factorMerger: A Set of Tools to Support Results From Post Hoc Testing
A moze jakis tytul bardziej zwiazany z wizualizacja?

Agnieszka Sitko*
Faculty of Mathematics, Informatics and Mechanics
University of Warsaw
and
Przemysław Biecek
Faculty of Mathematics, Informatics and Mechanics
University of Warsaw

July 14, 2017

Abstract

The text of your abstract. 200 or fewer words.

Keywords: *ANOVA*, hierarchical clustering, likelihood ratio test

*The authors gratefully acknowledge

1 Introduction

In this article we present the **factorMerger** package whose aim is to enrich results from *ANOVA* tests together with providing the variety of plots designed for deeper understanding analyzed models. The *ANOVA* method verifies the null hypothesis that a variable of interest y has the same distribution in all subpopulations. If this null hypothesis is rejected a more detailed analysis of differences among categorical variable levels might be needed. The traditional approach is to perform *pairwise post hoc tests* in order to verify which groups differ significantly.

One may find implementations of the traditional *post hoc tests* in many *R* packages. Package **agricolae** (de Mendiburu, 2016) offers a wide range of them. It gives one of the most popular *post hoc test*, Tukey HSD test (`HSD.test`), its less conservative version — Student-Newman-Keuls test (`SNK.test`) or Scheffe test (`scheffe.test`) which is robust to factor imbalance. These parametric tests are based on Student's t-distribution, thus, are reduced to Gaussian models only. In contrasts, **multcomp** package (Hothorn et al., 2008) can be used with generalized linear models (function `glht`) as it uses general linear hypothesis. Similarly to the **multcomp**, some implementations that accept `glm` objects are also given in **car** (`linearHypothesis`, Fox and Weisberg, 2011) and **lsmeans** (Lenth, 2016).

However, an undeniable disadvantage of single-step *post hoc tests* is the inconsistency of their results. For a fixed significance level, it is possible that mean in group A does not differ significantly from the one in group B, similarly with groups B and C. At the same time the difference between group A and C is detected. Then data partition is unequivocal and, as a consequence, impossible to put through.

The problem of clustering categorical variable into non-overlapping groups has already been present in the literature. First, J. Tukey proposed an iterative procedure of merging factor levels based on the studentized range distribution (Tukey, 1949). However, statistical test used in this approach made it limited to Gaussian models. *Collapse And Shrinkage in ANOVA* (*CAS-ANOVA*, Bondell and Reich, 2008) is an algorithm that extends categorical variable partitioning for generalized linear models in testing. It is based on the Tibshirani's *Fused LASSO* (Tibshirani et al., 2005) with the constraint taken on the pairwise differences

within a factor, which yields to their smoothing.

Delete or Merge Regressors algorithm (Prochenka, 2016, p. 37) is also adjusted to generalized linear models. It directly uses the hierarchical clustering to gain a hierarchical structure of a factor. At the beginning *DMR4glm* estimates models arising from the full model by pairwise merging or deleting factor levels. Each model is then compared with the reference model with a use of the *Likelihood Ratio Test*. Finally, the agglomerative clustering is performed taking LRT statistic as a distance — each step of the clustering produces a more generalized model with different factor structure. Experimental studies (Prochenka, 2016, p. 44–91) showed that the *Delete orMerge Regressors*'s performance is better than *CAS-ANOVA*'s when it comes to the model accuracy. *Delete or Merge Regressors*'s implementation may be found in the **DMR** package (Maj et al., 2013). The algorithm will be described in details in further sections.

In this article we will also present a more direct approach to the problem of hierarchical clustering, nazwa - wypadaloby miec jakas chwytliwa nazwe. Similarly to *DMR4glm*, nazwa procedure is motivated by the *Likelihod Ratio Test*. In each step it chooses a model with the highest *Likelihood Ratio Test* test p-value or, in other words, the highest likelihood. While this algorithm is more complex than *DMR4glm*, thanks to its dynamic adaptability, it maximizes the likelihood in the merging path[1]. What is more, it is easily expandable for non-parametric models (using permutation tests instead of *LRT*s).

Both *DMR4glm* and nazwa algorithms are implemented in the **factorMerger** package.

