Package 'psda'

May 24, 2020

Title Polygonal Symbolic Data Analysis

Version 1.4.0

Date 2020-05-24

Description

A toolbox in symbolic data framework as a statistical learning and data mining solution for symbolic polygonal data analysis. This study is a new approach in data analysis and it was proposed by Silva et al. (2019) <doi:10.1016/j.knosys.2018.08.009>. The package presents the estimation of main descriptive statistical measures, e.g, mean, covariance, variance, correlation and coefficient of variation.
In addition, a method to obtain polygonal data from classical data is presented. Empirical probability distribution function based on symbolic polygonal histogram and a regression model with its main measures are presented.

Depends R (>= 3.1)

License GPL-2

URL https://github.com/wagnerjorge/psda

BugReports https://github.com/wagnerjorge/psda/issues

Imports ggplot2, rgeos, plyr, sp, raster, stats

LazyData true

RoxygenNote 6.1.1

Suggests testthat, knitr, rmarkdown

VignetteBuilder knitr

NeedsCompilation no

Author Wagner Silva [aut, cre, ths],
Renata Souza [aut],
Francisco Cysneiros [aut]

Maintainer Wagner Silva <wjfs@cin.ufpe.br>

Repository CRAN

Date/Publication 2020-05-24 16:40:02 UTC

R topics documented:

`fitted.plr` *Extract Polygonal Linear Model Fitted Values*

Description

The function is used to calculate the fitted center and radius or fitted polygos from polygonal linear regression model.

Usage

```
## S3 method for class 'plr'
fitted(object, ..., polygon = FALSE, vertices)
```

Arguments

object	an object of the class "*plr*".
...	further arguments special methods could require.
polygon	logical. If *FALSE* the function returns the center and radius predicted for polygon. If *TRUE* the function returns an object of the class "*Polygonal*" representing the fitted polygons.
vertices	If *polygon* is *TRUE* a number of vertices should be defined. Besides, the number of vertices should be greater than 2 and equal to number of vertices chosen in symbolic polygonal variables.

Value

ans the fitted values for polygonal linear regression.

Examples

```
yp <- psim(10, 10) #simulate 10 polygons of 10 sides
xp1 <- psim(10, 10) #simulate 10 polygons of 10 sides
xp2 <- psim(10, 10) #simulate 10 polygons of 10 sides
e <- new.env()
e$yp <- yp
e$xp1 <- xp1
e$xp2 <- xp2
fit <- plr(yp~xp1+xp2-1, e)
fitted(fit) #shows the center and radius fitted from plr
fitted(fit, polygon = TRUE, vertices = 10) #Shows the polygon fitted from plr
```

---

longair *Airfares data (longair)*

Description

Longair data contains about quarterly average airfare and average weekly passengers for 4177 markets in 2001 of the U.S. Department of Transportation. The data can be seen in 'Polygonal data analysis: A new framework in symbolic data analysis' paper.

Usage

```
longair
```

Format

A data.frame with 1000 rows and 11 variables:

city1 City of boarding.

cit2 City of landing.

average_fare Average fare.

distance Distance between city of boarding and landing.

average_weekly_passengers Average weekly passengers.

market_leading_airline Market leading airline.

market_share Market share.

avarege_return_fare Average return fare

low_price_airline Lower price airline.

market_share2 Second market share

price Price of travel.

Source

https://www.sciencedirect.com/science/article/pii/S0950705118304052

`na.omit` *Handle Missing Values in Polygonal Objects*

Description

The function omits missing polygons.

Usage

```
na.omit(object, ...)
```

Arguments

`object`	objects of the class "*polygonal*".
`...`	further arguments special methods could require.

Value

polygons an object of the class "*polygonal*" without missing values.

Examples

```
y <- psim(5, 3)
y[[1]] <- NA
na.omit(y)
```

paggreg *Polygonal data aggregation*

Description

The function obtains symbolic data from classical data through the center and radius representation.

