

Calculate DFFITS in R

Authored by
stats writer

December 16, 2025

RECOMMENDED CITATION

stats writer (2025). # *Calculate DFFITS in R*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=107600>

In the field of statistical modeling, accurately assessing the stability and reliability of a model is paramount. A key challenge in this assessment is identifying **influential data points**--observations that, when removed, significantly alter the resulting model parameters. These observations can unduly skew the interpretation of the results and lead to erroneous conclusions.

The metric known as **DFFITS**, which stands for "Difference in Fits," is specifically designed to quantify this influence in **linear regression** analysis. DFFITS measures the difference between the predicted value for a specific observation when the model includes that observation versus when the model excludes it. Essentially, it provides a scaled measure of how much the fitted value changes when an observation is dropped.

Understanding DFFITS is crucial for robust model building. High DFFITS values signal potential outliers that exert disproportionate leverage on the **regression coefficients**. While these points are not necessarily "bad" data, they warrant closer investigation to determine if they represent true extremes or data entry errors, thereby guiding the decision on whether they should remain in the dataset or be treated separately.

This comprehensive tutorial will guide you through the process of calculating and visualizing **DFFITS** for individual observations within a regression model using the powerful statistical capabilities of **R**. We will utilize built-in functions to streamline this diagnostic process, ensuring your statistical models are both stable and trustworthy.

Defining DFFITS and Its Statistical Significance

DFFITS is closely related to other influence statistics like Cook's distance, but unlike Cook's distance, DFFITS retains the sign (positive or negative) of the influence, indicating the direction of the change in the fitted value. The formal definition involves the standardized residual and the leverage of the observation, providing a combined measure of deviation and potential impact.

Mathematically, the **DFFITS** value for the i -th observation is calculated as: $DFFITS_i = t_i \sqrt{h_{ii} / (1 - h_{ii})}$, where t_i is the externally standardized residual (or studentized residual) and h_{ii} is the leverage value (the diagonal element of the Hat matrix). The **leverage**, h_{ii} , measures how far an observation's predictor values are from the center of the other predictor values. High leverage points don't necessarily have a large influence, but they have the potential to do so if their actual outcome value deviates significantly from the model prediction.

The primary goal of calculating DFFITS is to identify observations that are both **outliers** (having large residuals) and **high leverage points**. When an observation exhibits both these characteristics, it becomes highly influential, meaning its presence dictates the slope or intercept of the regression line more than other points. Identifying and mitigating the effects of such points is a critical step in achieving a generalizable and reliable regression model.

Building and Examining the Initial Regression Model

To demonstrate the calculation of DFFITS, we will first establish a standard **regression model** using a well-known internal dataset in R. We choose the **mtcars** dataset, which contains data on 32 automobiles, capturing various aspects of their design and performance. Our objective is to model miles per gallon (mpg) as the response variable, based on two key predictors: displacement (disp) and horsepower (hp).

The modeling process begins by loading the dataset and utilizing the standard `lm()` function in **R** to fit the linear relationship. It is always good practice to review the model summary immediately after fitting to understand the initial parameter estimates, statistical significance of predictors, and overall model fit statistics such as R-squared and the F-statistic.

The following R code executes the model fitting and displays the resulting coefficient table, providing the foundation for our influence analysis:

Load the standard 'mtcars' dataset available in R's datasets package

```
data(mtcars)
```

```
# Fit a linear regression model: mpg ~ disp + hp
```

```
model <- lm(mpg~disp+hp, data=mtcars)
```

```
# View the detailed model summary, including coefficients and fit statistics
```

```
summary(model)
```

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 30.735904 1.331566 23.083 < 2e-16 ***

disp -0.030346 0.007405 -4.098 0.000306 ***

hp -0.024840 0.013385 -1.856 0.073679 .

