

# How can Cook's Distance be calculated in Python?

Authored by  
**stats writer**

April 23, 2024

## RECOMMENDED CITATION

stats writer (2024). *How can Cook's Distance be calculated in Python?*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=138567>

Cook's Distance is a statistical measure used to identify influential data points in a regression analysis. In Python, this can be calculated by first fitting a linear regression model using the statsmodels library, then accessing the "get\_influence" method to obtain the influence statistics. This method calculates the Cook's Distance for each data point in the model and returns a list of values. Alternatively, the "ols\_outlier\_test" function from the same library can be used to directly obtain the Cook's Distance values for all data points in the model. This allows for easy identification and removal of influential data points in a regression analysis using Python.

## Calculate Cook's Distance in Python

**Cook's distance is used to identify influential in a regression model.**

**The formula for Cook's distance is:**

$$D_i = (r_i^2 / p * MSE) * (h_{ii} / (1 - h_{ii})^2)$$

**where:**

**$r_i$  is the  $i$ th residual  
 $p$  is the number of coefficients in the regression model  
 $MSE$  is the mean squared error  
 $h_{ii}$  is the  $i$ th leverage value**

**Essentially Cook's distance measures how much all of the fitted values in the model change when the  $i$ th observation is deleted.**

**The larger the value for Cook's distance, the more influential a given observation.**

A general rule of thumb is that any observation with a Cook's distance greater than  $4/n$  (where  $n$  = total observations) is considered to be highly influential.

This tutorial provides a step-by-step example of how to calculate Cook's distance for a given regression model in Python.

Step 1: Enter the Data

First, we'll create a small dataset to work with in Python:

```
import pandas as pd

#create dataset
df = pd.DataFrame({'x': ,
'y': })
```

Step 2: Fit the Regression Model

Next, we'll fit a :

```
import statsmodels.api as sm

#define response variable
y = df
```

```
#define explanatory variable
```

```
x = df
```

```
#add constant to predictor variables
```

```
x = sm.add_constant(x)
```

```
#fit linear regression model
```

```
model = sm.OLS(y, x).fit()
```

Step 3: Calculate Cook's Distance

Next, we'll calculate Cook's distance for each observation in the model:

```
#suppress scientific notation
```

```
import numpy as np
```

```
np.set_printoptions(suppress=True)
```

```
#create instance of influence
```

```
influence = model.get_influence()
```

```
#obtain Cook's distance for each observation
```

```
cooks = influence.cooks_distance#display Cook's distances
```

```
print(cooks)
```

```
(array(),  
array())
```

By default, the `cooks_distance()` function displays an array of values for Cook's distance for each observation followed by an array of corresponding p-values.

For example:

```
Cook's distance for observation #1: .368 (p-value:  
.701)Cook's distance for observation #2: .061 (p-value:  
.941)Cook's distance for observation #3: .001 (p-value:  
.999)
```

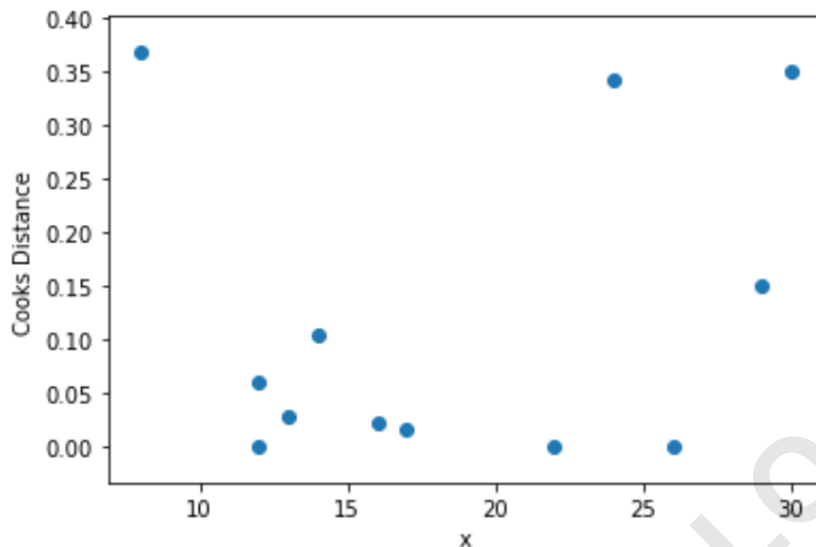
And so on.

Step 4: Visualize Cook's Distances

Lastly, we can create a scatterplot to visualize the values for the predictor variable vs. Cook's distance for each observation:

```
import matplotlib.pyplot as plt  
  
plt.scatter(df.x, cooks)  
plt.xlabel('x')
```

```
plt.ylabel('Cooks Distance')  
plt.show()
```



### Closing Thoughts

It's important to note that Cook's Distance should be used as a way to *identify* potentially influential observations. 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 if it's appropriate to delete it, leave it be, or simply replace it with an alternative value like the

**median.**

ARABPSYCHOLOGY.COM