

How to Calculate DFBETAS in R?

Authored by
stats writer

December 16, 2025

RECOMMENDED CITATION

stats writer (2025). *How to Calculate DFBETAS in R?*. PSYCHOLOGICAL SCALES.
Retrieved from <https://scales.arabpsychology.com/?p=107597>

In the realm of R programming for statistical analysis, understanding the influence of individual data points on a model is paramount for ensuring robustness and validity. When performing linear regression, we often encounter observations that exert disproportionate influence on the parameter estimates. Identifying and quantifying this influence is essential for high-quality data modeling.

The metric known as DFBETAS (Difference in Beta) serves as a critical diagnostic tool for assessing such influence. It specifically measures the standardized change in each coefficient estimate that results from removing a single observation from the dataset. A high DFBETAS value indicates that the corresponding observation is highly influential on that specific predictor's coefficient.

This comprehensive guide details the process of calculating, interpreting, and visualizing DFBETAS within the R environment. We will utilize built-in R functions and standard practices to pinpoint potentially problematic data points in a regression model.

Understanding DFBETAS and Influence Diagnostics

In statistics, particularly in regression analysis, it is vital to know how influential individual observations are on the overall model fitting. Highly influential observations can skew the results, leading to inaccurate coefficient estimates and potentially misleading conclusions about the relationship between variables. These data points often deserve closer scrutiny, as they might represent errors, outliers, or simply unique, high-leverage instances.

The DFBETAS metric addresses this by providing a standardized measure of change. For every observation, and for every predictor variable (including the intercept), DFBETAS calculates how much the regression coefficient changes when that specific observation is deleted. By standardizing this difference, we can compare the influence across different coefficients and models, enabling a systematic approach to identifying influential observations.

If an observation has a large positive DFBETAS for a certain predictor, it means that removing that observation causes the corresponding coefficient to decrease significantly. Conversely, a large negative DFBETAS implies that removing the observation causes the coefficient to increase. Understanding these diagnostics is a foundational requirement for rigorous statistical modeling.

Step 1: Setting Up the Model in R

To demonstrate the calculation of DFBETAS, we first need to establish a working linear regression model. We will use the readily available `mtcars` dataset built into R, which contains information about various automobile characteristics. Our objective is to predict miles per gallon (mpg) based on displacement (disp) and horsepower (hp). The `lm()` function is used to fit the model, and `summary()` provides an initial overview of the results, including the standard coefficient estimates.

The following code block executes the model fitting and displays the standard statistical summary, confirming the relationships between the predictors and the outcome variable before diagnostics are performed.

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

#view model summary
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
```

Step 2: Calculating DFBETAS for Each Observation

Once the model is built, we can proceed to calculate the DFBETAS values for every data point using the dedicated `dfbetas()` function available in base R. This function takes the fitted linear model object (`model`) as its argument and returns a matrix where rows correspond to observations (cars) and columns correspond to the coefficients (intercept, disp, hp).

We convert the resulting matrix into a data frame for easier manipulation and inspection using the `as.data.frame()` function. Inspecting this output allows us to see the precise numerical impact each observation has on the estimation of the intercept, displacement, and horsepower coefficients.

```
#calculate DFBETAS for each observation in the model
dfbetas <- as.data.frame(dfbetas(model))

