Package 'pAnalysis'

October 14, 2022

Type Package

Title Benchmarking and Rescaling R2 using Noise Percentile Analysis

Version 2.0

Date 2016-01-19

Author Joseph G Kreke, PhD; Harris, Inc.
Sangeet Khemlani, PhD; Naval Research Laboratory.
Greg Trafton, PhD; Naval Research Laboratory.

Maintainer Joseph G Kreke <jkreke2@gmail.com>

Description Provides the tools needed to benchmark the R2 value corresponding to a certain acceptable noise level while also providing a rescaling function based on that noise level yielding a new value of R2 we refer to as R2k which is independent of both the number of degrees of freedom and the noise distribution function.

Depends ggplot2, coin, grDevices, graphics, stats

License GPL-2

NeedsCompilation no

Repository CRAN

Date/Publication 2016-01-19 23:47:42

R topics documented:

`pAnalysis-package` *Benchmarking and Rescaling R2 using Noise Percentile Analysis*

Description

R-sqaured (R2), as a function of n datapoints of x and y, is a standard goodness-of-fit measure that has the unfortunate behavior of becoming more sensitive to noise as the number of degrees of freedom (n) decreases. The mean of R2 measuring just noise is 1/(n-1). However, the distributions of R2 values measuring just noise varies greatly for each n: they are neither uniform nor consistent in shape, especially at low n. At the next-to-lowest value of n, where n=3, the mean R2 value is 0.5 but the distribution of possible R2 values is symmetric about that point - rising from the mean (0.5) toward the extremes of both 0 and 1 - and every other possible value of R2 is more likely than the mean. When n=3, R2 values of 0 or 1 (the extremes) are more than 30 times more likely than the value of 0.5 (P(R2>0.999 or R2<0.001)=0.020; P(R2>0.499 and R2<0.501)=0.00069). For n=4 and higher, the distributions of R2 are not symmetric about the mean and high values of R2 are not as likely as they are at n=3 but there are still significant probabilities of achieving high R2 values. As n increases, the probability of obtaining high R2 values with just noise decreases sharply. We invite the reader to run the plotpdf() function for 3, 4 and 5 degrees of freedom. See plotpdf() examples for syntax.

Instead of judging the validity of a particular value of R2 by comparing it to the mean of the noise distributions (1/(n-1)), we consider how the percentiles of R2 - measuring noise only - vary with respect to n. For a given n, we conduct many measurements of R2 using numbers randomnly assigned according to a particular noise distribution function. Then, for a given percentile (p) of noise, we find the value of R2 that is above p percent of all R2 values which then becomes the baseline, R2p. Hence, if one knows the n, how the noise is distributed (dist) and what noise level to stay above (p), one can find the baseline noise (R2p) using the R2p function. We use the normal distribution (dist='normal') and the 95th percentile (p=0.95) as defaults. See plotcdf().

We also provide a function (R2pTable) that will output a table of R2p values based on several degrees of freedom and several percentiles you may want to have handy. Use a pctlist equal to the percentiles you would like to see, e.g. pctlist=c(0.9, 0.95, 0.99).

In addition, we also provide a function, R2k, one can use to rescale one or more measurements of R2 to a particular pct and n. One can argue that any value of R2 that equals R2p for a particular noise percentile (p) and number of degrees of freedom (n), must be equivalent to any other value of R2 if it equals R2p for a different n. (We do not presume the same can be said of different values of p.) In other words, all values of R2 along an R2p curve (see plotR2p()) sit at the border between acceptable and unacceptable noise. For a particular p, a measured value of R2 falling on the R2p curve has just as much chance (1-p) of being brought about by noise as any other value of R2 that falls on the same R2p curve (different n, same p). Therefore, any R2 value falling on the R2p curve is equivalent in terms of measuring goodness of fit. Values of R2 that sit above the R2p curve, then establish a ratio we define as R2k = (R2-R2p)/(1-R2p). This ratio, R2k, then establishes a line of equivalency: all values on this line reside at the same fractional distance away from the baseline and therefore have a measure that is equivalent to the original R2 measure. See plotR2k() and plotR2Equiv().