Usage

```
paggreg(data)
```

Arguments

data	A data frame with the first column of type factor.

Details

The class "aggregated" is composed by two data sets from center and range represetation. The first and second data set represent the center and radius, respectively.

Value

paggreg returns an objects of class "paggregated".

Examples

```
cat <- as.factor(sample(1:20, 1000, replace = TRUE))
cv <- runif(1000) #classical variable
cvc <- data.frame(category = cat, cv)
p <- paggreg(cvc)
```

parea *Polygonal Area*

Description

Compute the area of polygon.

Usage

```
parea(polygon)
```

Arguments

polygon	a matrix representing the polygon.

Value

a integer the area of polygon.

Examples

```
x <- psim(10, 10) #simulate 10 polygons of 10 sides
x <- x[[1]]
parea(x)
```

---

pconvex *Convex verification*

Description

Verify convexity of the polygons.

Usage

```
pconvex(polygon)
```

Arguments

polygon	A matrix with dimension l x 2, where l represent number of sides polygon.

Value

A boolean.

Examples

```
x <- psim(10, 10) #simulate 10 polygons of 10 sides
x <- x[[1]]
pconvex(x)
```

---

pcorr *Polygonal symbolic correlation*

Description

Compute the symbolic polygonal empirical correlation.

Usage

```
pcorr(polygons)
```

Arguments

polygons	A list of matrices of dimension l x 2, where l represent number of sides polygon.

Value

The method returns a integer.

Examples

```
x <- psim(10, 10) #simulate 10 polygons of 10 sides
pcorr(x)
```

---

pcov *Polygonal symbolic covariance*

Description

Compute the symbolic polygonal empirical covariance.

Usage

```
pcov(polygons)
```

Arguments

polygons	A list of polygonal datas.

Value

The method returns a integer.

Examples

```
x <- psim(10, 10) #simulate 10 polygons of 10 sides
pcov(x)
```

pfreq *Polygonal Symbolic Relative Frequency*

Description

Compute the bivariate relative frequency.

Usage

```
pfreq(pol)
```

Arguments

pol	A list of matrices of dimension l x 2, where l represent number of sides polygon.

Examples

```
x <- psim(10, 10) #simulate 10 polygons of 10 sides
frequency <- pfreq(x)
```

plr *Polygonal linear regression*

Description

plr is used to fit polygonal linear models.

Usage

```
plr(formula, data, model = TRUE, ...)
```

Arguments

formula	an object of class "formula": a symbolic description of the model to be fitted.
data	a environment that contains the variables of the study.
model	logicals. If TRUE the corresponding components of the fit are returned.
...	additional arguments to be passed to the low level polygonal linear regression fitting functions.

Details

Polygonal linear regression is the first model to explain the behavior of a symbolic polygonal variable in furnction to other polygonal variables, dependent and regressors, respectively. PLR is based on the least squares and uses the center and radius of polygons as representation them. The model is given by $y = X\beta + \epsilon$, where y, X, β, and ϵ is the dependent variable, matrix model, unknown parameters, and non-observed errors. In the model, the vector $y = (y_c^T, y_r)^T$, where y_c and y_r is the center and radius of center and radius. The matrix model $X = diag(X_c, X_r)$ for X_c and X_r describing the center and radius of regressors variables and finally, $\beta = (\beta_c^T, \beta_r^T)^T$. A detailed study about the model can be found in Silva et al.(2019).

Value

residuals is calculated as the response variable minus the fitted values.

rank the numeric rank of the fitted polygonal linear model.

call the matched call.

fitted.values the fitted mean values.

terms the terms.

coefficients a named vector of coefficients.

model the matrix model for center and radius.

References

Silva, W.J.F, Souza, R.M.C.R, Cysneiros, F.J.A. (2019) https://www.sciencedirect.com/science/article/pii/S0950705118304052.

