Package 'correlation'

May 5, 2020

Type Package

Title Methods for Correlation Analysis

Version 0.2.1

Maintainer Dominique Makowski <dom.makowski@gmail.com>

URL https://easystats.github.io/correlation/

BugReports https://github.com/easystats/correlation/issues

Description Lightweight package for computing different kinds of correlations, such as partial correlations, Bayesian correlations, multilevel correlations, polychoric correlations, biweight correlations, distance correlations and more. Relies on the easystats ecosystem (Lüdecke, Waggoner & Makowski (2019) <doi:10.21105/joss.01412>).

Imports stats, insight (>= 0.8.0), bayestestR (>= 0.5.0), parameters (>= 0.5.0), effectsize (>= 0.2.0)

Suggests BayesFactor, dplyr, energy, ggcorrplot, ggplot2, Hmisc, knitr, lme4, polycor, ppcor, psych, rmcorr, testthat, tidyr, covr, rstanarm, rmarkdown, see, WRS2

License GPL-3

Encoding UTF-8

LazyData true

RoxygenNote 7.1.0

NeedsCompilation no

Author Dominique Makowski [aut, cre] (<https://orcid.org/0000-0001-5375-9967>), Daniel Lüdecke [aut] (<https://orcid.org/0000-0002-8895-3206>), Mattan S. Ben-Shachar [ctb] (<https://orcid.org/0000-0002-4287-4801>), Indrajeet Patil [ctb] (<https://orcid.org/0000-0003-1995-6531>)

Repository CRAN

Date/Publication 2020-05-05 05:10:07 UTC

R topics documented:

`correlation` *Correlation Analysis*

Description

Performs a correlation analysis.

Usage

```
correlation(
  data,
  data2 = NULL,
  method = "pearson",
  p_adjust = "holm",
  ci = 0.95,
  bayesian = FALSE,
  bayesian_prior = "medium",
  bayesian_ci_method = "hdi",
  bayesian_test = c("pd", "rope", "bf"),
  redundant = FALSE,
  include_factors = FALSE,
  partial = FALSE,
  partial_bayesian = FALSE,
  multilevel = FALSE,
  robust = FALSE,
  ...
)
```

Arguments

data	A data frame.
data2	An optional data frame.
method	A character string indicating which correlation coefficient is to be used for the test. One of "pearson" (default), "kendall", or "spearman", "biserial", "polychoric", "tetrachoric", "biweight", "distance", "percentage" (for percentage bend correlation) or "shepherd" (for Shepherd's Pi correlation). Setting "auto" will attempt at selecting the most relevant method (polychoric when ordinal factors involved, tetrachoric when dichotomous factors involved, point-biserial if one dichotomous and one continuous and pearson otherwise).
p_adjust	Correction method for frequentist correlations. Can be one of "holm" (default), "hochberg", "hommel", "bonferroni", "BH", "BY", "fdr" or "none".
ci	Confidence/Credible Interval level. If "default", then it is set to 0.95 (95% CI).
bayesian	If TRUE, will run the correlations under a Bayesian framework. Note that for partial correlations, you will also need to set `partial_bayesian` to `TRUE` to obtain "full" Bayesian partial correlations. Otherwise, you will obtain pseudo-Bayesian partial correlations (i.e., Bayesian correlation based on frequentist partialization).
bayesian_prior	For the prior argument, several named values are recognized: "medium.narrow", "medium", "wide", and "ultrawide". These correspond to scale values of 1/sqrt(27), 1/3, 1/sqrt(3) and 1, respectively. See the `BayesFactor::correlationBF` function.
bayesian_ci_method	See arguments in `model_parameters` for `BayesFactor` tests.
bayesian_test	See arguments in `model_parameters` for `BayesFactor` tests.
redundant	Should the data include redundant rows (where each given correlation is repeated two times).
include_factors	If `TRUE`, the factors are kept and eventually converted to numeric or used as random effects (depending of `multilevel`). If `FALSE`, factors are removed upfront.
partial	Can be TRUE or "semi" for partial and semi-partial correlations, respectively.
partial_bayesian	If TRUE, will run the correlations under a Bayesian framework. Note that for partial correlations, you will also need to set `partial_bayesian` to `TRUE` to obtain "full" Bayesian partial correlations. Otherwise, you will obtain pseudo-Bayesian partial correlations (i.e., Bayesian correlation based on frequentist partialization).
multilevel	If `TRUE`, the factors are included as random factors. Else, if `FALSE` (default), they are included as fixed effects in the simple regression model.
robust	If TRUE, will rank-transform the variables prior to estimating the correlation. Note that, for instance, a Pearson's correlation on rank-transformed data is equivalent to a Spearman's rank correlation. Thus, using `robust=TRUE` and `method="spearman"` is redundant. Nonetheless, it is an easy way to increase the robustness of the correlation (as well as obtaining Bayesian Spearman rank Correlations).
...	Arguments passed to or from other methods.

