

Package ‘exactmaxsel’

May 29, 2009

Version 1.0-4

Date 2009-05-29

Title Maximally selected statistics for binary response variables - Exact methods

Author Anne-Laure Boulesteix <boulesteix@imse.med.tu-muenchen.de>

Maintainer Anne-Laure Boulesteix <boulesteix@imse.med.tu-muenchen.de>

Depends R (>= 2.0.0), combinat

Suggests

Description This package computes the exact distribution of some maximally selected statistics in the following setting: the ‘response’ variable is binary, the splitting variable may be nominal, ordinal or continuous. Currently, the package implements the chi-square statistic and the Gini-index.

License GPL (>= 2)

URL <http://cran.r-project.org/web/packages/exactmaxsel/index.html>

Repository CRAN

Date/Publication 2009-05-29 13:19:48

R topics documented:

birth	2
boundary	3
Fcat	5
Ford	6
Ford2	7
ginigain	8
maxsel	10
maxsel.test	11

Index	14
--------------	-----------

 birth

Birth data set by Boulesteix (2006)

Description

A data set containing 25 qualitative and quantitative variables for n=501 births in 2003 and 2004.

Usage

```
data(birth)
```

Details

The missing values are of two types: 1. Some of the variables do not concern all the births. For example, the variable `Episiotomy` is not relevant for cesarean births. The type `NA` is used for these entries. Such observations should be removed when analysing these variables. It does not make sense to impute these missing values. 2. Some questionnaires were returned incomplete, because the mother forgot some questions or did not know the answer (especially for the head circumference).

Value

A data frame with following variables

<code>IndexMother</code>	Index of the mother (unordered)
<code>Sex</code>	Sex of the baby (1:male,2:female)
<code>Weight</code>	Weight of the baby in g.
<code>Height</code>	Height of the baby in cm.
<code>Head</code>	Head circumference of the baby in cm.
<code>Month</code>	Month of birth (coded as 1:january,...,12:december).
<code>Year</code>	Year of birth.
<code>Country</code>	Country of birth (FR:France,BE:Belgium,CH:Switzerland,CA:Canada, etc).
<code>Term</code>	Duration of the pregnancy in week.
<code>AgeMother</code>	Age of the mother in year.
<code>Previous</code>	Number of previous deliveries.
<code>WeightBefore</code>	Weight of the mother before pregnancy in kg.
<code>HeightMother</code>	Height of the mother in cm.
<code>WeightEnd</code>	Weight of the mother at the end of the pregnancy.
<code>Twins</code>	Twin pregnancy? (0:no,1:yes)
<code>Intensive</code>	Number of days spent by the baby in the neonatology intensive care unit.
<code>Cesarean</code>	Cesarean birth? (0:no,1:yes)
<code>Planned</code>	If cesarean birth, was the cesarean planned before the beginning of labor? (0:no,1:yes)

Episiotomy	If vaginal birth, did the mother have an episiotomy? (0:no,1:yes)
Tear	If vaginal birth, did the mother tear? (0:no,1:yes)
Operative	If vaginal birth, was it a vaginal operative birth (i.e. with forceps,vacuum extractor,etc)? (0:no,1:yes)
Induced	If vaginal delivery or emergency cesarean, was the labor induced medically? (0:no,1:yes)
Membranes	Did the membranes rupture before the beginning of labor? (0:no,1:yes)
Rest	Was the mother prescribed bed rest for one month or more? (0:no,1:yes)
Presentation	Presentation of the baby at birth (1:cephalic,2:breech,3:other,e.g. transverse).

Source

This data set was collected from internet users via french-speaking pregnancy forums in August 2004 by Anne-Laure Boulesteix.

References

A.-L. Boulesteix (2006), Maximally selected chi-square statistics for ordinal variables, *Biometrical Journal* 48:451-462.

Examples

```
# load exactmaxsel library
library(exactmaxsel)

# load data set
data(birth)

# Display the two first observations
birth[1:2,]
```

boundary

Computes the coordinates of the boundaries

Description

The function `boundary` is an internal function that computes the greatest (and also the smallest if `lower=TRUE`) number of observations from class $Y=1$ in the left node that lead to an association criterion $\leq c$. Remark: in general, the numbers output by `boundary` are not integers.

