Calculate Standardized Residuals in R

Authored by

Mohammed looti

November 6, 2025

RECOMMENDED CITATION

Mohammed looti (2025). *Calculate Standardized Residuals in R*. PSYCHOLOGICAL STATISTICS. Retrieved from https://statistics.arabpsychology.com/?p=11459

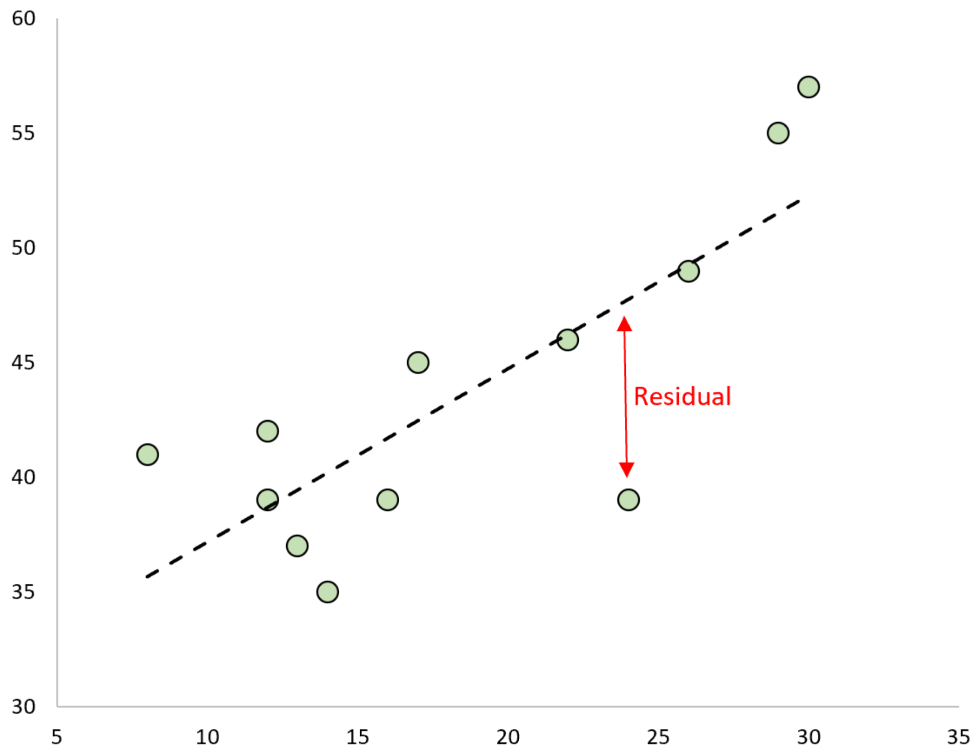
Understanding Residuals and Their Importance

In statistical modeling, particularly regression analysis, a residual represents the difference between an observed data point and the value predicted by the fitted regression model. Essentially, it quantifies the error of prediction for that specific observation.

The basic calculation for a **residual** is straightforward:

Residual = Observed value - Predicted value

When visualizing a simple linear model, the residual corresponds to the vertical distance between the actual data point and the regression line, illustrating how far the model's prediction is from the truth for that observation.



The Role of Standardized Residuals

While raw residuals are useful, they are scale-dependent. To effectively identify unusual observations or potential outliers, we often rely on a refinement known as the standardized residual. Standardizing residuals helps ensure that they follow a common scale, making it easier to compare them across different models or datasets.

The formula for calculating the standardized residual (r_i) for the i^{th} observation is given by:

ri = ei / s(ei) = **ei / RSE√1-hii**

Where these components represent crucial diagnostic metrics:

ei: The raw residual for the ith observation.
RSE: The Residual Standard Error of the overall model.
hii: The leverage of the ith observation, which measures how far the observed value's predictor variables are from the mean of the predictor variables.

A common rule of thumb in statistical practice is to flag any standardized residual whose absolute value exceeds 3 as a potential outlier, warranting further investigation. This tutorial demonstrates the precise steps required to calculate these crucial values using the statistical software R.

Step 1: Preparing the Data in R

The first step in any R analysis is to prepare the dataset. For this example, we will construct a small dataset containing 12 observations, with 'x' as the predictor variable and 'y' as the response variable.

We use the `data.frame()` function to structure our variables. The code snippet below illustrates both the creation and immediate viewing of the resulting data frame:

#create data
data <- data.frame(x=c(8, 12, 12, 13, 14, 16, 17, 22, 24, 26, 29, 30),
y=c(41, 42, 39, 37, 35, 39, 45, 46, 39, 49, 55, 57))

#view data
data

x y
1 8 41
2 12 42
3 12 39
4 13 37
5 14 35
6 16 39
7 17 45
8 22 46
9 24 39
10 26 49

11 29 55
12 30 57

Step 2: Fitting the Simple Linear Regression Model

With the data prepared, the next logical step is to fit the chosen statistical framework--in this case, a simple linear regression model. We utilize R's powerful built-in function, `lm()` (for linear model), specifying that 'y' should be modeled as a function of 'x'.

Running the `summary()` function on the fitted model provides comprehensive output, including coefficient estimates, standard errors, and initial statistics about the raw residuals.