Details

Correlation Types:

- **Pearson's correlation**: The covariance of the two variables divided by the product of their standard deviations.
- **Spearman's rank correlation**: A non-parametric measure of rank correlation (statistical dependence between the rankings of two variables). The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not).
- **Kendall's rank correlation**: In the normal case, the Kendall correlation is preferred than the Spearman correlation because of a smaller gross error sensitivity (GES) and a smaller asymptotic variance (AV), making it more robust and more efficient. However, the interpretation of Kendall's tau is less direct than that of Spearman's rho, in the sense that it quantifies the difference between the % of concordant and discordant pairs among all possible pairwise events.
- **Biweight midcorrelation**: A measure of similarity between samples that is median-based, rather than mean-based, thus is less sensitive to outliers, and can be a robust alternative to other similarity metrics, such as Pearson correlation.
- **Distance correlation**: Distance correlation measures both linear and nonlinear association between two random variables or random vectors. This is in contrast to Pearson's correlation, which can only detect linear association between two random variables.
- **Percentage bend correlation**: Introduced by Wilcox (1994), it is based on a down-weight of a specified percentage of marginal observations deviating from the median (by default, 20%).
- **Shepherd's Pi correlation**: Equivalent to a Spearman's rank correlation after outliers removal (by means of bootstrapped mahalanobis distance).
- **Point-Biserial and biserial correlation**: Correlation coefficient used when one variable is continuous and the other is dichotomous (binary). Point-serial is equivalent to a Pearson's correlation, while Biserial should be used when the binary variable is assumed to have an underlying continuity. For example, anxiety level can be measured on a continuous scale, but can be classified dichotomously as high/low.
- **Polychoric correlation**: Correlation between two theorised normally distributed continuous latent variables, from two observed ordinal variables.
- **Tetrachoric correlation**: Special case of the polychoric correlation applicable when both observed variables are dichotomous.

Partial Correlation: **Partial correlations** are estimated as the correlation between two variables after adjusting for the (linear) effect of one or more other variable. The correlation test is then run after having partialized the dataset, independently from it. In other words, it considers partialization as an independent step generating a different dataset, rather than belonging to the same model. This is why some discrepancies are to be expected for the t- and p-values, CIs, BFs etc (but *not* the correlation coefficient) compared to other implementations (e.g., `ppcor`). (The size of these discrepancies depends on the number of covariates partialled-out and the strength of the linear association between all variables.)

Multilevel correlations are a special case of partial correlations where the variable to be adjusted for is a factor and is included as a random effect in a mixed model.

Notes:

- Kendall and Spearman correlations when `bayesian=TRUE`: These are technically Pearson Bayesian correlations of rank transformed data, rather than pure Bayesian rank correlations (which have different priors).

Value

A correlation object that can be displayed using the `print`, `summary` or `table` methods.

Multiple tests correction: About multiple tests corrections.