In addition to the comprehensive algorithm which tries uniting all feasible pairs of levels in a step, also a *successive version* is provided. In the *successive version* only levels which are relatively close can be merged (levels distance is dependent on the model chosen). While the basic approach (all vs. all comparisons) may result in a slightly better partition from the statistical point of view, proposed extension (all vs. subsequent comparisons) seems to be more graceful when it comes to the interpretation. Moreover, the former algorithm is more computationally expensive.

More detailed description of algorithms implemented in **factorMerger** is given in the section Methodology.

[1]Although it may be shown that the *DMR* algorithm is a consistent model selection method, its performance on smaller datasets is undefined. TODO....

Przydaloby sie tez cos napisac o wizualizacji i moze o problemach z post hocami

2 Methodology

Merging procedures implemented in the **factorMerger** package begin with the full model — with all levels of a given factor included — and iteratively merge one pair of levels until the factor is constant. Uniting two groups reduces by one the number of subsets defined by the factor. In a single iteration pairs *worth uniting* are considered and the one which optimizes an objective function is joined. Objective functions use likelihood-based statistics. We will specify them later on.

The **factorMerger** package gives the ability to perform analysis for the wide family of models and choose from the broad spectrum of merging approaches. Depending on the problem statement, some parts of the merging procedure may differ. The general sketch of the algorithm is described below.

Algorithm 1 The outline of the merging procedure

function MERGEFACTORS($response, factor, family, successive, method$)
  $pairsSet := generatePairs(response, factor, successive)$
  $\mathcal{M} := createModel(response, factor, family)$
  while $|levels(factor)| > 1$ **do**
    $toBeMerged := \text{argmax}_{\text{pair} \in \text{pairsSet}} objectiveFunction(pair, response, factor)$
    $\mathcal{M} := updateModel(\mathcal{M},\ toBeMerged)$
    $factor := mergeLevels(factor, pair)$
    $pairsSet := joinPair(pairsSet, pair)$
  end while
end function

In the article we will denote:

- the sample size as n,
- the initial number of factor levels as k,

- the binary matrix representing group membership as $X = \{x_{ij}\}_{i,j=1}^{n,k}$.

$$x_{ij} = \begin{cases} 1 & \text{if } i-\text{th observation belongs to group } j, \\ 0 & \text{otherwise.} \end{cases}$$

  We assume that groups do not overlap.

- the response vector as $y = (y_1, \ldots, y_n)$ or the response matrix as $Y = \{y_{ij}\}_{i,j=1}^{n,m}$,

- the effects — vector of a length k or $k \times m$ matrix — as β.

2.1 Model family

In the current version the package supports parametric models:

- single dimensional Gaussian (with the argument `family = "gaussian"`),

- multi dimensional Gaussian — Gaussian model with multiple outputs y (with the argument `family = "gaussian"`)[2],

- binomial (with the argument `family = "binomial"`),

- survival (with the argument `family = "survival"`).

Each case has its own method of estimating model parameters and a specific likelihood formula.

Single dimensional Gaussian model A convenient and commonly made assumption in the analysis of variance is the normality of the errors. Here we will consider a linear model in which beta coefficients represent group means. The model may be written in vector form as

$$y = X\beta + \epsilon, \quad \epsilon \sim \mathcal{N}\left(0, \sigma^2\right).$$

[2]Both single dimensional and multi dimensional Gaussian models use `family = "gaussian"`. However, multi dimensional model uses different functions for likelihood estimation and may require additional preprocessing, thus, it is considered as a separate category.

Under the above assumptions we may formulate the likelihood of the Gaussian linear model (Friedman et al., 2001, p.31)

$$L\left(\beta,\sigma|y\right)=\left(2\pi\sigma^{2}\right)^{-\frac{n}{2}}\exp\left(-\frac{1}{2}\left(y-X\beta\right)^{T}\left(y-X\beta\right)/\sigma^{2}\right)$$

and the corresponding logarithm of the likelihood

$$l\left(\beta,\sigma|y\right)=-\frac{n}{2}\log\left(2\pi\right)-\frac{n}{2}\log\left(\sigma^{2}\right)-\frac{1}{2}\left(y-X\beta\right)^{T}\left(y-X\beta\right)/\sigma^{2}.$$

To calculate the loglikelihood in the package we use `logLik.lm{stats}`.