R2k has several important features. 1. Its range of possible values is negative infinity to +1. A negative value is a quick indicator that the associated R2 measure is indistinguishable from noise

and a positive value means it is above the noise whose magnitude indicates how far it is above the noise. 2. It is independent of n, which means it can be directly compared to R2ks obtained from other R2 measurements using different n. 3. It is independent of the noise distribution. Once the R2p value is obtained for a given set of parameters (n, p, dist), the associated, rescaled R2k values can be directly compared. However, R2k values coming from different noise baselines (R2ps) can not be directly compared.

Author(s)

Joseph G Kreke, PhD; Harris, Inc. Sangeet Khemlani, PhD; Naval Research Laboratory. Greg Trafton, PhD; Naval Research Laboratory.

Maintainer: Joseph G Kreke <jkreke2@gmail.com>

References

Khemlani, Sangeet; Kreke, Joseph; Trafton, Greg. "Using Percentile Analysis to Baseline Noise in R-squared". Harris, Inc; Naval Research Laboratory. (in draft)

`cap1` *cap1*

Description

Simple function to capitalize letters

Usage

```
cap1(x)
```

Arguments

`x`	Character variable

Value

The output of cap 1 is a capitalized character variable

Author(s)

Joseph G. Kreke, PhD

Examples

```
uncappedtitle <- "this title"
  cappedtitle <- cap1(uncappedtitle)
```

pcdfs *R2 values with corresponding probability density and cumulative density functions.*

Description

This function builds a data frame of all possible R2 values over its range of 0 to 1, with corresponding values of probability (pdf) and cumulative probability (cdf) for a given number of degrees of freedom. R2 is divided uniformly over its range into bins whose width is determined by the number of decimal places chosen (default=3). The number of samples is determined by order (10^order). Values of the cumulative density function (cdf) are used to calculate the baseline noise level, R2p.

Usage

```
pcdfs(dof, order = 6, ndecimals = 3, dist = "normal", par1 = 0, par2 = 1)
```

Arguments

dof	an integer greater than 1
order	a positive number used to set the order of magnitide of the number of samples (default is 6)
ndecimals	a positive integer describing the number of decimal places desired in the results
dist	a character string identifying the noise distribution. The current list of possible distributions is, 'normal', 'uniform', 'lognormal', 'poisson' and 'binomial'.
par1	one of two parameters used to define the noise distribution For 'normal', par1 = mean, For 'uniform', par1 = min, For 'lognormal', par1 = logmean, For 'poisson', par1=lambda, For 'binomial', par1=size
par2	the second of two parameters used to define the noise distribution For 'normal', par2 = std dev, For 'uniform', par2 = max, For 'lognormal', par2 = log std dev, For 'poisson', par2=(not used), For 'binomial', par2=probability

Value

pcdfs returns a data frame with columns "R2", "pdf" and "cdf". R2 is the full range of values that R2 can possibly have (from 0 to 1) divided by 10^bw where bw (bin width). binwidth is determined by ndecimals so 10^bw = 10^(-ndecimals). pdf is the probability density function – the probability of obtaining a specific range of values of R2 corresponding to one of the bins. Values range from 0 to 1. cdf is the cumulative pdf. Values of cdf also range from 0 to 1.

Author(s)

Joseph G. Kreke, PhD

Examples

```
R2df <- pcdfs(dof=8, order=6, ndecimals=3, dist="uniform")
R2df <- pcdfs(5)
```

plotcdf *Plot several Cumulative Density Functions*

Description

Plots the cumulative probability density function for a given number of degrees of freedom (dof) and a noise distribution function

Usage

```
plotcdf(dof, order = 4, dist = "normal", ...)
```

Arguments

dof	the degrees of freedom of interest
order	the order of magnitude of the number of samples desired for the plot
dist	the noise distribution: 'normal', 'uniform', 'lognormal', 'poisson', 'binomial'
...	other arguments used in pcdfs().