Examples

```
library(correlation)
cor <- correlation(iris)

cor
summary(cor)
summary(cor, redundant = TRUE)

# Grouped dataframe
if (require("dplyr")) {
  iris %>%
    group_by(Species) %>%
    correlation()
}

# automatic selection of correlation method
correlation(mtcars[-2], method = "auto")
```

---

cor_test *Correlation test*

Description

This function performs a correlation test between two variables.

Usage

```
cor_test(
  data,
  x,
  y,
  method = "pearson",
  ci = 0.95,
  bayesian = FALSE,
  bayesian_prior = "medium",
  bayesian_ci_method = "hdi",
  bayesian_test = c("pd", "rope", "bf"),
```

```
  include_factors = FALSE,
  partial = FALSE,
  partial_bayesian = FALSE,
  multilevel = FALSE,
  robust = FALSE,
  ...
)
```

Arguments

`data`	A data frame.
`x, y`	Names of two variables present in the data.
`method`	A character string indicating which correlation coefficient is to be used for the test. One of "pearson" (default), "kendall", or "spearman", "biserial", "polychoric", "tetrachoric", "biweight", "distance", "percentage" (for percentage bend correlation) or "shepherd" (for Shepherd's Pi correlation). Setting "auto" will attempt at selecting the most relevant method (polychoric when ordinal factors involved, tetrachoric when dichotomous factors involved, point-biserial if one dichotomous and one continuous and pearson otherwise).
`ci`	Confidence/Credible Interval level. If "default", then it is set to 0.95 (95% CI).
`bayesian, partial_bayesian`	If TRUE, will run the correlations under a Bayesian framework. Note that for partial correlations, you will also need to set `partial_bayesian` to `TRUE` to obtain "full" Bayesian partial correlations. Otherwise, you will obtain pseudo-Bayesian partial correlations (i.e., Bayesian correlation based on frequentist partialization).
`bayesian_prior`	For the prior argument, several named values are recognized: "medium.narrow", "medium", "wide", and "ultrawide". These correspond to scale values of 1/sqrt(27), 1/3, 1/sqrt(3) and 1, respectively. See the `BayesFactor::correlationBF` function.
`bayesian_ci_method, bayesian_test`	See arguments in `model_parameters` for `BayesFactor` tests.
`include_factors`	If `TRUE`, the factors are kept and eventually converted to numeric or used as random effects (depending of `multilevel`). If `FALSE`, factors are removed upfront.
`partial`	Can be TRUE or "semi" for partial and semi-partial correlations, respectively.
`multilevel`	If `TRUE`, the factors are included as random factors. Else, if `FALSE` (default), they are included as fixed effects in the simple regression model.
`robust`	If TRUE, will rank-transform the variables prior to estimating the correlation. Note that, for instance, a Pearson's correlation on rank-transformed data is equivalent to a Spearman's rank correlation. Thus, using `robust=TRUE` and `method="spearman"` is redundant. Nonetheless, it is an easy way to increase the robustness of the correlation (as well as obtaining Bayesian Spearman rank Correlations).
`...`	Arguments passed to or from other methods.

Details

Correlation Types:

- **Pearson's correlation**: The covariance of the two variables divided by the product of their standard deviations.
- **Spearman's rank correlation**: A non-parametric measure of rank correlation (statistical dependence between the rankings of two variables). The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not).
- **Kendall's rank correlation**: In the normal case, the Kendall correlation is preferred than the Spearman correlation because of a smaller gross error sensitivity (GES) and a smaller asymptotic variance (AV), making it more robust and more efficient. However, the interpretation of Kendall's tau is less direct than that of Spearman's rho, in the sense that it quantifies the difference between the % of concordant and discordant pairs among all possible pairwise events.
- **Biweight midcorrelation**: A measure of similarity between samples that is median-based, rather than mean-based, thus is less sensitive to outliers, and can be a robust alternative to other similarity metrics, such as Pearson correlation.
- **Distance correlation**: Distance correlation measures both linear and nonlinear association between two random variables or random vectors. This is in contrast to Pearson's correlation, which can only detect linear association between two random variables.
- **Percentage bend correlation**: Introduced by Wilcox (1994), it is based on a down-weight of a specified percentage of marginal observations deviating from the median (by default, 20%).
- **Shepherd's Pi correlation**: Equivalent to a Spearman's rank correlation after outliers removal (by means of bootstrapped mahalanobis distance).
- **Point-Biserial and biserial correlation**: Correlation coefficient used when one variable is continuous and the other is dichotomous (binary). Point-serial is equivalent to a Pearson's correlation, while Biserial should be used when the binary variable is assumed to have an underlying continuity. For example, anxiety level can be measured on a continuous scale, but can be classified dichotomously as high/low.
- **Polychoric correlation**: Correlation between two theorised normally distributed continuous latent variables, from two observed ordinal variables.
- **Tetrachoric correlation**: Special case of the polychoric correlation applicable when both observed variables are dichotomous.

