# What constitutes an influential observation in statistics?

Authored by

**stats writer**

April 28, 2024

An influential observation in statistics refers to a data point that has a significant impact on the results of statistical analysis. This observation holds a high degree of leverage and can greatly affect the overall conclusions and outcomes of a study. Influential observations can arise due to extreme values, outliers, or influential data points that do not follow the expected pattern of the data. It is important to identify and properly handle influential observations in statistics to ensure accurate and reliable results.

## What is an Influential Observation in Statistics?

**In statistics, an influential observation is an observation in a dataset that, when removed, dramatically changes the of a regression model.**

**The most common way to measure the influence of observations is to use Cook's distance, which quantifies how much all of the fitted values in a regression model change when the ith observation is deleted.**

**As a rule of thumb, any observation with a Cook's distance greater than 1 is considered to be an observation with high leverage.**

**The following example shows how to calculate and interpret Cook's distance for a given dataset to detect potential influential observations.**

## Example: Detecting Influential Observations

## Suppose we have the following dataset with 14 values:

| x | y |
|---|---|
| 1 | 23 |
| 2 | 24 |
| 3 | 23 |
| 4 | 19 |
| 5 | 34 |
| 7 | 35 |
| 3 | 36 |
| 2 | 36 |
| 12 | 34 |
| 11 | 32 |
| 15 | 38 |
| 14 | 41 |
| 17 | 42 |
| 22 | 180 |



## Now suppose we fit a . The regression output is shown below:

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
|---|---|---|---|---|
| Intercept | 8.47 | 13.58 | 0.62 | 0.54 |
| x | 4.05 | 1.28 | 3.17 | 0.01 |

## Using statistical software, we can calculate the following values for Cook's distance for each observation:

| x | y | Cook's Distance |
|---|---|---|
| 1 | 23 | 0.014 |
| 2 | 24 | 0.006 |
| 3 | 23 | 0.001 |
| 4 | 19 | 0.002 |
| 5 | 34 | 0.002 |
| 7 | 35 | 0.002 |
| 3 | 36 | 0.019 |
| 2 | 36 | 0.038 |
| 12 | 34 | 0.032 |
| 11 | 32 | 0.023 |
| 15 | 38 | 0.103 |
| 14 | 41 | 0.05 |
| 17 | 42 | 0.202 |
| 22 | 180 | 3.693 |

**Notice that the last observation has a value significantly greater than 1 for Cook's distance, which tells us that it's an influential observation.**

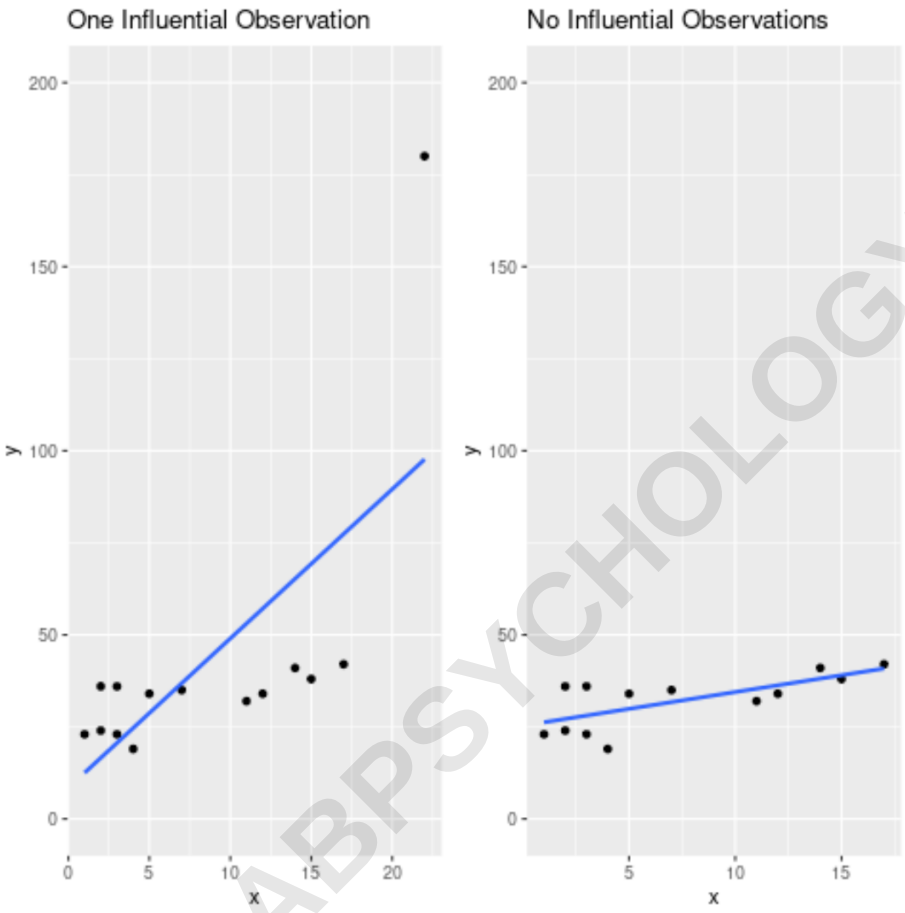
**Suppose we remove this value from the dataset and fit a new simple linear regression model. The output for this model is shown below:**

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
|---|---|---|---|---|
| Intercept | 25.34 | 2.61 | 9.71 | 0.00 |
| x | 0.91 | 0.28 | 3.20 | 0.01 |

**Notice that the regression coefficients for the intercept and x both changed dramatically. This tells us that removing the influential observation from the dataset**

**completely changed the fitted regression model.**

**The following plots show the difference between these two fitted regression equations:**






**Notes**

**It's important to note that Cook's distance should be used as a way to *identify* potentially influential observations. However, just because an observation is influential doesn't necessarily mean that it should be deleted from the dataset.**

**First, you should verify that the observation isn't a result of a data entry error or some other odd occurrence. If it turns out to be a legit value, you can then decide to deal with it in one of the following ways:**

**Delete it from the dataset.Leave it in the dataset.Replace it with an alternative value like the mean or median.**

**Depending on your specific scenario, one of these options may make more sense than the others.**

### How to Calculate Cook's Distance in Practice

**The following tutorials explain how to calculate Cook's distance for a given dataset in Python and R:**