#fit model
model <- lm(y ~ x, data=data)

#view model summary
summary(model)

Call:
lm(formula = y ~ x, data = data)

Residuals:
Min 1Q Median 3Q Max
-8.7578 -2.5161 0.0292 3.3457 5.3268

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.6309 3.6189 8.188 9.6e-06 ***
x 0.7553 0.1821 4.148 0.00199 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.442 on 10 degrees of freedom
Multiple R-squared: 0.6324, Adjusted R-squared: 0.5956
F-statistic: 17.2 on 1 and 10 DF, p-value: 0.001988

This output confirms the successful fitting of the linear regression model. Note the section detailing the distribution of the raw residuals, which will now be transformed into standardized values in the following step.

Step 3: Calculating and Interpreting Standardized Residuals

To calculate the standardized residuals, which account for differing variances among the raw residuals, we use R's dedicated function: `rstandard()`. This function automatically performs the complex calculations involving leverage and the residual standard error.

#calculate the standardized residuals
standard_res <- rstandard(model)

#view the standardized residuals
standard_res

1 2 3 4 5 6
1.40517322 0.81017562 0.07491009 -0.59323342 -1.24820530 -0.64248883
7 8 9 10 11 12
0.59610905 -0.05876884 -2.11711982 -0.06655600 0.91057211 1.26973888

For ease of analysis and diagnosis, it is best practice to append these calculated standardized values back to the original data frame. We use `cbind()` to combine the original predictor and response variables with the new diagnostic metric.

#column bind standardized residuals back to original data frame
final_data <- cbind(data, standard_res)

#view data frame
x y standard_res
1 8 41 1.40517322
2 12 42 0.81017562
3 12 39 0.07491009
4 13 37 -0.59323342
5 14 35 -1.24820530
6 16 39 -0.64248883
7 17 45 0.59610905
8 22 46 -0.05876884
10 26 49 -0.06655600
11 29 55 0.91057211
12 30 57 1.26973888

To efficiently identify observations that might exert undue influence on the model, we sort the results based on the absolute magnitude of the standardized residual, ranking them from largest to smallest. This highlights potential outliers.

#sort standardized residuals descending
final_data

```
x y standard_res
1 8 41 1.40517322
12 30 57 1.26973888
11 29 55 0.91057211
2 12 42 0.81017562
7 17 45 0.59610905
3 12 39 0.07491009
8 22 46 -0.05876884
10 26 49 -0.06655600
4 13 37 -0.59323342
6 16 39 -0.64248883
5 14 35 -1.24820530
9 24 39 -2.11711982
```

Reviewing the sorted list, we observe that the maximum absolute standardized residual is 2.117. Since this value is less than the critical threshold of 3, we conclude that none of the observations in this dataset are classified as extreme outliers according to this standard diagnostic criterion.

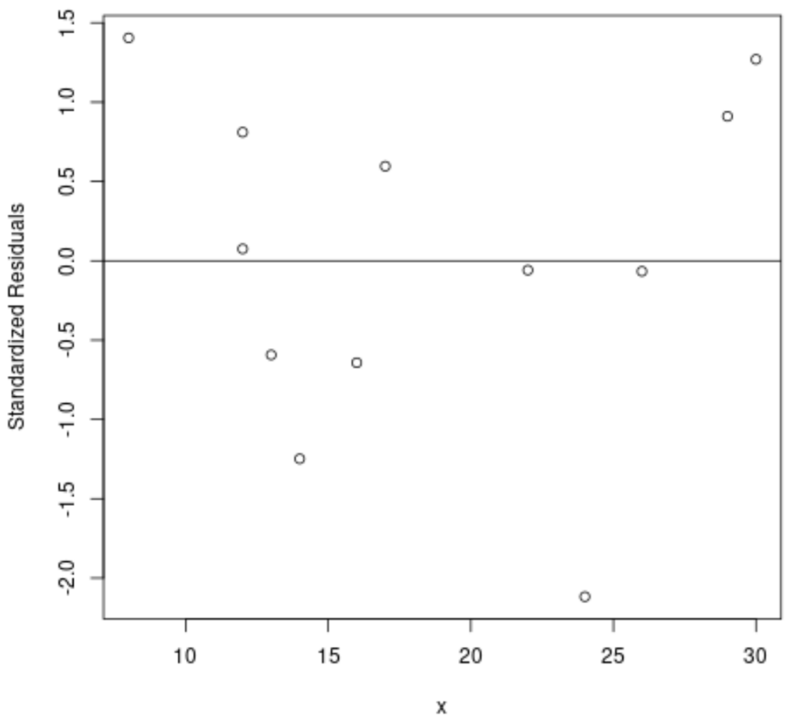
Step 4: Visualizing Residual Diagnostics

A crucial aspect of regression analysis is visually inspecting the standardized residuals to check for linearity assumptions, homoscedasticity, and independence. Plotting the predictor variable ('x') against the standardized residuals allows us to detect patterns that might indicate model deficiencies.

Using the base plotting functions in R, we generate a scatterplot and overlay a horizontal line at y=0, which represents the ideal condition where residuals are centered around zero across all predictor values.

#plot predictor variable vs. standardized residuals
plot(final_data$x, standard_res, ylab='Standardized Residuals', xlab='x')

```
#add horizontal line at 0
abline(0, 0)
```



If the plot exhibits a random scatter of points above and below the zero line, as observed in this visualization, it generally suggests that the model assumptions regarding linearity and homoscedasticity are reasonably satisfied for this dataset.

Additional Resources for Regression Diagnostics

To further deepen your understanding of regression diagnostics and outlier detection in R:

Explore advanced concepts of studentized and standardized residuals.
Learn about other influence measures, such as Cook's Distance and DFFITS.

What Are Standardized Residuals?