Partial Correlation: **Partial correlations** are estimated as the correlation between two variables after adjusting for the (linear) effect of one or more other variable. The correlation test is then run after having partialized the dataset, independently from it. In other words, it considers partialization as an independent step generating a different dataset, rather than belonging to the same model. This is why some discrepancies are to be expected for the t- and p-values, CIs, BFs etc (but *not* the correlation coefficient) compared to other implementations (e.g., `ppcor`). (The size of these discrepancies depends on the number of covariates partialled-out and the strength of the linear association between all variables.)

Multilevel correlations are a special case of partial correlations where the variable to be adjusted for is a factor and is included as a random effect in a mixed model.

Notes:

- Kendall and Spearman correlations when `bayesian=TRUE`: These are technically Pearson Bayesian correlations of rank transformed data, rather than pure Bayesian rank correlations (which have different priors).

Examples

```
library(correlation)

cor_test(iris, "Sepal.Length", "Sepal.Width")
cor_test(iris, "Sepal.Length", "Sepal.Width", method = "spearman")
cor_test(iris, "Sepal.Length", "Sepal.Width", method = "kendall")
cor_test(iris, "Sepal.Length", "Sepal.Width", method = "biweight")
cor_test(iris, "Sepal.Length", "Sepal.Width", method = "distance")
cor_test(iris, "Sepal.Length", "Sepal.Width", method = "percentage")
cor_test(iris, "Sepal.Length", "Sepal.Width", method = "shepherd")
cor_test(iris, "Sepal.Length", "Sepal.Width", bayesian = TRUE)

# Tetrachoric
data <- iris
data$Sepal.Width_binary <- ifelse(data$Sepal.Width > 3, 1, 0)
data$Petal.Width_binary <- ifelse(data$Petal.Width > 1.2, 1, 0)
cor_test(data, "Sepal.Width_binary", "Petal.Width_binary", method = "tetrachoric")

# Biserial
cor_test(data, "Sepal.Width", "Petal.Width_binary", method = "biserial")

# Polychoric
data$Petal.Width_ordinal <- as.factor(round(data$Petal.Width))
data$Sepal.Length_ordinal <- as.factor(round(data$Sepal.Length))
cor_test(data, "Petal.Width_ordinal", "Sepal.Length_ordinal", method = "polychoric")

# When one variable is continuous, will run 'polyserial' correlation
cor_test(data, "Sepal.Width", "Sepal.Length_ordinal", method = "polychoric")

# Robust (these two are equivalent)
cor_test(iris, "Sepal.Length", "Sepal.Width", method = "pearson", robust = TRUE)
cor_test(iris, "Sepal.Length", "Sepal.Width", method = "spearman", robust = FALSE)

# Partial
cor_test(iris, "Sepal.Length", "Sepal.Width", partial = TRUE)
cor_test(iris, "Sepal.Length", "Sepal.Width", multilevel = TRUE)

cor_test(iris, "Sepal.Length", "Sepal.Width", partial_bayesian = TRUE)
```

`cor_to_ci` *Convert correlation to p-values and CIs*

Description

Get statistics, p-values and confidence intervals (CI) from correlation coefficients.