Multi dimensional Gaussian model In the multi dimensional generalization of the Gaussian model we will observe multiple outputs Y. It takes the following form

$$Y=X\beta+E,\quad E\sim\mathcal{N}(0,\Sigma),$$

where $\beta=\{\beta\}_{i,j=1}^{k,m}$ is an $k\times m$ effects matrix and $E=(\epsilon_1,\epsilon_2,...,\epsilon_m)$ is an m-dimensional error.

Now, we may write the likelihood

$$L\left(\beta,\Sigma|Y\right)=\left(|2\pi\Sigma|\right)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\left(Y-X\beta\right)^{T}\Sigma^{-1}\left(Y-X\beta\right)\right)$$

and its logarithm

$$l\left(\beta,\sigma|Y\right)=-\frac{n}{2}\log\left(2\pi\right)-\frac{n}{2}\log\left(|\Sigma|\right)-\frac{1}{2}\left(Y-X\beta\right)^{T}\Sigma^{-1}\left(Y-X\beta\right).$$

Unfortunately, **stats** or any commonly used R package do not support multiple responses in the loglikelihood calculation for linear Gaussian models. In the package we use `logLik.mlm` implementation introduced in the **Atools** package (B¡b6¿ck, 2014) and the `dmvnorm{mvtnorm}` (Genz and Bretz, 2009) implementation for multivariate normal density estimation.

Binomial model In the binomial case the observed response y will take one of two possible disjoint outcomes (success or failure). We will be interested whether probabilities of

success in given subpopulations are statistically the same. Let us introduce the assumption

$$y \sim \mathcal{B}(p, n)$$

where $\mathcal{B}(p, n)$ is the binomial distribution with the probability of success p and the number of trials n. We consider the logit model

$$\ln\left(\frac{p}{1-p}\right) = X\beta.$$

Let $z = \sum_{i=1}^{n} y_i$. Now the likelihood may be written as follows (Czepiel, 2002)

$$L(\beta|y) = \frac{n!}{z!\,(n-z)!} p^z (1-p)^{n-z}.$$

Thus, the logarithm of the likelihood is expressed as

$$l(\beta|y) = zX\beta - n\log\left(1 + \exp^{X\beta}\right).$$

TODO: Policzy? loglik, czy na pewno dobrze.

To calculate the loglikelihood for the binomial model `logLik.glm{stats}` is used.

Survival model Survival analysis is a branch of statistics for analyzing the time until a specified event takes to happen (such as a patient death or a machine failure). Usually we assume that survival time of an individual depends on two factors: the underlying baseline hazard function and a set of explanatory variables. In the package for survival analysis applications the *Cox proportional hazard regression* introduced by D. R. Cox (Cox, 1992) is used. In this model we assume that the baseline hazard function is the same for the whole population and the effects of predictors are multiplicatively related to the individual hazard.

Let $\lambda(t)$ be the hazard function and let $\lambda_0(t)$ be the baseline hazard function, where t denotes time. Then the Cox model has the following form

$$\lambda(t) = \lambda_0(t) \cdot \exp(X\beta).$$

Here $y = (y_1, \ldots, y_n)$ is a vector of observed times in the sample. For i-th observation y_i can be either event time or censoring time. We say that an observation is censored if

we do not have complete information about its status in the context of our interest (for example, the patient died of causes other than the disease of our consideration or was lost to follow-up). Let $C = (C_1, \ldots, C_n)$ be a vector of indicators that the time corresponds to the event. To be more presice,

$$C_i = \begin{cases} 1 & \text{if for } i-\text{th observation the event occured,} \\ 0 & \text{if } i-\text{th observation was censored.} \end{cases}$$

Then we may construct the partial likelihood (as a function of β only) in the following way

$$L(\beta|y) = \prod_{i:C_i=1} \left[\frac{\exp(X_i\beta)}{\sum_{j:y_j \geq y_i} \exp(X_i\beta)} \right]$$

and the corresponding loglikelihood

$$l(\beta|\, y) = \sum_{i:C_i=1} \left(X_i\beta - \log \left(\sum_{j:y_j \geq y_i} \exp(X_i\beta) \right) \right).$$

The Cox regression is implemented in the **survival** package (`coxph{survival}`, Terry M. Therneau and Patricia M. Grambsch, 2000) and the loglikelihood of the model one may find by accessing the field `loglik` of a `coxph.object`.