Usage

```
boundary(x, n0, n1, c, statistic, lower=TRUE)
```

Arguments

<code>x</code>	the number of observations in the left node
<code>n0</code>	the number of observations in class $Y=0$
<code>n1</code>	the number of observations in class $Y=1$
<code>c</code>	the value of the criterion that should not be exceeded.
<code>statistic</code>	the association measure. Currently, only <code>statistic="chi2"</code> (chi-square statistic) and <code>statistic="gini"</code> (the Gini-gain from machine learning) are implemented.
<code>lower</code>	Should the lower boundary also be computed?

Details

This function should not be called by the user in practice.

Value

a list with

<code>upper</code>	the upper boundary (greatest allowed value).
<code>lower</code>	the lower boundary (smallest allowed value).

Author(s)

Anne-Laure Boulesteix (http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/020_professuren/boulesteix/).

References

- A.-L. Boulesteix (2006), Maximally selected chi-square statistics for ordinal variables, *Biometrical Journal* 48:451-462.
- A.-L. Boulesteix (2006), Maximally selected chi-square statistics and binary splits of nominal variables, *Biometrical Journal* 48:838-848.
- C. Strobl, A.-L. Boulesteix and T. Augustin (2007), Unbiased split selection for classification trees based on the Gini index, *Computational Statistics and Data Analysis* 52:483-501.

See Also

[Ford](#), [Fcat](#).

Examples

```
# load exactmaxsel library
library(exactmaxsel)

boundary(10, 30, 30, c=3, statistic="chi2", lower=TRUE)
```

Fcat	<i>Distribution of maximally selected statistics for multicategorical variables</i>
------	---

Description

The function `Fcat` computes the distribution of the maximally selected association criterion of interest (either the chi-square statistic or the Gini-gain in the current version) when Y is binary and X has unordered categorical values, given n_0 , n_1 and A .

Usage

```
Fcat(c, n0, n1, A, statistic)
```

Arguments

<code>c</code>	the value at which the distribution function has to be computed.
<code>n0</code>	the number of observations in class $Y=0$.
<code>n1</code>	the number of observations in class $Y=1$.
<code>A</code>	a vector of length K giving the number of observations with $X=1, \dots, X=K$.
<code>statistic</code>	the association measure used as criterion to select the best split. Currently, only <code>statistic="chi2"</code> (chi-square statistic) and <code>statistic="gini"</code> (the Gini-gain from machine learning) are implemented.

Details

Suppose the response Y is binary ($Y=0,1$) and the predictor X has K unordered categorical values ($X=1, \dots, K$). The criterion is maximized over all the binary splittings of the set $\{1, \dots, K\}$. For example, if $K=4$, the criterion is thus maximized over the splittings $\{1\}\{2,3,4\}$, $\{1,2\}\{3,4\}$, $\{1,2,3\}\{4\}$, $\{1,2,4\}\{3\}$, $\{1,4\}\{2,3\}$, $\{1,3,4\}\{2\}$, $\{1,3\}\{2,4\}$.

Value

the value of the distribution function at c .

Author(s)

Anne-Laure Boulesteix (http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/020_professuren/boulesteix/)

References

A.-L. Boulesteix (2006), Maximally selected chi-square statistics and binary splits of nominal variables, *Biometrical Journal* 48:838-848.

See Also

[Ford](#), [Ford2](#), [maxsel](#).

Examples

```
# load exactmaxsel library
library(exactmaxsel)

Fcat(c=4,n0=15,n1=10,A=c(6,10,9),statistic="chi2")
Fcat(c=5,n0=15,n1=15,A=c(5,8,7,10),statistic="gini")
```

Ford

Distribution of maximally selected statistics for (at least) ordinally scaled variables

Description

The function `Ford` computes the distribution of the maximally selected association criterion of interest (either the chi-square statistic or the Gini-gain in the current version) when Y is binary and X has ordered values, given n_0 , n_1 and A . Note that X must be AT LEAST ordinally scaled, i.e. continuous variables are also allowed as an extreme special case.

Usage

```
Ford(c, n0, n1, A, statistic)
```

Arguments

<code>c</code>	the value at which the distribution function has to be computed.
<code>n0</code>	the number of observations in class $Y=0$.
<code>n1</code>	the number of observations in class $Y=1$.
<code>A</code>	a vector of length K giving the number of observations with $X=1, \dots, X=K$. In the special case of a continuous X variable taking distinct values in the available sample, A takes the form $A = \text{rep}(1, N)$, where $N = n_0 + n_1$.
<code>statistic</code>	the association measure used as criterion to select the best split. Currently, only <code>statistic="chi2"</code> (chi-square statistic) and <code>statistic="gini"</code> (the Gini-gain from machine learning) are implemented.