Usage

```
cor_to_ci(cor, n, ci = 0.95, method = "pearson")

cor_to_p(cor, n, method = "pearson")
```

Arguments

cor	A correlation matrix or coefficient.
n	The sample size (number of observations).
ci	Confidence/Credible Interval level. If "default", then it is set to 0.95 (95% CI).
method	A character string indicating which correlation coefficient is to be used for the test. One of "pearson" (default), "kendall", or "spearman", "biserial", "polychoric", "tetrachoric", "biweight", "distance", "percentage" (for percentage bend correlation) or "shepherd" (for Shepherd's Pi correlation). Setting "auto" will attempt at selecting the most relevant method (polychoric when ordinal factors involved, tetrachoric when dichotomous factors involved, point-biserial if one dichotomous and one continuous and pearson otherwise).

Value

A list containing a p-value and the statistic or the CI bounds.

Examples

```
cor.test(iris$Sepal.Length, iris$Sepal.Width)
cor_to_p(-0.1175698, n = 150)
cor_to_p(cor(iris[1:4]), n = 150)
cor_to_ci(-0.1175698, n = 150)
cor_to_ci(cor(iris[1:4]), n = 150)

cor.test(iris$Sepal.Length, iris$Sepal.Width, method = "spearman")
cor_to_p(-0.1667777, n = 150, method = "spearman")
cor_to_ci(-0.1667777, ci = 0.95, n = 150)

cor.test(iris$Sepal.Length, iris$Sepal.Width, method = "kendall")
cor_to_p(-0.07699679, n = 150, method = "kendall")
```

cor_to_cov *Convert a correlation to covariance*

Description

Convert a correlation to covariance

Usage

```
cor_to_cov(cor, sd = NULL, variance = NULL, tol = .Machine$double.eps^(2/3))
```

Arguments

cor	A correlation matrix, or a partial or a semipartial correlation matrix.
sd, variance	A vector that contains the standard deviations, or the variance, of the variables in the correlation matrix.
tol	Relative tolerance to detect zero singular values.

Value

A covariance matrix.

Examples

```
cor <- cor(iris[1:4])
cov(iris[1:4])

cor_to_cov(cor, sd = sapply(iris[1:4], sd))
cor_to_cov(cor, variance = sapply(iris[1:4], var))
```

cor_to_pcor *Correlation Matrix to (Semi) Partial Correlations*

Description

Convert a correlation matrix to a (semi)partial correlation matrix. Partial correlations are a measure of the correlation between two variables that remains after controlling for (i.e., "partialling" out) all the other relationships. They can be used for graphical Gaussian models, as they represent the direct interactions between two variables, conditioned on all remaining variables. This means that the squared partial correlation between a predictor X1 and a response variable Y can be interpreted as the proportion of (unique) variance accounted for by X1 relative to the residual or unexplained variance of Y that cannot be accounted by the other variables.

Usage

```
cor_to_pcor(cor, tol = .Machine$double.eps^(2/3))

pcor_to_cor(pcor, tol = .Machine$double.eps^(2/3))

cor_to_spcor(cor = NULL, cov = NULL, tol = .Machine$double.eps^(2/3))

spcor_to_cor(
  spcor = NULL,
  cov = NULL,
  semi = FALSE,
  tol = .Machine$double.eps^(2/3)
)
```

Arguments

cor, pcor, spcor	A correlation matrix, or a partial or a semipartial correlation matrix.
tol	Relative tolerance to detect zero singular values.
cov	A covariance matrix (or a vector of the SD of the variables). Required for semi-partial correlations.
semi	Semi-partial correlations.

Details

The semi-partial correlation is similar to the partial correlation statistic. However, it represents (when squared) the proportion of (unique) variance accounted for by the predictor X1, relative to the total variance of Y. Thus, it might be seen as a better indicator of the "practical relevance" of a predictor, because it is scaled to (i.e., relative to) the total variability in the response variable.