#display DFBETAS for each observation
dfbetas

(Intercept) disp hp
```

Mazda RX4 -0.1174171253 0.030760632 1.748143e-02
 Mazda RX4 Wag -0.1174171253 0.030760632 1.748143e-02
 Datsun 710 -0.1694989349 0.086630144 -3.332781e-05
 Hornet 4 Drive 0.0577309674 0.078971334 -8.705488e-02
 Hornet Sportabout -0.0204333878 0.237526523 -1.366155e-01
 Valiant -0.1711908285 -0.139135639 1.829038e-01
 Duster 360 -0.0312338677 -0.005356209 3.581378e-02
 Merc 240D -0.0312259577 -0.010409922 2.433256e-02
 Merc 230 -0.0865872595 0.016428917 2.287867e-02
 Merc 280 -0.1560683502 0.078667906 -1.911180e-02
 Merc 280C -0.2254489597 0.113639937 -2.760800e-02
 Merc 450SE 0.0022844093 0.002966155 -2.855985e-02
 Merc 450SL 0.0009062022 0.001176644 -1.132941e-02
 Merc 450SLC 0.0041566755 0.005397169 -5.196706e-02
 Cadillac Fleetwood 0.0388832216 -0.134511133 7.277283e-02
 Lincoln Continental 0.0483781688 -0.121146607 5.326220e-02
 Chrysler Imperial -0.1645266331 0.236634429 -3.917771e-02
 Fiat 128 0.5720358325 -0.181104179 -1.265475e-01
 Honda Civic 0.3490872162 -0.053660545 -1.326422e-01
 Toyota Corolla 0.7367058819 -0.268512348 -1.342384e-01
 Toyota Corona -0.2181110386 0.101336902 5.945352e-03
 Dodge Challenger -0.0270169005 -0.123610713 9.441241e-02
 AMC Javelin -0.0406785103 -0.141711468 1.074514e-01
 Camaro Z28 0.0390139262 0.012846225 -5.031588e-02
 Pontiac Firebird -0.0549059340 0.574544346 -3.689584e-01
 Fiat X1-9 0.0565157245 -0.017751582 -1.262221e-02
 Porsche 914-2 0.0839169111 -0.028670987 -1.240452e-02
 Lotus Europa 0.3444562478 -0.402678927 2.135224e-01
 Ford Pantera L -0.1598854695 -0.094184733 2.320845e-01
 Ferrari Dino -0.0343997122 0.248642444 -2.344154e-01
 Maserati Bora -0.3436265545 -0.511285637 7.319066e-01
 Volvo 142E -0.1784974091 0.132692956 -4.433915e-02

Each row in the output represents one specific car. The values indicate the change in the corresponding coefficient estimate when that car's data is temporarily excluded from the model fitting process. For instance, the value of 0.7367 for the Toyota Corolla under the (Intercept) column suggests that if this observation were removed, the intercept value would decrease by 0.7367 units relative to its standard error.

Step 3: Establishing the Influence Threshold

Interpreting the raw DFBETAS values requires a comparison against a generally accepted threshold. Statisticians often use a rule of thumb to determine when an observation is considered an influential observation. The common threshold applied is $2/\sqrt{n}$, where n represents the total number of observations in the dataset.

This threshold provides a standardized cutoff: if the absolute value of any DFBETAS measure exceeds this value, that observation is flagged as potentially having undue influence on the corresponding coefficient. If multiple observations surpass this threshold, they warrant mandatory investigation. For the `mtcars` dataset, which has 32 observations ($n=32$), the threshold calculation is straightforward in R.

```
#find number of observations
```

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

```
#calculate DFBETAS threshold value
```

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

```
thresh
```

```
0.3535534
```

Thus, any DFBETAS value greater than 0.3535534 in absolute terms indicates a strong influence on the respective coefficient estimate. This numerical boundary helps transition the analysis from raw values to actionable insights regarding data quality and model stability.

Step 4: Visualizing and Interpreting DFBETAS

While examining the numerical matrix is useful, visualizing the DFBETAS values relative to the calculated threshold provides a clearer and more immediate understanding of the distribution of influence. We can generate diagnostic plots for each predictor using R's plotting capabilities.

The following code sets up a plotting region with two rows and one column (`par(mfrow=c(2,1))`) to display the diagnostic plots for the `disp` and `hp` coefficients side-by-side (or one above the other). Horizontal dashed lines are added to represent the positive and negative influence thresholds, making it easy to spot observations that cross the critical boundary.