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

Residual standard error: 3.127 on 29 degrees of freedom

Multiple R-squared: 0.7482, Adjusted R-squared: 0.7309

F-statistic: 43.09 on 2 and 29 DF, p-value: 2.062e-09

Calculating DFFITS for Each Observation

Once the **linear regression** model is fitted, calculating the DFFITS values for every observation is straightforward in R. We can use the generic `dffits()` function, which is available in base R when

applied to an `lm` object. This function takes the fitted model object as its sole argument and automatically computes the standardized measure of influence for each data point.

The output of the `dffits()` function is a numeric vector, where each element corresponds to an observation in the original dataset (in this case, named after the specific car model). For easier handling and inspection, especially when combining it with the original data or sorting, it is often useful to convert this vector into a data frame.

The following script calculates the DFFITS values and then displays the output, allowing us to inspect the raw influence scores associated with each car in the `mtcars` dataset. Note the variation in influence, ranging from small negative values to larger positive values, indicating differing degrees and directions of impact on the model fit.

Calculate DFFITS for each observation in the fitted model.

```
dffits <- as.data.frame(dffits(model))
```

```
# Display the DFFITS values, indexed by the row names (car models)
```

```
dffits
```

```
dffits(model)
```

```
Mazda RX4 -0.14633456
```

```
Mazda RX4 Wag -0.14633456
```

```
Datsun 710 -0.19956440
```

```
Hornet 4 Drive 0.11540062
```

```
Hornet Sportabout 0.32140303
```

```
Valiant -0.26586716
```

```
Duster 360 0.06282342
```

```
Merc 240D -0.03521572
```

```
Merc 230 -0.09780612
```

```
Merc 280 -0.22680622
```

```
Merc 280C -0.32763355
```

```
Merc 450SE -0.09682952
```

```
Merc 450SL -0.03841129
```

```
Merc 450SLC -0.17618948
```

```
Cadillac Fleetwood -0.15860270
```

```
Lincoln Continental -0.15567627
```

```
Chrysler Imperial 0.39098449
```

```
Fiat 128 0.60265798
```

```
Honda Civic 0.35544919
```

```
Toyota Corolla 0.78230167
```

```
Toyota Corona -0.25804885
```

Dodge Challenger -0.16674639
AMC Javelin -0.20965432
Camaro Z28 -0.08062828
Pontiac Firebird 0.67858692
Fiat X1-9 0.05951528
Porsche 914-2 0.09453310
Lotus Europa 0.55650363
Ford Pantera L 0.31169050
Ferrari Dino -0.29539098
Maserati Bora 0.76464932
Volvo 142E -0.24266054

Determining the Critical Influence Threshold

While raw DFFITS values are informative, we need a standard benchmark to determine which observations qualify as truly influential. A commonly accepted rule of thumb for identifying potentially problematic **influential data points** is comparing the absolute DFFITS value against a specific threshold. If an observation's DFFITS score exceeds this threshold, it merits closer examination.

The conventional threshold for DFFITS is calculated using the formula: $2\sqrt{p/n}$, where p represents the number of predictor variables (or degrees of freedom for the model parameters, excluding the intercept) and n represents the total number of observations in the dataset. This threshold provides a context-specific benchmark based on the size and complexity of the model being evaluated.

In our `mtcars` example, we have two predictor variables (`disp` and `hp`), meaning $p=2$, and a total of 32 observations, so $n=32$. We use **R** to calculate this threshold precisely, which helps us move from subjective inspection to objective assessment of influence.

Find the number of predictor variables (p) in the model. This is the length of coefficients minus the intercept.

```
p <- length(model$coefficients)-1
```

Find the total number of observations (n) in the dataset.

```
n <- nrow(mtcars)
```

Calculate the standard DFFITS threshold value: $2 * \sqrt{p/n}$

```
thresh <- 2*sqrt(p/n)
```

```
thresh
```

0.5

Sorting and Analyzing Influential Data Points

With the calculated DFFITS values and the established threshold of 0.5, the next logical step is to systematically identify which observations exceed this critical boundary. Sorting the DFFITS values in descending order allows us to quickly view the most impactful observations at the top of the list.