Details

Suppose the response Y is binary ($Y=0,1$) and the predictor X has K ordered categorical values ($X=1, \dots, K$). The criterion is maximized over all the binary splittings of the set $\{1, \dots, K\}$ that preserve the ordering. For $K=3$, the criterion is thus maximized over the splittings $\{1,2\}\{3\}$ and $\{1\}\{2,3\}$. Note that X may also be a substantially continuous variable that is observed at a discrete scale and thus has ties.

Value

the value of the distribution function at c .

Author(s)

Anne-Laure Boulesteix (http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/020_professuren/boulesteix/)

References

A.-L. Boulesteix (2006), Maximally selected chi-square statistics for ordinal variables, *Biometrical Journal* 48:451-462.

See Also

`Fcat`, `Ford2`, `maxsel`.

Examples

```
# load exactmaxsel library
library(exactmaxsel)

Ford(c=4, n0=15, n1=10, A=c(6, 10, 9), statistic="chi2")
Ford(c=0.02, n0=15, n1=15, A=c(5, 8, 7, 10), statistic="gini")
```

Ford2

Distribution of maximally selected statistics for (at least) ordinally scaled variables in the two-cutpoint context

Description

The function `Ford2` computes the distribution of the maximally selected association criterion of interest (either the chi-square statistic or the Gini-gain in the current version) when Y is binary and X has ordered values, given n_0 , n_1 and A , in the case of a non-monotonic association represented by two cutpoints.

Usage

```
Ford2(c, n0, n1, A, statistic)
```

Arguments

c the value at which the distribution function has to be computed.
 n_0 the number of observations in class $Y=0$.
 n_1 the number of observations in class $Y=1$.
 A a vector of length K giving the number of observations with $X=1, \dots, X=K$.

`statistic` the association measure used as criterion to select the best split. Currently, only `statistic="chi2"` (chi-square statistic) and `statistic="gini"` (the Gini-gain from machine learning) are implemented.

Details

Suppose the response Y is binary ($Y=0,1$) and the predictor X has K ordered categorical values ($X=1,\dots,K$). The criterion is maximized over all the binary splittings of the set $\{1,\dots,K\}$ that are obtained from at most two cutpoints. For example, with $K=4$, the criterion is maximized over the splittings $\{1,2,3\}\{4\}$, $\{1,2\}\{3,4\}$, $\{1\}\{2,3,4\}$, $\{1,2,4\}\{3\}$, $\{1,4\}\{2,3\}$ and $\{1,3,4\}\{2\}$.

Value

the value of the distribution function at c .

Author(s)

Anne-Laure Boulesteix (http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/020_professuren/boulesteix/)

References

A.-L. Boulesteix and C. Strobl (2006), Maximally selected chi-square statistics and non-monotonic associations: an exact approach based on two cutpoints. *Computational Statistics and Data Analysis* 51:6295-6306.

See Also

[Ford](#), [Fcat](#), [maxsel](#).

Examples

```
# load exactmaxsel library
library(exactmaxsel)

Ford2(c=4, n0=15, n1=15, A=c(6, 10, 9, 5), statistic="chi2")
Ford2(c=0.02, n0=15, n1=15, A=c(5, 8, 7, 10), statistic="gini")
```

ginigain

Computation of the Gini gain from a 2x2 contingency table

Description

The function `ginigain` computes the Gini gain (also denoted as impurity reduction) resulting by splitting into the left and the right nodes whose counts are given in the contingency table `mat`. See Strobl et al. (2006) for a more precise definition.

Usage

```
ginigain(mat)
```

Arguments

mat a 2x2 matrix corresponding to a two-dimensional contingency table. The first row and the second row correspond to Y=0 and Y=1, respectively. The first column and the second column correspond to the left and right nodes, respectively.

Details

Note that, in contrast to the chi-square statistic, the Gini gain does not treat X and Y symmetrically.

Value

the (positive) value of the Gini gain.

Author(s)

Anne-Laure Boulesteix (http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/020_professuren/boulesteix/)

References

L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone (1984), Classification and Regression Trees, Wadsworth, Monterey, CA.

C. Strobl, A.-L. Boulesteix and T. Augustin (2006), Unbiased split selection for classification trees based on the Gini index, Computational Statistics and Data Analysis 52:483-501.

See Also

[maxsel.test](#).