Value

The output of plotcdf() is a ggplot object

Author(s)

Joseph G. Kreke, PhD

Examples

```
plt <- plotcdf(dof=10, dist="lognormal")
plt <- plotcdf(4,order=5,dist='binomial',par1=10,par2=0.75)
```

plotpdf *Plot Several Probability Density Functions*

Description

Plots the probability density function for a given number of degrees of freedom (dof) and a noise distribution function

Usage

```
plotpdf(dof, order = 4, dist = "normal", ...)
```

Arguments

dof	the number of degrees of freedom
order	the order of magnitude of the number of samples desired for the plot
dist	the noise distribution function. "normal" by default)
...	other arguments used in calls to pcdfs()

Value

The output of plotpdf is a ggplot object

Author(s)

Joseph G. Kreke, PhD

Examples

```
plt <- plotpdf(3)
plt <- plotpdf(5,order=6)
```

---

plotR2Equiv *Plotting Equivalent R2s across a range of degrees of freedom.*

Description

For given values of R2, degrees of freedom (dof) and a percentile noise level(pct), this will plot the noise baseline (R2p) and equivalent R2 based on R2K.

Usage

```
plotR2Equiv(R2, dof, pct = 0.95, order = 4, plot_pctr2 = F, ...)
```

Arguments

R2	a number between 0 and 1
dof	an integer number >= 3
pct	percentile of allowable noise expressed as a number between 0 and 1. Default is 0.95.
order	order of magnitude of the number of samples
plot_pctr2	adds the plot of R2p equal to R2
...	other arguments used in calls to pcdfs()

Value

The output of plotR2Equiv() is a ggplot object

Author(s)

Joseph G. Kreke, PhD

Examples

```
plt <- plotR2Equiv(R2=0.83, dof=10, pct=0.99)
plt <- plotR2Equiv(0.7,5)
```

`plotR2k` *Plot R2k for a single measured R2 and a single noise percentile across a range of degrees of freedom*

Description

This function plots R2k values presuming that the same R2 value was obtained using varying numbers of degrees of freedom. Provide the R2 value of interest and the desired noise baseline level (pct).

Usage

```
plotR2k(R2, doflist = c(2:30), pct = 0.95, order = 4, ndecimals = 3, ...)
```

Arguments

`R2`	a number between 0 and 1
`doflist`	dof list - a vector of integers > 1
`pct`	percentile of allowable noise expressed as a number between 0 and 1. Default is 0.95.
`ndecimals`	the number of desired decimal places in the result
`order`	order of magnitude of the number of samples
`...`	other arguments used by pcdfs()

Value

The output of this function is a ggplot object.

Author(s)

Joseph G. Kreke, PhD

Examples

```
plt = plotR2k(R2=0.77, pct=0.90)
plt = plotR2k(0.5)
```

plotR2p *Plot several noise baselines (R2p)*

Description

Plots R2 values at several baseline noise levels (pct). Measured R2 values above the baseline can be distinguished from noise while those R2 values below the baseline can not.

Usage

```
plotR2p(doflist = c(2:30), pctlist = c(0.95), order = 4, ndecimals = 3, ...)
```

Arguments

doflist	a vector of degrees of freedom, integer numbers >=2
pctlist	a vector of percentiles of acceptable noise expressed as numbers between 0 and 1
order	a single real number > 3 and < 7. Defaults are 5 and 6)
ndecimals	the number of decimal places desired for the result. an integer number > 0.
...	other arguments used by pcdfs()

Value

The output of this function is a ggplot object

Author(s)

Joseph G. Kreke, PhD

Examples

```
plt <- plotR2p(doflist=c(2:30), pctlist=0.95, order=4)
```

R2 *R-squared*

Description

Simple measure of R-squared

Usage

```
R2(x, y)
```

Arguments

x	a vector of real numbers
y	a vector of real numbers; must be the same length as x

