

How can standardized residuals be calculated in Python?

Authored by
stats writer

April 23, 2024

RECOMMENDED CITATION

stats writer (2024). *How can standardized residuals be calculated in Python?*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=138551>

Standardized residuals are a statistical measure used to assess the deviation of observed data points from the expected values in a regression model. In Python, standardized residuals can be calculated by first obtaining the predicted values using the regression model and then subtracting them from the actual observed values. The resulting values are then divided by the standard error of the regression, which is calculated by taking the square root of the sum of squared residuals divided by the degrees of freedom. This process can be easily implemented in Python using various libraries such as statsmodels or scikit-learn, making it a convenient tool for analyzing and interpreting regression models.

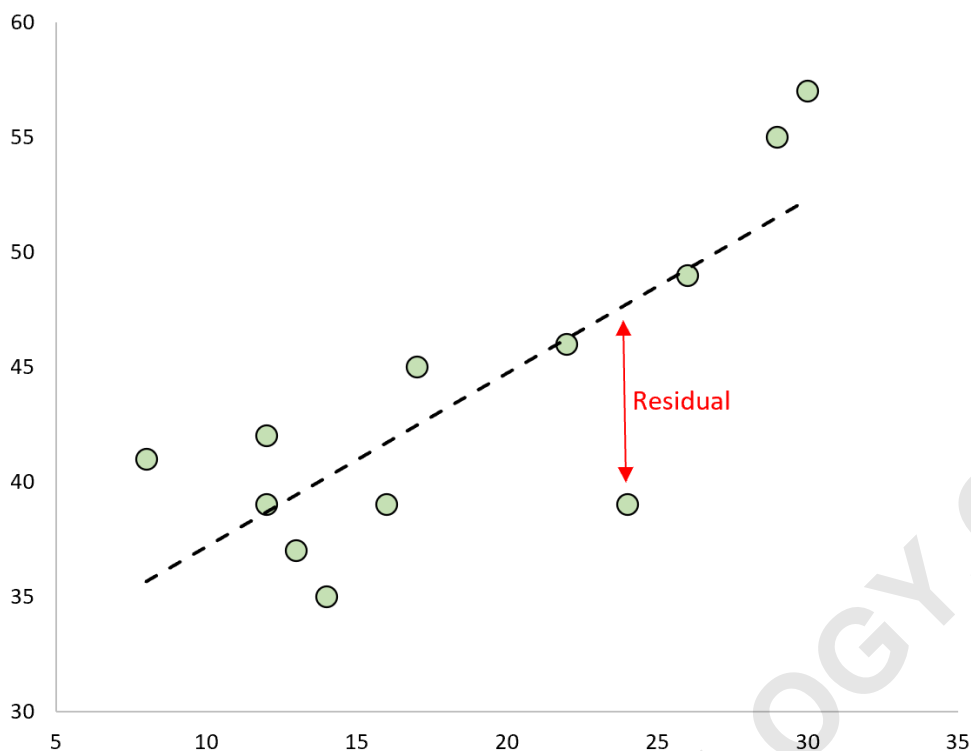
Calculate Standardized Residuals in Python

A residual is the difference between an observed value and a predicted value in a regression model.

It is calculated as:

Residual = Observed value - Predicted value

If we plot the observed values and overlay the fitted regression line, the residuals for each would be the vertical distance between the observation and the regression line:



One type of residual we often use to identify outliers in a regression model is known as a standardized residual.

It is calculated as:

$$r_i = e_i / s(e_i) = e_i / RSE \sqrt{1 - h_{ii}}$$

where:

e_i : The i th residual
 RSE : The residual standard error of the model
 h_{ii} : The leverage of the i th observation

In practice, we often consider any standardized residual

with an absolute value greater than 3 to be an outlier.

This tutorial provides a step-by-step example of how to calculate standardized residuals 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 the Standardized Residuals

Next, we'll calculate the standardized residuals of the model:

```
#create instance of influence  
influence = model.get_influence()  
  
#obtain standardized residuals  
standardized_residuals =  
influence.resid_studentized_internal#display  
standardized residuals  
print(standardized_residuals)
```

From the results we can see that none of the standardized residuals exceed an absolute value of 3. Thus, none of the observations appear to be outliers.

Step 4: Visualize the Standardized Residuals

Lastly, we can create a scatterplot to visualize the values for the predictor variable vs. the standardized residuals:

```
import matplotlib.pyplot as plt
```

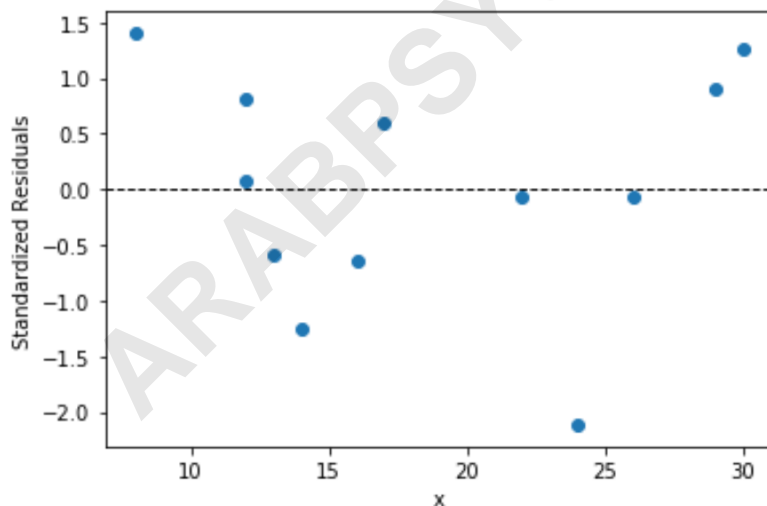
```
plt.scatter(df.x, standardized_residuals)
```

```
plt.xlabel('x')
```

```
plt.ylabel('Standardized Residuals')
```

```
plt.axhline(y=0, color='black', linestyle='--', linewidth=1)
```

```
plt.show()
```



What Are Standardized Residuals?