Examples

```
# load exactmaxsel library
library(exactmaxsel)

# Define matrix
A<-matrix(c(15,20,22,12),2,2)

# Compute Gini gain
ginigain(A)
```

maxsel

Maximally selected criterion

Description

The function `maxsel` computes the maximal value of the criterion of interest (either the Gini-gain or the chi-square statistic) over some candidate binary splits. The candidate binary splits depend on `type` (see details).

Usage

```
maxsel(x, y=NULL, type, statistic)
```

Arguments

<code>x</code>	a numeric vector of length <code>n</code> giving the values of the variable <code>X</code> for the considered <code>n</code> observations. The classes must be coded as <code>1,...,K</code> . Alternatively, <code>x</code> can be a <code>2 x K</code> matrix corresponding to a contingency table, where the two rows are for the values of <code>Y</code> (<code>Y=0,1</code>) and the <code>K</code> columns are for the values of <code>X</code> (<code>X=1,...,K</code>). In this case, <code>y</code> must be set to <code>y=NULL</code> .
<code>y</code>	a numeric vector of length <code>n</code> giving the class (response variable <code>Y</code>) of the considered observations. The classes must be coded as <code>0</code> and <code>1</code> . If <code>x</code> is a contingency table, <code>y</code> must be set to <code>y=NULL</code> .
<code>type</code>	the type of the considered binary splits. <code>type="ord"</code> corresponds to an ordinal <code>X</code> variable, <code>type="cat"</code> corresponds to a categorical <code>X</code> variable with unordered categories, <code>type="ord2"</code> corresponds to an ordinal <code>X</code> variable with 2 cutpoints (non-monotonous association).
<code>statistic</code>	the association measure used as criterion to select the best split. Currently, only <code>statistic="chi2"</code> (chi-square statistic) and <code>statistic="gini"</code> (the Gini-gain from machine learning) are implemented.

Details

For example, let us consider a variable `X` with the possible values `{1,2,3,4}`. If `type="ord"`, the set of candidate splits consists of `{1}{2,3,4}`, `{1,2}{3,4}` and `{1}{2,3,4}`. If `type="cat"`, the set of candidate splits consists of `{1}{2,3,4}`, `{1,2}{3,4}`, `{1,2,3}{4}`, `{1,2,4}{3}`, `{1,4}{2,3}`, `{1,3,4}{2}`, `{1,3}{2,4}`. If `type="ord2"`, the set of candidate splits consists of `{1}{2,3,4}`, `{1,2}{3,4}`, `{1,2,3}{4}`, `{1,2,4}{3}`, `{1,4}{2,3}`, `{1,3,4}{2}`.

Value

the value of the maximally selected criterion.

Author(s)

Anne-Laure Boulesteix (http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/020_professuren/boulesteix/)

References

- A.-L. Boulesteix (2006), Maximally selected chi-square statistics for ordinal variables, *Biometrical Journal* 48:451-462.
- A.-L. Boulesteix (2006), Maximally selected chi-square statistics and binary splits of nominal variables, *Biometrical Journal* 48:838-848.
- C. Strobl, A.-L. Boulesteix and T. Augustin (2007), Unbiased split selection for classification trees based on the Gini index, *Computational Statistics and Data Analysis* 52:483-501.
- A.-L. Boulesteix and C. Strobl (2006), Maximally selected chi-square statistics and non-monotonic associations: an exact approach based on two cutpoints. *Computational Statistics and Data Analysis* 51:6295-6306.

See Also

[maxsel.test](#).

Examples

```
# load exactmaxsel library
library(exactmaxsel)

# First case: x and y are data vectors
# Simulate x and y
x<-sample(4,30,replace=TRUE)
y<-sample(c(0,1),30,replace=TRUE)

maxsel.test(x=x,y=y,type="ord",statistic="chi2")
maxsel.test(x=x,y=y,type="cat",statistic="gini")

# Second case: x is a contingency table, y=NULL.
x<-matrix(c(8,10,40,13,15,4),2,4,byrow=TRUE)
maxsel.test(x=x,y=NULL,type="ord",statistic="chi2")
maxsel.test(x=x,y=NULL,type="cat",statistic="gini")
```

maxsel.test

Test of independence based on maximally selected statistics

Description

The function `maxsel.test` computes the probability that the maximally selected criterion is \leq than the value observed from the data, under the null-hypothesis of no association between X and Y, given the numbers of observations with $Y=0, Y=1, X=1, \dots, X=K$. The candidate binary splits over which the criterion is maximized depend on `type` (see details). If p denotes the output of the function `maxsel.test`, $1-p$ may be seen as the p-value of an independence test.