Examples

```
yp <- psim(10, 10) #simulate 10 polygons of 10 sides
xp1 <- psim(10, 10) #simulate 10 polygons of 10 sides
xp2 <- psim(10, 10) #simulate 10 polygons of 10 sides
e <- new.env()
e$yp <- yp
e$xp1 <- xp1
e$xp2 <- xp2
fit <- plr(yp~xp1+xp2, e)
```

---

pmean *Polygonal empiric mean*

Description

Compute the polygonal empirical mean for polygonal variable.

Usage

```
pmean(polygons)
```

Arguments

polygons	A list of matrices of dimension l x 2, where l represent number of sides polygon.

Value

The method returns a vector containing the symbolic polygonal empirical mean in first and second dimension, respectively.

Examples

```
x <- psim(10, 10) #simulate 10 polygons of 10 sides
pmean(x)
```

---

pmean_id *Polygonal symbolic internal mean*

Description

Compute the symbolic polygonal empirical mean for only one observation (classes).

Usage

```
pmean_id(polygon)
```

Arguments

polygon	a matrix representing the polygon.

Value

a polygonal empiric mean of a polygon.

Examples

```
x <- psim(10, 10) #simulate 10 polygons of 10 sides
x <- x[[1]]
pmean_id(x)
```

pplot *Plot polygonal symbolic variable*

Description

Prints all overlaid polygons in the display. The polygons obtained through classes.

Usage

```
pplot(polygon, center = FALSE, color = "black")
```

Arguments

polygon	A list of matrices with dimension l x 2 where l represents vertices number of polygon.
center	logical. iF FALSE(the default) the center of polygon is not displayed.
color	A string that describes the color of center.

Examples

```
x <- psim(10, 10) #simulate 10 polygons of 10 sides
pplot(x, center = TRUE, color = 'red')
```

print.plr *Print method for Polygonal Linear Regression*

Description

`print.plr` is the `plr` method of the generic print function which prints its argument.

Usage

```
## S3 method for class 'plr'
print(x, digits = max(3L, getOption("digits") - 3L), ...)
```

Arguments

x	the object to be printed.
digits	a non-null value for `digits` specifies the minimum number of significant digits to be printed in values.
...	further arguments passed to or from other methods.

Examples

```
yp <- psim(10, 10) #simulate 10 polygons of 10 sides
xp1 <- psim(10, 10) #simulate 10 polygons of 10 sides
xp2 <- psim(10, 10) #simulate 10 polygons of 10 sides
e <- new.env()
e$yp <- yp
e$xp1 <- xp1
e$xp2 <- xp2
fit <- plr(yp~xp1 + xp2, data = e)
fit
```

`print.summary.plr` *Print Summary Polygonal Linear Regression*

Description

print arguments of the class "*summary.plr*" and returns it *invisibly* (via `invisible` (x)).

Usage

```
## S3 method for class 'summary.plr'
print(x, digits = max(3L, getOption("digits") -
  3L), concise = FALSE, ...)
```

Arguments

`x`	an object of the class "*summary.plr*".
`digits`	non-null value for `digits` specifies the minimum number of significant digits to be printed in values.
`concise`	a *logical* used to determine the type of digits.
`...`	further arguments special methods could require.

Examples

```
yp <- psim(50, 10) #simulate 50 polygons of 10 sides
xp1 <- psim(50, 10) #simulate 50 polygons of 10 sides
xp2 <- psim(50, 10) #simulate 50 polygons of 10 sides
e <- new.env()
e$yp <- yp
e$xp1 <- xp1
e$xp2 <- xp2
fit <- plr(yp~xp1 + xp2, data = e)
s <- summary(fit)
s
```

psim *Polygonal symbolic data simulation*

Description

Simulate a polygonal variable with one or more individuals.

Usage

```
psim(n, vertices)
```

Arguments

n	number of simulated polygons.
vertices	number of vertex of the polygon.