The Likelihood Ratio Test statistics The substantial part of **factorMerger**'s algorithms is calculating the *Likelihood Ratio Test statistics*. In this paragraph we will formulate its definition.

Let us assume that a factor C divides population into l subgroups. Let us also denote the effect of a group i on the response as β_i. We define h_{ij}, a constraint on groups i and j claiming that their group effects are statistically the same, as

$$h_{ij} : \beta_i = \beta_j, \;\; i \neq j \;\; i, j \in \{1, 2, ..., l\}. \tag{1}$$

Let us take $\mathcal{M}_0$ — the model with no constraints and $\mathcal{M}_{h_{ij}}$ — the model under h_{ij}. For both models estimate an estimator of group effects and denote them as $\hat{\beta}_{\mathcal{M}_0}, \hat{\beta}_{\mathcal{M}_{h_{ij}}}$, respectively.

Then, the **Likelihood Ratio Test statistic** is

$$LRT(\mathcal{M}_{h_{ij}}|\mathcal{M}_0) = 2 \cdot l(\hat{\beta}_{\mathcal{M}_0}|y) - 2 \cdot l(\hat{\beta}_{\mathcal{M}_{h_{ij}}}|y), \tag{2}$$

where $l(\cdot|y)$ is the log-likelihood function conditioned by observed y.

The higher the $LRT(\mathcal{M}_{h_{ij}}|\mathcal{M}_0)$, the more likely it is that the constraint h_{ij} is rejected. One may interpret the $LRT(\mathcal{M}_{h_{ij}}|\mathcal{M}_0)$ as a distance between group i and j.

As for all i, j $(i \neq j \;\; i, j \in \{1, 2, ..., l\})\mathcal{M}_{h_{ij}}$ are nested in $\mathcal{M}_0$, the likelihood of $\mathcal{M}_{h_{ij}}$ for fixed i, j is not greater than the $\mathcal{M}_0$'s likelihood. If $\mathcal{H}$ is a set of all considered constraints defined in (1), a constraint

$$\text{argmin}_{h \in \mathcal{H}} LRT(\mathcal{M}_h|\mathcal{M}_0) = \text{argmax}_{h \in \mathcal{H}} l(\mathcal{M}_h) \tag{3}$$

will reduce the likelihood the least and, therefore, minimizing the LRT distance between subgroups is equivalent to maximizing the likelihood.

Asymptotic behaviour of the *LRT statistic* An advantageous result by Samuel S. Wilks (Wilks, 1938) shows that for a linear constraint h the $LRT(\mathcal{M}_h|\mathcal{M}_0)$ tends asymptotically to chi-squared distribution with degrees of freedom equal to the difference in degrees of freedom between $\mathcal{M}_0$ and $\mathcal{M}_h$ as number of observations approaches infinity. This convergence will be used to evaluate model's *statistical correctness* in visualizations.

2.2 Pairs considered

In the Algorithm 1 to achieve maximal values of the likelihood in the merging path all feasible pairs should be considered while performing a single step. However, computing an objective function can be expensive and, especially for big datasets, it may be beneficial to limit the set of tested hypotheses. Let us also remark that it is more likely that a pair of levels i and j will be chosen to merge if corresponding effects estimators $(\hat{\beta}_i, \hat{\beta}_j)$ are close.

TODO: Czy to potrzebuje rozwini?cia?

Therefore, in the package we implement two strategies of merging — either comprehensive or limited. They are called as follows:

- *all-to-all* (with the argument `successive = FALSE`),

- *successive* (with the argument `successive = TRUE`).

The version *all-to-all* considers all possible pairs of factor levels. In the *successive* approach factor levels are preliminarily sorted and then only consecutive groups can be united.

It is possible that the *all-to-all* strategy will give a better merging path, though, intuitively, the difference should not be significant.

TODO: Czy ja tak moge pisac?