Value

R2 output is a number between 0 and 1

Author(s)

Joseph G. Kreke, PhD

Examples

```
x=c(1,2,3,4,5,6)
y=c(1.2, 2.1, 2.9, 3.9, 5.3, 6.0)
r2 <- R2(x,y)
```

R2k *Conversion of standard R2 to a noise/dof-independent value*

Description

This function converts a vector of R2 values to a vector of noise-baselined, dof-independent and noise distribution-independent values. The resulting R2k values may vary from -inf to +1 where any negative value indicates it is indistinguishable from noise and should be discarded. Positive values indicate the R2k value is distinguishable from noise and allow direct comparison to other R2k values that may have been arrived at from models of different degrees of freedom.

Usage

```
R2k(R2, dof, pct=0.95, ndecimals=3,...)
```

Arguments

R2	a vector of real numbers between 0 and 1
dof	the number of degrees of freedom; an integer.
pct	percentile of allowable noise expressed as a number between 0 and 1. Default is 0.95.
ndecimals	the number of decimal places in the result
...	other arguments used in calls to pcdfs()

Value

R2k is a value between 0 and 1

Author(s)

Joseph G. Kreke, PhD

Examples

```
r2a <- 0.839
dof <- 10
r2ka <- R2k(r2a, dof)
r2b <- runif(n=20,min=0.71,max=0.73)
r2kb <- R2k(r2b, dof)
```

R2p — *Calculation of baseline noise level (R2p) at a single value of degrees of freedom (dof)*

Description

This function determines the value of R2, called R2p here, below which a certain percentile level of noise is present. Any models with R2 values below this baseline R2 value are therefore indistinguishable from noise.

Usage

```
R2p(dof, pct = 0.95, ndecimals = 3,...)
```

Arguments

dof	degrees of freedom; an integer
pct	percentile of allowable noise expressed as a number between 0 and 1. Default is 0.95.
ndecimals	the number of decimal places in the result
...	other arguments used by pcdfs()

Value

R2p is a real number between 0 and 1

Author(s)

Joseph G. Kreke, PhD

Examples

```
pct <- 0.95
dof <- 10
  r2p <- R2p(dof, pct)
```

R2pTable *Creates a table R2p values for combinations of degrees of freedom and percentiles*

Description

R2pTable builds a table (a data frame) of baseline noise levels (R2p values) for each combination of degree of freedom and percentile. A matrix is created with the number of rows equal to the length of doflist and the number of columns equal to the length of pctlist. The elements of this matrix are the results of calls to the R2p function with arguments of each of combination of the elements of doflist and pctlist. Additional arguments desired for R2p can be passed along through these calls. The resulting matrix is converted to a data frame. Although it takes a few seconds longer, we recommend using order=5 for sufficient accuracy. (order=4 is the default to meet the CRAN recommendation that default functions should take no more than a few seconds.)

Usage

```
R2pTable(doflist = NULL, pctlist = NULL, order = 4, ndecimals = 2,...)
```

Arguments

doflist	a vector of integers greater than 1
pctlist	a vector of percentiles of acceptable noise expressed as numbers between 0 and 1
order	order of magnitude of samples
ndecimals	the number of decimal places in the result
...	refers to any argument used by calls with the R2pTable routine, specifically, R2p() and pcdfs()

Details

R2pTable can be used to generate a handy table of R2p values. R2pTable is also useful for generating a table used for plotting R2p for several values of pct. However, when generating many values, the processing time increases and it might take awhile to build the table. It takes about 1min to generate R2ps for 60 degrees of freedom with order=5 and one value of pct.

Value

R2pTable returns a data frame of R2p values – each column corresponds to a different percentile and each row's name corresponds to a different degree of freedom.

Note

Running R2pTable with defaults takes about 20s on a MacBook Pro laptop.

Author(s)

Joseph G. Kreke, PhD

Examples

```
tab <- R2pTable(doflist=c(3,4,5),pctlist=c(0.7,0.8,0.9))
```

Index