# Residual Standard Deviation/Error: Guide for Beginners

Regression Analysis

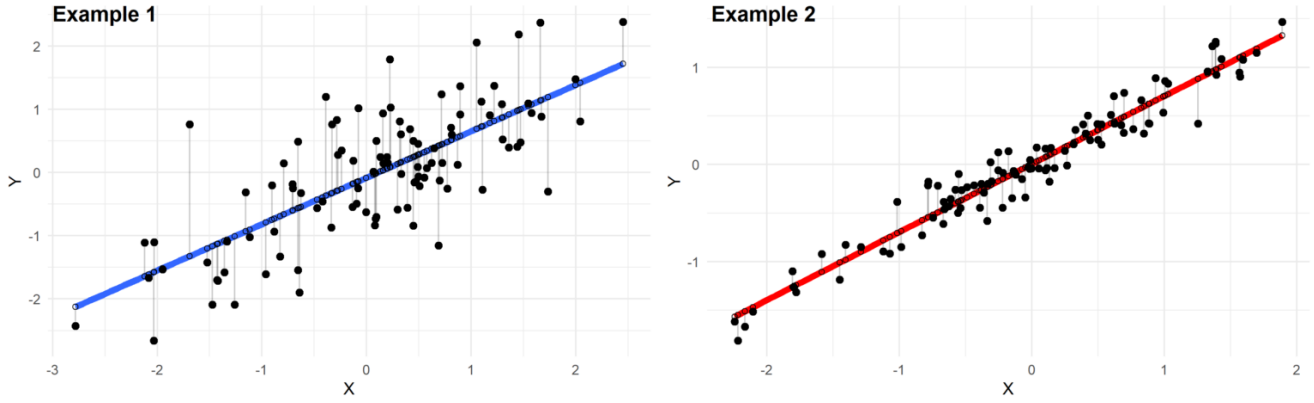
The **residual standard deviation** (or **residual standard error**) is a measure used to assess how well a linear regression model fits the data. (The other measure to assess this goodness of fit is $R^2$).

But before we discuss the residual standard deviation, let's try to assess the goodness of fit graphically.

Consider the following linear regression model:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Plotted below are examples of 2 of these regression lines modeling 2 different datasets:




Just by looking at these plots we can say that the linear regression model in "example 2" fits the data better than that of "example 1".

This is because in "example 2" the points are closer to the regression line. Therefore, using a linear regression model to approximate the true values of these points will yield smaller errors than "example 1".

In the plots above, the gray vertical lines represent the error terms — the

difference between the model and the true value of Y.

Mathematically, the error of the $i^{th}$ point on the x-axis is given by the equation: $(Y_i - \hat{Y}_i)$, which is the difference between the true value of Y ($Y_i$) and the value predicted by the linear model ($\hat{Y}_i$) — this difference determines the length of the gray vertical lines in the plots above.

Now that we developed a basic intuition, next we will try to come up with a statistic that quantifies this goodness of fit.

## Residual standard deviation vs residual standard error vs RMSE

The simplest way to quantify how far the data points are from the regression line, is to calculate the average distance from this line:

$$Average\ distance = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)}{n}$$

Where *n* is the sample size.

But, because some of the distances are positive and some are negative (certain points are above the regression line and others are below it), these distances will cancel each other out — meaning that the average distance will be biased low.

In order to remedy this situation, one solution is to take the square of this distance (which will always be a positive number), calculate the sum of these squared distances for all data points, and then take the square root of this sum to obtain the Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}}$$

We can take this equation one step further:

Instead of dividing by the sample size *n*, we can divide by the degrees of freedom *df* to obtain an unbiased estimation of the standard deviation of the error term ε. (If you're having trouble with this idea, I recommend these 4 videos from Khan Academy which provide a simple explanation mainly through simulations instead of math equations).

The quantity obtained is sometimes called the residual standard **deviation** (as referred to it in the textbook *Data Analysis Using Regression and Multilevel Hierarchical Models* by Andrew Gelman and Jennifer Hill). Other textbooks refer to it as the residual standard **error** (for example *An Introduction to Statistical Learning* by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani).

In the statistical programming language R, calling the function *summary* on the linear model will calculate it automatically.

$$Residual\ standard\ error = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{df}}$$

The degrees of freedom *df* is equal to the sample size minus the number of parameters we're trying to estimate.

For example, if we're estimating 2 parameters $\beta_0$ and $\beta_1$ as in:

$Y = \beta_0 + \beta_1 X + \varepsilon$

Then, df = n – 2

Otherwise if we're estimating 3 parameters, as in:

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

Then, df = n – 3

And so on...

Now that we have a statistic that measures the goodness of fit of a linear model, next we will discuss how to interpret it in practice.

---