The observations that exceed the threshold warrant detailed scrutiny. This investigation might involve examining the raw data for potential typos, verifying the measurement process for those specific points, or analyzing if the observation represents a true but rare event that might require specialized modeling techniques (such as robust **linear regression**). If the influential point is deemed valid and representative, we might choose to keep it; if it's due to error or represents an anomaly unrelated to the population of interest, removal or transformation might be necessary.

Running the sort command in R clearly highlights the potential issues:

Sort observations by DFFITS score in descending order to identify the most influential points.

dffits),]

```
dffits(model)
Toyota Corolla 0.78230167
Maserati Bora 0.76464932
Pontiac Firebird 0.67858692
Fiat 128 0.60265798
Lotus Europa 0.55650363
Chrysler Imperial 0.39098449
Honda Civic 0.35544919
Hornet Sportabout 0.32140303
Ford Pantera L 0.31169050
Hornet 4 Drive 0.11540062
Porsche 914-2 0.09453310
Duster 360 0.06282342
Fiat X1-9 0.05951528
Merc 240D -0.03521572
Merc 450SL -0.03841129
Camaro Z28 -0.08062828
Merc 450SE -0.09682952
Merc 230 -0.09780612
```

Mazda RX4 -0.14633456
Mazda RX4 Wag -0.14633456
Lincoln Continental -0.15567627
Cadillac Fleetwood -0.15860270
Dodge Challenger -0.16674639
Merc 450SLC -0.17618948
Datsun 710 -0.19956440
AMC Javelin -0.20965432
Merc 280 -0.22680622
Volvo 142E -0.24266054
Toyota Corona -0.25804885
Valiant -0.26586716
Ferrari Dino -0.29539098
Merc 280C -0.32763355

The output confirms that five specific car models--Toyota Corolla, Maserati Bora, Pontiac Firebird, Fiat 128, and Lotus Europa--all possess DFFITS values significantly greater than the 0.5 threshold. These are the primary **influential data points** that demand immediate attention from the analyst. Their inclusion or exclusion will materially affect the estimated relationship between engine characteristics and fuel efficiency.

Visualizing DFFITS and the Influence Threshold

While numerical sorting provides precision, a graphical representation of the DFFITS values offers immediate insight into the distribution of influence across the dataset. Visualization is a powerful diagnostic tool that can quickly highlight observations deviating significantly from the norm.

We utilize the standard `plot()` function in R, applying it directly to the `dffits(model)` object. Specifying `type = 'h'` generates a spike plot, which is highly effective for displaying individual observation metrics. The X-axis represents the observation index (row number), and the Y-axis represents the calculated DFFITS score.

To make the identification of influential points unambiguous, we overlay horizontal lines onto the plot corresponding to the absolute threshold values (+0.5 and -0.5). Any vertical line (spike) that extends beyond these horizontal dashed lines indicates an observation whose removal would significantly impact the fit of the **regression model**.

