Package 'bulkQC'

April 27, 2024

Type Package

Title Quality Control and Outlier Identification in Bulk for Multicenter Trials

Version 1.1

Date 2024-04-25

Author Joseph Rigdon <jrigdon@wakehealth.edu>

Maintainer Joseph Rigdon <jrigdon@wakehealth.edu>

Description

Multicenter randomized trials involve the collection and analysis of data from numerous study participants across multiple sites. Outliers may be present. To identify outliers, this package examines data at the individual level (univariate and multivariate) and site-level (with and without covariate adjustment). Methods are outlined in further detail in Rigdon et al (to appear).

License GPL-3

Depends isotree, stddiff

NeedsCompilation no

Repository CRAN

Date/Publication 2024-04-27 17:50:02 UTC

R topics documented:

bulkQC-package *Quality Control and Outlier Identification in Bulk for Multicenter Trials*

Description

Multicenter randomized trials involve the collection and analysis of data from numerous study participants across multiple sites. Outliers may be present. To identify outliers, this package examines data at the individual level (univariate and multivariate) and site-level (with and without covariate adjustment). Methods are outlined in further detail in Rigdon et al (to appear).

Details

Package:	bulkQC
Type:	Package
Version:	1.1
Date:	2024-04-24
License:	GPL-3

Author(s)

Joseph Rigdon <jrigdon@wakehealth.edu>

References

Tukey J. Exploratory Data Analysis. 1st edition. Reading, Mass: Pearson; 1977. 712 p.

Cortes D. Explainable outlier detection through decision tree conditioning. arXiv:200100636 [cs, stat] [Internet]. 2020 Jan 2 [cited 2021 Nov 12]; Available from: http://arxiv.org/abs/2001.00636

Yang D, Dalton JE. A unified approach to measuring the effect size between two groups using SAS. 2012;6

getOutliers *Function to obtain values outside of whiskers on boxplot*

Description

Within vector of continuous data, identifies and outputs those values sufficiently smaller than the first quartile (Q1) or larger than the third quartile (Q3). Used in consort with ind_uni function to identify individual univariate outliers.

Usage

```
getOutliers(x, mult = 1.5)
```

Arguments

x	A vector of continuous data
mult	A multiplier on the interquartile range (IQR) to flag outliers that are mult*IQR less than Q1 or mult*IQR greater than Q3; default is 1.5

Value

Returns the subset of observations in x that are mult*IQR less than Q1 or mult*IQR greater than Q3

References

Tukey J. Exploratory Data Analysis. 1st edition. Reading, Mass: Pearson; 1977. 712 p.

See Also

ind_uni

Examples

```
ex = c(rnorm(95), -10, -8, 7, 9, 11)
getOutliers(ex)
getOutliers(ex, mult=3)
```

---

ind_multi *Identifies individual multivariate outliers*

Description

Discovers potential individual multivariate outliers by identifying and returning those observations with outlier score greater than a threshold. The outlier score is calculated using single isolation forests.

Usage

```
ind_multi(d0, exclude = c("pid", "site"), thresh = 0.7, n_uniq = 10)
```

Arguments

d0	A data frame with columns as variables and rows as observations
exclude	A vector of names of variables to exclude in outlier identification
thresh	Threshold (0-1) that an outlier score must exceed to be flagged for further investigation
n_uniq	Number of unique observations of a variable needed for outlier identification to be performed

Details

The function evaluates multivariate observations from each row consisting of those variables not excluded by the 'exclude' argument above. For each multivariate observation, an outlier score is calculated using single isolation forests. Those multivariate observations that are isolated earliest in a decision tree have a lower tree depth, in turn have higher outlier scores, and are thought more likely to be outliers.