Usage

```
maxsel.test(x, y=NULL, type, statistic)
```

Arguments

<code>x</code>	a numeric vector of length <code>n</code> giving the values of the variable <code>X</code> for the considered <code>n</code> observations. The classes must be coded as <code>1,...,K</code> . Alternatively, <code>x</code> can be a <code>2 x K</code> matrix corresponding to a contingency table, where the two rows are for the values of <code>Y</code> (<code>Y=0,1</code>) and the <code>K</code> columns are for the values of <code>X</code> (<code>X=1,...,K</code>). In this case, <code>y</code> must be set to <code>y=NULL</code> .
<code>y</code>	a numeric vector of length <code>n</code> giving the class (response variable <code>Y</code>) of the considered observations. The classes must be coded as <code>0</code> and <code>1</code> . If <code>x</code> is a contingency table, <code>y</code> must be set to <code>y=NULL</code> .
<code>type</code>	the type of the considered binary splits. <code>type="ord"</code> corresponds to an ordinal <code>X</code> variable, <code>type="cat"</code> corresponds to a categorical <code>X</code> variable with unordered categories, <code>type="ord2"</code> corresponds to an ordinal <code>X</code> variable with 2 cutpoints (non-monotonous association).
<code>statistic</code>	the association measure used as criterion to select the best split. Currently, only <code>statistic="chi2"</code> (chi-square statistic) and <code>statistic="gini"</code> (the Gini-gain from machine learning) are implemented.

Details

For example, let us consider a variable `X` with the possible values `{1,2,3,4}`. If `type="ord"`, the set of candidate splits consists of `{1}{2,3,4}`, `{1,2}{3,4}` and `{1}{2,3,4}`. If `type="cat"`, the set of candidate splits consists of `{1}{2,3,4}`, `{1,2}{3,4}`, `{1,2,3}{4}`, `{1,2,4}{3}`, `{1,4}{2,3}`, `{1,3,4}{2}`, `{1,3}{2,4}`. If `type="ord2"`, the set of candidate splits consists of `{1}{2,3,4}`, `{1,2}{3,4}`, `{1,2,3}{4}`, `{1,2,4}{3}`, `{1,4}{2,3}`, `{1,3,4}{2}`.

Value

the probability that the maximally selected criterion is \leq than the value observed from the data, under the null-hypothesis of no association between `x` and `y`, given the numbers of observations with `Y=0, Y=1, X=1, ..., X=K`.

Author(s)

Anne-Laure Boulesteix (http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/020_professuren/boulesteix/)

References

- A.-L. Boulesteix (2006), Maximally selected chi-square statistics for ordinal variables, *Biometrical Journal* 48:451-462.
- A.-L. Boulesteix (2006), Maximally selected chi-square statistics and binary splits of nominal variables, *Biometrical Journal* 48:838-848.
- C. Strobl, A.-L. Boulesteix and T. Augustin (2007), Unbiased split selection for classification trees based on the Gini index, *Computational Statistics and Data Analysis* 52:483-501.
- A.-L. Boulesteix and C. Strobl (2006), Maximally selected chi-square statistics and non-monotonic associations: an exact approach based on two cutpoints. *Computational Statistics and Data Analysis* 51:6295-6306.

See Also[maxsel.](#)**Examples**

```
# load exactmaxsel library
library(exactmaxsel)

# First case: x and y are data vectors
# Simulate x and y
x<-sample(4,30,replace=TRUE)
y<-sample(c(0,1),30,replace=TRUE)

maxsel.test(x=x,y=y,type="ord",statistic="chi2")
maxsel.test(x=x,y=y,type="cat",statistic="gini")

# Second case: x is a contingency table, y=NULL.
x<-matrix(c(8,10,40,13,15,4),2,4,byrow=TRUE)
maxsel.test(x=x,y=NULL,type="ord",statistic="chi2")
maxsel.test(x=x,y=NULL,type="cat",statistic="gini")
```

Index

*Topic **datasets**

birth, [2](#)

*Topic **htest**

boundary, [3](#)

Fcat, [5](#)

Ford, [6](#)

Ford2, [7](#)

ginigain, [8](#)

maxsel, [10](#)

maxsel.test, [11](#)

birth, [2](#)

boundary, [3](#)

Fcat, [4](#), [5](#), [7](#), [8](#)

Ford, [4](#), [6](#), [6](#), [8](#)

Ford2, [6](#), [7](#), [7](#)

ginigain, [8](#)

maxsel, [6–8](#), [10](#), [13](#)

maxsel.test, [9](#), [11](#), [11](#)