Value

The (semi) partial correlation matrix.

Examples

```
cor <- cor(iris[1:4])

# Partialize
cor_to_pcor(cor)
cor_to_spcor(cor, cov = sapply(iris[1:4], sd))

# Inverse
round(pcor_to_cor(cor_to_pcor(cor)) - cor, 2) # Should be 0
```

`distance_mahalanobis` *Mahalanobis distance and confidence interval (CI)*

Description

The Mahalanobis distance (in squared units) measures the distance in multivariate space taking into account the covariance structure of the data. Because a few extreme outliers can skew the covariance estimate, the bootstrapped version is considered as more robust.

Usage

```
distance_mahalanobis(data, ci = 0.95, iterations = 1000, robust = TRUE, ...)
```

Arguments

`data`	A data frame.
`ci`	Confidence/Credible Interval level. If "default", then it is set to 0.95 (95% CI).
`iterations`	The number of draws to simulate/bootstrap (when `robust` is `TRUE`).
`robust`	If `TRUE`, will run a bootstrapped version of the function with i iterations.
`...`	Arguments passed to or from other methods.

Value

Description of the Mahalanobis distance.

References

- Schwarzkopf, D. S., De Haas, B., & Rees, G. (2012). Better ways to improve standards in brain-behavior correlation analysis. Frontiers in human neuroscience, 6, 200.

Examples

```
distance_mahalanobis(iris[, 1:4])
distance_mahalanobis(iris[, 1:4], robust = FALSE)
```

is.cor *Check if matrix ressembles a correlation matrix*

Description

Check if matrix ressembles a correlation matrix

Usage

```
is.cor(x)
```

Arguments

x	A matrix.

Value

TRUE of the matrix is a correlation matrix or FALSE otherwise.

isSquare *Check if Square Matrix*

Description

Check if Square Matrix

Usage

```
isSquare(m)
```

Arguments

m	A matrix.

Value

TRUE of the matrix is square or FALSE otherwise.

matrix_inverse *Matrix Inversion*

Description

Performs a Moore-Penrose generalized inverse (also called the Pseudoinverse).

Usage

```
matrix_inverse(m, tol = .Machine$double.eps^(2/3))
```

Arguments

m	Matrix for which the inverse is required.
tol	Relative tolerance to detect zero singular values.

Value

An inversed matrix.

See Also

pinv from the pracma package

Examples

```
m <- cor(iris[1:4])
matrix_inverse(m)
```

simulate_simpson *Simpson's paradox dataset simulation*

Description

Simpson's paradox, or the Yule-Simpson effect, is a phenomenon in probability and statistics, in which a trend appears in several different groups of data but disappears or reverses when these groups are combined.

Usage

```
simulate_simpson(n = 100, r = 0.5, groups = 3, difference = 1)
```

Arguments

n	The number of observations for each group to be generated.
r	A value or vector corresponding to the desired correlation coefficients.
groups	Number of groups.
difference	Difference between groups.

Value

A dataset.

Examples

```
data <- simulate_simpson(n = 100, groups = 5, r = 0.5)

library(ggplot2)
ggplot(data, aes(x = V1, y = V2)) +
  geom_point(aes(color = Group)) +
  geom_smooth(aes(color = Group), method = "lm") +
  geom_smooth(method = "lm")
```

---

z_fisher *Fisher z-transformation*

Description

The Fisher z-transformation converts the standard Pearson's r to a normally distributed variable z'. It is used to compute confidence intervals to correlations. The z' variable is different from the z-statistic.

Usage

```
z_fisher(r = NULL, z = NULL)
```

Arguments

r, z	The r or the z' value to be converted.

Value

The transformed value.

References

Zar, J.H., (2014). Spearman Rank Correlation: Overview. Wiley StatsRef: Statistics Reference Online. doi:10.1002/9781118445112.stat05964

Examples

```
z_fisher(r = 0.7)
z_fisher(z = 0.867)
```

Index