Value

nID	The number of observations evaluated
nVar	The number of variables evaluated
data	A data frame containing those observations deemed to be potential outliers that appends the outliers with the excluded variables to aid in interpretation, and includes an outlier score for each row

References

Cortes D. Explainable outlier detection through decision tree conditioning. arXiv:200100636 [cs, stat] [Internet]. 2020 Jan 2 [cited 2021 Nov 12]; Available from: http://arxiv.org/abs/2001.00636

Examples

```
data(iris)
iris2 = iris
iris2$pid = 1:dim(iris2)[1]
ind_multi(iris2, exclude=c("pid", "Species"), thresh=0.7, n_uniq=10)
ind_multi(iris2, exclude=c("pid", "Species"), thresh=0.6, n_uniq=10)
```

---

ind_uni *Identifies individual univariate outliers*

Description

Discovers potential individual univariate outliers by identifying and returning those observations outside of the whiskers on a boxplot

Usage

```
ind_uni(d0, exclude = c("pid", "site"), n_uniq = 10, m = 1.5)
```

Arguments

d0	A data frame with columns as variables and rows as observations
exclude	A vector of names of variables to exclude in outlier identification
n_uniq	Number of unique observations of a variable needed for outlier identification to be performed
m	A multiplier on the interquartile range (IQR) to flag outliers that are mult*IQR less than Q1 or mult*IQR greater than Q3; default is 1.5

Value

nID	The number of observations evaluated
nVar	The number of variables evaluated
data	A data frame containing those observations deemed to be potential outliers that appends the outliers with the excluded variables to aid in interpretation

References

Tukey J. Exploratory Data Analysis. 1st edition. Reading, Mass: Pearson; 1977. 712 p.

See Also

getOutliers

Examples

```
data(iris)
iris2 = iris
iris2$pid = 1:dim(iris2)[1]
ind_uni(iris2, exclude=c("pid", "Species"), m=1.5)
ind_uni(iris2, exclude=c("pid", "Species"), m=3)
```

---

nVal *Unique values in a vector*

Description

Returns number of unique values in a vector

Usage

```
nVal(x)
```

Arguments

x	A vector of observations of a variable

Value

The number of unique values in the vector

Examples

```
nVal(runif(10))
```

`site_outliers` *Identifies site level outliers*

Description

Discovers potential site level outliers by using unadjusted and adjusted regression models and standardized difference calculations.

Usage

```
site_outliers(d0, exclude = c("pid"), siteID = "site", covs = c("age"), threshG = 0.001,
thresh2 = 0.05, threshS = 0.5, n_uniq = 10, n_dec = 4, n_decS = 2)
```

Arguments

`d0`	A data frame with columns as variables and rows as observations
`exclude`	A vector of names of variables to exclude in outlier identification
`siteID`	The name of the variable in the data frame that identifies sites
`covs`	A vector of covariates to adjust for in the adjusted regression models
`threshG`	P-value threshold for global test equal means across sites
`thresh2`	P-value threshold for comparison of reference site vs. all other sites
`threshS`	Standardized difference threshold above which a site difference is deemed meaningfully large
`n_uniq`	Number of unique observations of a variable needed for outlier identification to be performed
`n_dec`	Number of decimals to display for p-values in output
`n_decS`	Number of decimals to display for standardized differences in output

Details

The function compares the distribution of a given variable across sites by first conducting a global test of equal means (without and with adjustment for covariates of interest). Among those variables where the null hypothesis of equal means across sites is rejected, the function then compares each site vs. all other sites using unadjusted and adjusted comparisons. The unadjusted comparisons include a two-sample t-test with equal variance and a standardized difference calculation. The adjusted comparisons include a linear regression model with an indicator variable for reference site and user-specified covariates, and an adjusted standardized difference calculated as the model coefficient for site divided by the model estimated root mean squared error.

Value

`overall`	A matrix with rows as variables where global test of equal means is rejected and columns as the corresponding p-values from the unadjusted and adjusted statistical tests

`sitewise_P` For the variables identified by the global tests (columns), the unadjusted p-values (from two-sample t-test) comparing each site to all other sites (rows). Values above threshold printed as missing.

`sitewise_P_adj` For the variables identified by the global tests (columns), the adjusted p-values (from linear regression model) comparing each site to all other sites (rows). Values above threshold printed as missing.

`sitewise_StDf` For the variables identified by the global tests (columns), the unadjusted standardized differences comparing each site to all other sites (rows). Values below threshold printed as missing.

`sitewise_StDf_adj` For the variables identified by the global tests (columns), the adjusted standardized differences comparing each site to all other sites (rows). Values below threshold printed as missing.

References

Yang D, Dalton JE. A unified approach to measuring the effect size between two groups using SAS. 2012;6

Examples

```
data(iris)
iris2 = iris
iris2$temp = rnorm(dim(iris2)[1]) #for covariate adjustment
site_outliers(iris2, site="Species", covs=c("temp"))
```

Index