Details

The argument `radius` should have all values greater than zero. Otherwise, we cannot construct the polygons that compose the symbolic polygonal random variable. Besides, the size of the `center` vector should be equal to `range` vector.

Value

A list of polygons.

Examples

```
number_polygons <- 10
psim(number_polygons, 4)
```

psmi *Polygonal internal second moment*

Description

Caltulate symbolic polygonal internal second moment for polygonal data.

Usage

```
psmi(polygon)
```

Arguments

polygon	a matrix that represents a polygonal variable.

Value

The internal variance.

Examples

```
x <- psim(5, 3) #simulate 5 polygons of 3 sides
psmi(x[[1]])
```

`psymbolic` *Polygonal Symbolic Data*

Description

The function obtain a symbolic polygonal variables from data of class 'paggregated', i.e aggregated data. For this, the researcher need to select the number of vertices.

Usage

```
psymbolic(pdata, vertices)
```

Arguments

`pdata`	an object of the class 'paggregated' that represents the representation of symbolic polygonal data.
`vertices`	the number of vertices for the polygon.

Details

psymbolic converts data represented by center and radius representation in symbolic polygonal data. It is importat that the researcher considers a positive number for radius. Besides, the variable vertices should be greater than 2 for the number of vertices.

When the object of class 'paggregated' is composed by a vector for center and one vector for radius a simple symbolic variable is obtained.

Value

psdata is an object of class 'polygonal-variables', i.e. an environment, where for each object in the environment is a list with the polygons(matrix with dimention l times 2, where l represents the number of vertices).

Examples

```
## Obtaining a simple symbolic polygonal variable
cat1 <- as.factor(sample(1:20, 1000, replace = TRUE))
cv1 <- runif(1000) #classical variable
cvc1 <- data.frame(category = cat1, variable = cv1)
pol1 <- paggreg(cvc1)
out <- psymbolic(pol1, 6) #Hexagon
out$X1

## Obtaining three (or more) symbolic polygonal variables
cat2 <- as.factor(sample(1:20, 1000, replace = TRUE))
cv2 <- matrix(runif(3000), ncol = 3) #classical variable
cvc2 <- data.frame(category = cat2, cv2)
pol2 <- paggreg(cvc2)
out2 <- psymbolic(pol2, 8) #Octagon
out2$X1
out2$X2
out2$X3
```

---

pvar *Polygonal symbolic variance*

Description

Estime the symbolic polygonal empirical variance.

Usage

```
pvar(polygons)
```

Arguments

polygons	A list of matrices of dimension l x 2 where l represent number of sides polygon.

Value

The method returns a bi-dimensional vector.

Examples

```
x <- psim(8, 12) #simulate 8 polygons of 12 sides
pvar(x)
```

pvari *Polygonal internal variance*

Description

Caltulate the symbolic polygonal internal variance for a polygonal data.

Usage

```
pvari(polygon)
```

Arguments

polygon	a matrix that represents a polygonal variable.

Value

The internal variance.

Examples

```
x <- psim(10, 10) #simulate 10 polygons of 10 sides
pvari(x[[1]])
```

rmsea *Root mean squared error of area*

Description

Root mean squared error of area is a meansure proposed by Silva et al. (2019). It is used to evaluate the performance of symbolic polygonal linear regression model (plr).

Usage

```
rmsea(observed, fitted)
```

Arguments

observed	is the response variable of polygonal linear regression model.
fitted	are the polygons obtained from polygonal linear regression model as fitted values of the response variable.

Value

rmsea the value of the root mean squared error of area.

References

Silva, W.J.F, Souza, R.M.C.R, Cysneiros, F.J.A. (2019) https://www.sciencedirect.com/science/article/pii/S0950705118304052.