The *successive* merging In the *successive* version of the algorithm at the very beginning levels of a categorical variable are sorted. The order depends on the model family chosen. In most cases it is associated with the beta coefficients estimators. The detailed rules of ordering levels are given below.

model	**metric**
one-dimensional Gaussian	average
multi-dimensional Gaussian	average of the isoMDS transformation
binomial	proportion of successes
survival	log hazard ratio

Table 1: Factor ordering by model family

For single dimensional Gaussian and binomial models groups are sorted by means and proportions of success, respectively. In the survival case we use log hazard ratios with one level set as the reference level. Multi dimensional Gaussian model needs additional preprocessing. First, group means are computed. Then they are projected into one dimensional space with the use of the Kruskal's non-metric multidimensional scaling. The **factorMerger** uses `isoMDS` implementation from the package **MASS** (Venables and Ripley, 2002).

Having set the factor order, we may decrease number of comparisons in each step in a way that a particular level is tested only against its closest neighbours.

2.3 Objective functions

The **factorMerger** package for each model family and merging strategy implements two types of a single iteration of the algorithm. They use one of the following:

- *Likelihood Ratio Test* (with the argument `method = "LRT"`),
- *agglomerative clustering with constant distance matrix* (based on the *DMR4glm* algorithm, with the argument `method = "hclust"`).

Ujednolicic algorytmy tak, zeby byly spojne z Alg 1

2.4 The *Likelihood Ratio Test*-based merging

The *Likelihood Ratio Test*-based approach minimizes likelihood reduction in the merging path. It may be summarized as follows.

TODO: Rozwin??... (Analogia do LRT testw, ale mo?na upro?ci? do samego loglik)

Algorithm 2 Merging with the *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

2.5 The *DMR4glm*-based merging

TODO: Wst?pny opis

Algorithm 3 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

We acknowledge the financial support from the *NCN Opus grant 2016/21/B/ST6/02176*.

Problem z kodowaniem bibliografii.

References

A. B¡b6¿ck. *Atools: Atools*, 2014. URL `https://R-Forge.R-project.org/projects/biostat/`. R package version 0.2/r191.

H. D. Bondell and B. J. Reich. Simultaneous factor selection and collapsing levels in ANOVA. *Department of Statistics, North Carolina State University*, 2008.

D. R. Cox. Regression models and life-tables. In *Breakthroughs in statistics*, pages 527–541. Springer, 1992.

S. A. Czepiel. Maximum likelihood estimation of logistic regression models: theory and implementation. *Available at czep. net/stat/mlelr. pdf*, 2002.

F. de Mendiburu. *agricolae: Statistical Procedures for Agricultural Research*, 2016. URL `https://CRAN.R-project.org/package=agricolae`. R package version 1.2-4.

J. Fox and S. Weisberg. *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, second edition, 2011. URL `http://socserv.socsci.mcmaster.ca/jfox/Books/Companion`.

J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

A. Genz and F. Bretz. *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Springer-Verlag, Heidelberg, 2009. ISBN 978-3-642-01688-2.

T. Hothorn, F. Bretz, and P. Westfall. Simultaneous inference in general parametric models. *Biometrical Journal*, 50(3):346–363, 2008.

R. V. Lenth. Least-squares means: The R package lsmeans. *Journal of Statistical Software*, 69(1):1–33, 2016. doi: 10.18637/jss.v069.i01.

A. Maj, A. Prochenka, and P. Pokarowski. *DMR: Delete or Merge Regressors for linear model selection.*, 2013. URL `https://CRAN.R-project.org/package=DMR`. R package version 2.0.

A. Prochenka. *Delete or Merge Regressors algorithm*, chapter 4.3, 5–6, pages 37, 44–91. 2016.

Terry M. Therneau and Patricia M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, New York, 2000. ISBN 0-387-98784-3.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society*, pages 01–108, 2005.

J. Tukey. Comparing Individual Means in the Analysis of Variance. *BIOMETRICS*, pages 99–114, 1949.

W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. URL `http://www.stats.ox.ac.uk/pub/MASS4`. ISBN 0-387-95457-0.

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. ISBN 978-0-387-98140-6. URL `http://ggplot2.org`.

S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.