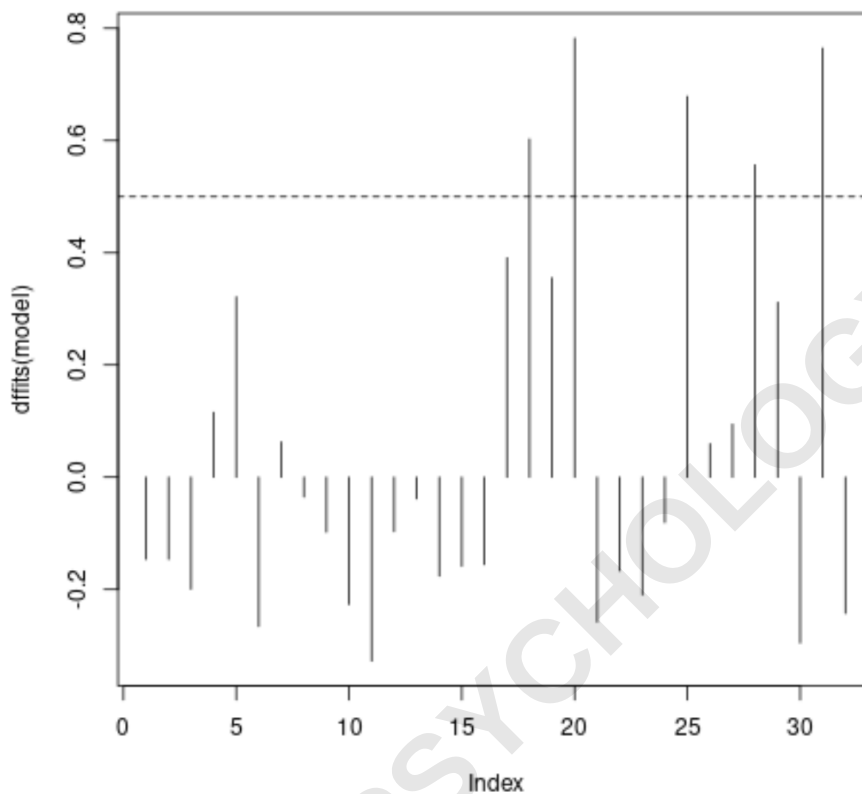
Generate a spike plot of DFFITS values for each observation
plot(dffits(model), type = 'h')

```
# Add a horizontal dashed line indicating the positive threshold
```

```
abline(h = thresh, lty = 2)
```

```
# Add a horizontal dashed line indicating the negative threshold
```

```
abline(h = -thresh, lty = 2)
```



As clearly demonstrated by the plot, several spikes rise above the upper dashed line (the positive threshold). These correspond precisely to the five observations identified numerically. The **DFFITS** visualization confirms that these observations possess high leverage and high residual values, making them extremely critical in determining the estimated regression line.

The x-axis displays the index of each observation in the dataset and the y-value displays the corresponding DFFITS value for each observation.

Interpreting DFFITS Results and Mitigating Influence

A large positive **DFFITS** value implies that removing the observation would cause the fitted value to decrease, suggesting the observation pulled the regression line upwards. Conversely, a large negative DFFITS value implies that removing the observation would cause the fitted value to increase, meaning the observation pulled the regression line downwards. The directionality of

DFFITS provides richer diagnostic information than scalar measures of influence.

Upon identifying these influential points (Toyota Corolla, Maserati Bora, etc.), the next step in robust statistical analysis involves determining the appropriate course of action. Possible interventions include:

Data Verification: Re-examining the raw data source for transcription errors or measurement inaccuracies specific to the influential point.

Model Re-specification: If the point is valid but influential due to complex relationships, it might suggest that the current **linear regression** model is misspecified, perhaps requiring interaction terms, polynomial terms, or a transformation of variables.

Robust Methods: Employing robust regression techniques (e.g., M-estimation) that automatically downweight the influence of outliers and high leverage points, offering a more stable estimate of the population parameters.

Exclusion and Reporting: If the observation is confirmed to be a legitimate outlier representing a population not covered by the rest of the sample, it may be justifiable to exclude it, provided this exclusion is clearly documented and justified in the final report.

By meticulously calculating and visualizing DFFITS, data analysts can move beyond simple model fitting and ensure the resulting statistical inferences are not dictated by a small subset of observations, leading to more generalized and scientifically sound conclusions. This rigorous diagnostic process is fundamental to producing high-quality predictive and explanatory models in **R**.

Conclusion: Enhancing Model Reliability through Influence Diagnostics

The ability to calculate and interpret diagnostic metrics like **DFFITS** is essential for any practitioner working with **regression models**. By quantifying the stability of the model coefficients when individual observations are removed, DFFITS provides a clear warning sign regarding potential issues stemming from **influential data points**.

As demonstrated using the `mtcars` dataset in R, identifying these high-influence observations is not an endpoint, but the beginning of a deeper statistical inquiry. Whether the solution involves correcting errors, utilizing robust regression, or adjusting the model specification, the DFFITS measure ensures that analysts are aware of which data points are exerting the most control over the final model estimates.

Mastering these influence diagnostics ensures that your models are not only statistically significant but also robust against undue leverage, providing a stable foundation for predictive modeling and inference.