482.13
(SlvR): 479.88
(Swdn): 479.85
(Hngr): 477.82
(Crot): 471.21
(Serb): 451.83
(Blgr): 439.89
(Mntn): 405.21
PISA 2012
Results in mathematics by country

Models:
- constant
- full
- best

GIC penalty

GIC penalty = 12.5
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.
Narodowe Centrum Nauki
Any questions?

Agnieszka Sitko

ag.agnieszka.sitko@gmail.com

06-07-2017