```
#specify 2 rows and 1 column in plotting region
```

```
par(mfrow=c(2,1))
```

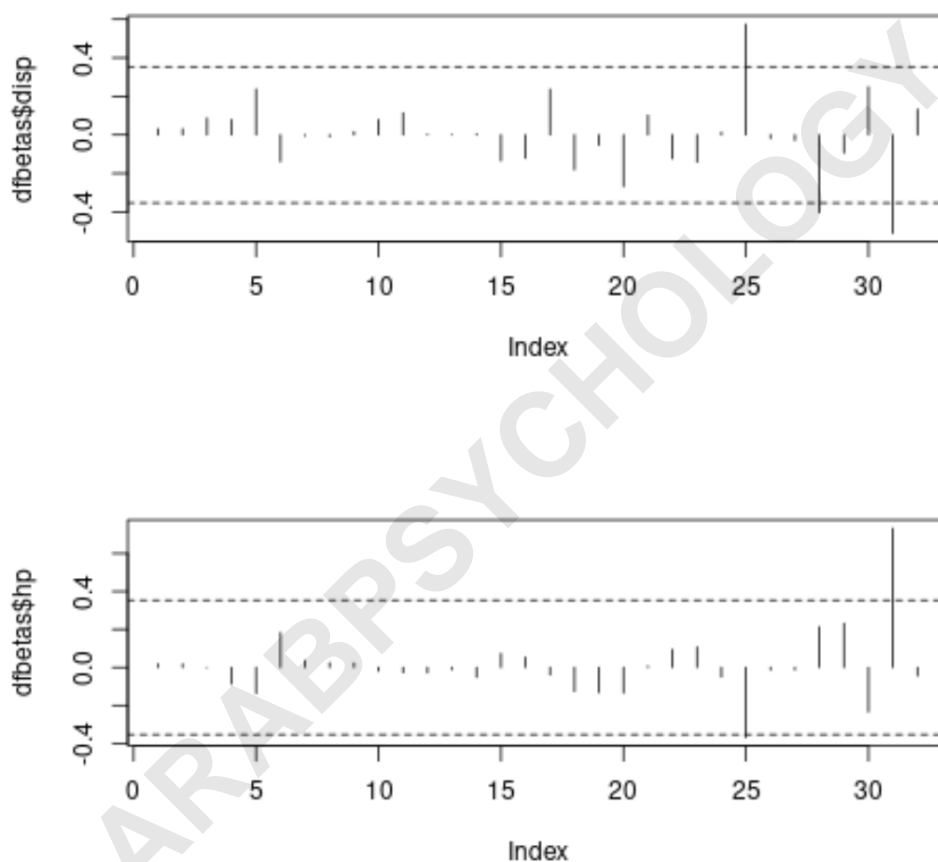
```
#plot DFBETAS for disp with threshold lines
```

```
plot(dfbetas$disp, type='h')
abline(h = thresh, lty = 2)
abline(h = -thresh, lty = 2)
```

```
#plot DFBETAS for hp with threshold lines
```

```
plot(dfbetas$hp, type='h')
abline(h = thresh, lty = 2)
abline(h = -thresh, lty = 2)
```

The resulting image visually captures the influence profile:



In these plots, the x-axis represents the index of the observation (from 1 to 32), and the y-axis shows the magnitude of the DFBETAS value. Observations whose vertical lines extend beyond the dashed threshold lines (both positive and negative) are deemed highly influential observations on the respective coefficient.

Interpreting the Visualization Results

Analyzing the generated plots reveals crucial information about data leverage. In the top plot

(Disp), we observe several bars extending beyond the 0.3535534 threshold. Specifically, three observations exceed this absolute value, suggesting they significantly alter the slope estimated for the displacement variable. Similarly, the bottom plot (HP) shows two distinct observations that cross the influence threshold for the horsepower coefficient.

These specific data points require immediate follow-up. They are not necessarily errors, but their removal would substantially change the model parameters. The investigation might involve:

Verifying data entry accuracy for those specific records.

Examining if these data points represent a sub-population or a specific condition not adequately captured by the current model structure.

Considering robust regression methods that are less sensitive to influential observations.

The calculated DFBETAS values provide the precise evidence needed to justify these deeper investigations, contributing to a more reliable and defensible linear regression model.

How to Calculate DFFITS in R