How can correlation be calculated in R when there are missing values present in the data?

Authored by

stats writer

June 24, 2024

RECOMMENDED CITATION

stats writer (2024). *How can correlation be calculated in R when there are missing values present in the data?*. PSYCHOLOGICAL SCALES. Retrieved from https://scales.arabpsychology.com/?p=151382

ARABPSYCHOLOGY.COM

In R, correlation can be calculated even when there are missing values present in the data. This can be done by using the "cor" function and specifying the desired method for handling missing values, such as "pairwise.complete.obs" or "pairwise.na.ignore". These methods will calculate the correlation between each pair of variables using only the available data points and ignore any missing values. This allows for a more accurate and complete analysis of the relationship between variables, even when some data is missing.

Calculate Correlation in R with Missing Values

You can use the following methods to calculate correlation coefficients in R when one or more variables have missing values:

Method 1: Calculate Correlation Coefficient with Missing Values Present

cor(x, y, use='complete.obs')

Method 2: Calculate Correlation Matrix with Missing Values Present

cor(df, use='pairwise.complete.obs')

The following examples show how to use each method in practice.

Example 1: Calculate Correlation Coefficient with Missing Values

ARABPSYCHOLOGY.COM

Present

Suppose we attempt to use the cor() function to calculate the Pearson correlation coefficient between two variables when missing values are present:

```
#create two variables
x <- c(70, 78, 90, 87, 84, NA, 91, 74, 83, 85)
y <- c(90, NA, 79, 86, 84, 83, 88, 92, 76, 75)

#attempt to calculate correlation coefficient between x and y
cor(x, y)
```

NA

The cor() function returns NA since we didn't specify how to handle missing values.

To avoid this issue, we can use the argument use='complete.obs' so that R knows to only use pairwise observations where both values are present:

```
#create two variables
x <- c(70, 78, 90, 87, 84, NA, 91, 74, 83, 85)
```

ARABPSYCHOLOGY.COM

```
y <- c(90, NA, 79, 86, 84, 83, 88, 92, 76, 75)

#calculate correlation coefficient between x and y
cor(x, y, use='complete.obs')
```

```
-0.4888749
```

The correlation coefficient between the two variables turns out to be -0.488749.

Note that the cor() function only used pairwise combinations where both values were present when calculating the correlation coefficient.

Example 2: Calculate Correlation Matrix with Missing Values Present

Suppose we attempt to use the cor() function to create a for a data frame with three variables when missing values are present:

```
#create data frame with some missing values
df <- data.frame(x=c(70, 78, 90, 87, 84, NA, 91, 74, 83, 85),
y=c(90, NA, 79, 86, 84, 83, 88, 92, 76, 75),
z=c(57, 57, 58, 59, 60, 78, 81, 83, NA, 90))
```

ARABPSYCHOLOGY.COM

```
#attempt to create correlation matrix for variables in data frame
cor(df)

x y z
x 1 NA NA
y NA 1 NA
z NA NA 1
```

To avoid this issue, we can use the argument use='pairwise.complete.obs' so that R knows to only use pairwise observations where both values are present:

```
#create data frame with some missing values
df <- data.frame(x=c(70, 78, 90, 87, 84, NA, 91, 74, 83, 85),
y=c(90, NA, 79, 86, 84, 83, 88, 92, 76, 75),
z=c(57, 57, 58, 59, 60, 78, 81, 83, NA, 90))

#create correlation matrix for variables using only pairwise complete observations
cor(df, use='pairwise.complete.obs')

x y z
```

ARABPSYCHOLOGY.COM

x 1.0000000 -0.4888749 0.1311651

y -0.4888749 1.0000000 -0.1562371

z 0.1311651 -0.1562371 1.0000000

The correlation coefficients for each pairwise combination of variables in the data frame are now shown.

ARABPSYCHOLOGY.COM