Examples

```
yp <- psim(10, 10) #simulate 10 polygons of 10 sides
xp1 <- psim(10, 10) #simulate 10 polygons of 10 sides
xp2 <- psim(10, 10) #simulate 10 polygons of 10 sides
e <- new.env()
e$yp <- yp
e$xp1 <- xp1
e$xp2 <- xp2
fit <- plr(yp~xp1+xp2-1, e)
yp_fitted <- fitted(fit, polygon = TRUE, vertices = 10) #Shows the polygon fitted from plr
rmsea(yp, yp_fitted)
```

---

saeb2017 *Basic Education Assessment System 2017 (SAEB 2017)*

Description

The dataset describes information about the Brazilian Basic Education Assessment System (SAEB) and infrastructure of the schools in 2017.

Usage

```
saeb2017
```

Format

A data.frame with 4037 observations (rows) and 13 variables, each row represent a county. One column indicates the county identification, the first six are the center of the polygons, and the six last are the radius of polygons. In datails:

county Identification of the county that participate of the SAEB.

proficiency_lp_center Leverage of Portuguese language proficiency score.

proficiency_mt_center Leverage of the Mathematics.

classroom_center Leverage number of classroom.

classroom_used_center Leverage number of classroom used.

employess Leverage number of employees of the schools.

proficiency_lp_radius Dispersion of the Portuguese language proficiency score.

proficiency_mt_radius Dispersion of the Mathematics proficiency score.

classroom_radius Classrooms dispersion.

classroom_used_radius Classrooms used dispersion.

computers_radius Dispertion of the computer numbers.

employees_radius Dispersion of the employess numbers.

spolygon *Symbolic Polygon*

Description

The function obtains a simple symbolic polygon from center and radius representation.

Usage

```
spolygon(center, radius, vertices)
```

Arguments

center	a integer that represents the barycenter of polygon.
radius	a integer that represents the radius of polygon.
vertices	represents the number of vertices for the polygon.

Value

matrix that represents the polygon.

Examples

```
spolygon(2.5, 3, 5) #pentagon
```

summary.plr *Summarizing Polygonal Linear Regression*

Description

summary method for class plr.

Usage

```
## S3 method for class 'plr'
summary(object, digits = max(3L, getOption("digits") - 3L),
  ...)
```

Arguments

object	an object of the class plr, usually, a result of a call to plr.
digits	a non-null value for digits specifies the minimum number of significant digits to be printed in values.
...	further arguments passed to or from other methods.

Value

residuals calculated as the response variable minus the fitted values.

sigma the given by square root of the estimated variance of the random error

$$\sigma^2 = \frac{\sum i = 1^n (y_i - \hat{y}_i)^2}{n - p - 1}$$

where p is two times the number of independent variables.

call the matched call.

aliased named logical vector showing if the original coefficients are aliased.

terms the `terms`.

coefficients a p x 4 matrix with columns for the estimated coefficient, its standard error, z-statistic and corresponding (two-sided) p-value.

Examples

```
yp <- psim(50, 10) #simulate 50 polygons of 10 sides
xp1 <- psim(50, 10) #simulate 50 polygons of 10 sides
xp2 <- psim(50, 10) #simulate 50 polygons of 10 sides
e <- new.env()
e$yp <- yp
e$xp1 <- xp1
e$xp2 <- xp2
fit <- plr(yp~xp1 + xp2, data = e)
s <- summary(fit)
```

---

wnba2014 *Women National Basketball Association 2014 (WNBA 2014).*

Description

The data set contains information about the season 2014. The data can be seen in 'Polygonal data analysis: A new framework in symbolic data analysis' paper.

Usage

```
wnba2014
```

Format

A data.frame with 4022 rows and 6 variables:

player_id Identification of player.

team_pts Number of points made by team.

opp_pts Number of points made by opponent.

minutes Minutes played.

fgatt Field goal attempts.

efficiency Efficiency.

Source

https://www.sciencedirect.com/science/article/pii/S0950705118304052

Index