How can missing values be imputed in R? Can you provide examples of imputation methods in R?

Authored by

stats writer

April 20, 2024

Missing values are a common occurrence in datasets and can significantly impact the accuracy and reliability of statistical analyses. In R, missing values can be imputed, or filled in, using various methods to ensure complete data for analysis. These methods include mean imputation, median imputation, mode imputation, and multiple imputation. Mean imputation replaces missing values with the mean of the available data for the respective variable. Similarly, median imputation uses the median instead of the mean. Mode imputation replaces missing values with the most frequently occurring value in the dataset. Multiple imputation involves creating multiple imputed datasets and combining them for analysis. These methods can be implemented using built-in functions and packages in R, such as the "mice" package for multiple imputation. Overall, imputation in R allows for the handling of missing values and enables more accurate and comprehensive analysis of datasets.

Impute Missing Values in R (With Examples)

Often you may want to replace missing values in the columns of a data frame in R with the mean or the median of that particular column.

To replace the missing values in a single column, you can use the following syntax:

```
df$col <- mean(df$col, na.rm=TRUE)
```

And to replace the missing values in multiple columns, you can use the following syntax:

```
for(i in 1:ncol(df)) {
df)] <- mean(df, na.rm=TRUE)
}
```

This tutorial explains exactly how to use these functions in practice.

Example 1: Replace Missing Values with Column Means

The following code shows how to replace the missing values in the first column of a data frame with the mean value of the first column:

```
#create data frame
df <- data.frame(var1=c(1, NA, NA, 4, 5),
var2=c(7, 7, 8, 3, 2),
var3=c(3, 3, 6, 6, 8),
var4=c(1, 1, 2, 8, 9))

#replace missing values in first column with mean of first column
df$var1 <- mean(df$var1, na.rm=TRUE)

#view data frame with missing values replaced
df

var1 var2 var3 var4
1 1.000000 7 3 1
2 3.333333 7 3 1
3 3.333333 8 6 2
```

4 4.000000 3 6 8

5 5.000000 2 8 9

The mean value in the first column was 3.333, so the missing values in the first column were replaced with 3.333.

The following code shows how to replace the missing values in each column with the mean of its own column:

```
#create data frame
df <- data.frame(var1=c(1, NA, NA, 4, 5),
var2=c(7, 7, 8, NA, 2),
var3=c(NA, 3, 6, NA, 8),
var4=c(1, 1, 2, 8, 9))

#replace missing values in each column with column means
for(i in 1:ncol(df)) {
df)] <- mean(df, na.rm=TRUE)
}

#view data frame with missing values replaced
df
```

var1 var2 var3 var4

1 1.000000 7 5.666667 1

2 3.333333 7 3.000000 1

3 3.333333 8 6.000000 2

4 4.000000 6 5.666667 8

5 5.000000 2 8.000000 9

Example 2: Replace Missing Values with Column Medians

The following code shows how to replace the missing values in the first column of a data frame with the median value of the first column:

```
#create data frame
df <- data.frame(var1=c(1, NA, NA, 4, 5),
var2=c(7, 7, 8, NA, 2),
var3=c(NA, 3, 6, NA, 8),
var4=c(1, 1, 2, 8, 9))

#replace missing values in first column with median of first column
df$var1 <- median(df$var1, na.rm=TRUE)

#view data frame with missing values replaced
df
```

var1 var2 var3 var4

1 1 7 NA 1

2 4 7 3 1

3 4 8 6 2

4 4 NA NA 8

5 5 2 8 9

The median value in the first column was 4, so the missing values in the first column were replaced with 4.

The following code shows how to replace the missing values in each column with the median of its own column:

```
#create data frame
df <- data.frame(var1=c(1, NA, NA, 4, 5),
var2=c(7, 7, 8, NA, 2),
var3=c(NA, 3, 6, NA, 8),
var4=c(1, 1, 2, 8, 9))

#replace missing values in each column with column medians
for(i in 1:ncol(df)) {
df)] <- median(df, na.rm=TRUE)
```

```
}

#view data frame with missing values replaced
df

var1 var2 var3 var4
1 1 7 6 1
2 4 7 3 1
3 4 8 6 2
4 4 7 6 8
5 5 2 8 9
```

How to Loop Through Column Names in R

How to Calculate the Mean of Multiple Columns in R

How to Sum Specific Columns in R