

[Guide] Residual Properties Guide

The following document contains the explanations for the properties of residuals that we have claimed thus far.

You do not need to know why these properties are true! These explanations are not in scope.

For some of these properties they take a lot of higher level math to prove that you do not need to know!

Author: Yanay Rosen

We've told you that residuals have some special properties, here are some explanations of why those properties are true.

These properties are in scope but the proofs are not, they are just fun :)

The most intuitive proof is the proof in part 2 for why the residuals have no trend- highly recommended!

The rest of the proofs involve a bit of math, in order from least math to most math:

First proof in Part 1 of why residuals have mean of zero, proof in Part 3 of why the SD of residuals is equal to the RMSE, second proof in part 1 of why residuals have mean of zero.

0) What is a residual?

A residual is an error made by the line of best fit. The line of best fit is the line that minimizes mean squared error (also minimizes root mean squared error!). An error is equal to the actual value minus the predicted value.

1) The mean of residuals is zero

Proving that the average of residuals is zero is simple-- the line of best fit is the line that minimizes mean squared error (also minimizes root mean squared error!). Since residuals are errors, we can get the mean squared error in the following manner:

$$\text{mse}_{\text{original}} = \text{mean}(\text{residuals}^2)$$

Let's start with the assumption that the mean of the residuals is zero. We will show that any variations from this assumption all have outcomes that support our claim.

If we add some bias term b to the residuals, such as shifting them up by 1 or down by 5 for example, let's see what happens to the mse.

$$\text{mse}_{\text{new}} = \text{mean}((\text{residuals} + b)^2)$$

$$\text{mse}_{\text{new}} = \text{mean}(\text{residuals}^2 + 2 * \text{residuals} * b + b^2)$$

Because of some special properties of the mean we can rewrite that as:

$$\text{mse}_{\text{new}} = \text{mean}(\text{residuals}^2) + \text{mean}(2 * \text{residuals} * b) + \text{mean}(b^2)$$

$\text{mean}(b^2) = b^2$ since b is a constant and the mean of a constant is just that constant. Note that since we are squaring this value it is positive no matter what the value of b is!

$$\text{mean}(\text{residuals}^2) = \text{mse}_{\text{original}} \text{ from before.}$$

So what is $\text{mean}(2 * \text{residuals} * b)$ equal to? Well it turns out we can rewrite that as $2 * b * \text{mean}(\text{residuals}) = 2 * b * 0 = 0$.

Therefore, no matter what value we set the mean of our hypothetical residuals to, the MSE of best fit line will be equal to:

$$\text{mean}(\text{residuals}^2) + b^2$$

Remember that our goal in finding the line of best fit is to minimize the MSE. So, we want to make the value above as small as possible. The smallest this value will be is when $b = 0$, thus the average of the residuals must be 0!

Here's another more mathematical proof of this property, based on this [response](#).

To make this proof easier to read we are going to use some fancy math terminology.

The mean of x can be written as \bar{x} , read as "x bar".

The standard deviation of x can be written as σ_x read as "SD x" or "sigma x".

We can write the prediction for y_i as \hat{y}_i , read as "y hat sub i".

For this proof instead of writing "residuals" we will use the variable e .

You read $\sum_{i=1}^n (x_i)$ as "the sum from i equals 1 to i equals n of the ith element of x"

Note that $\frac{1}{n} \sum_{i=1}^n (x_i)$ represents the mean of x .

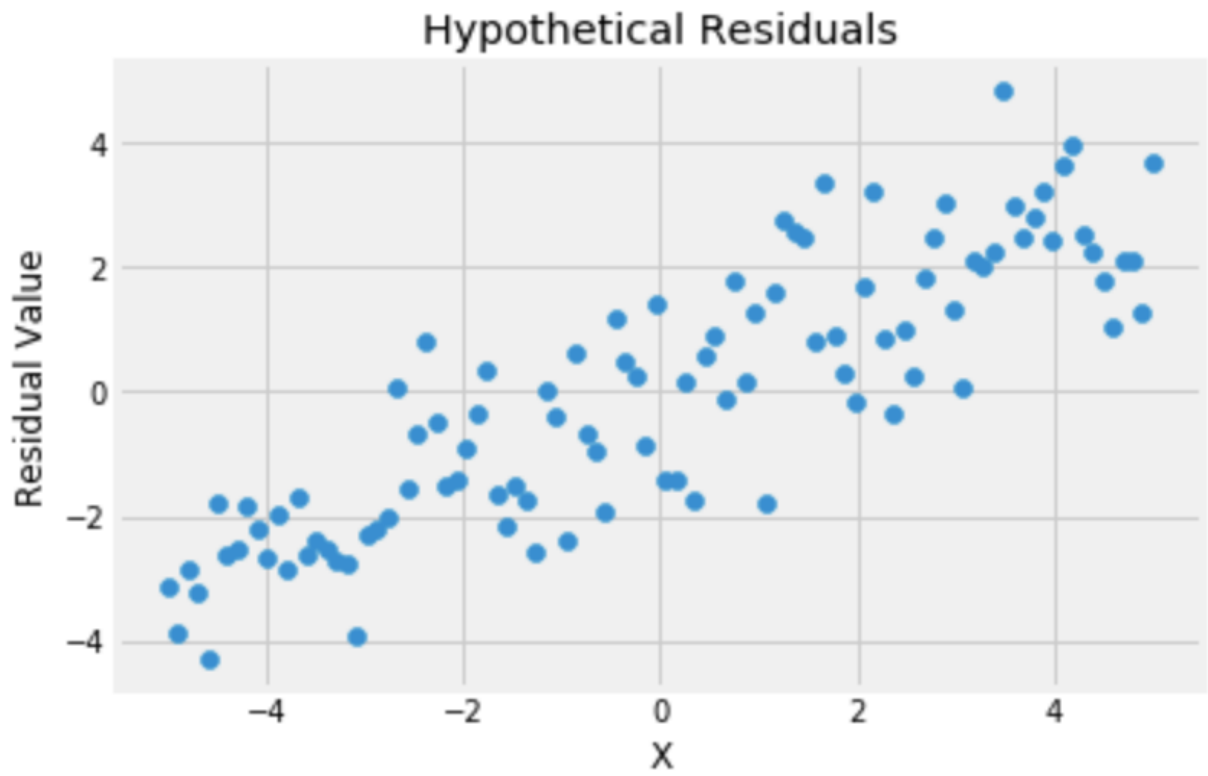
Math	Explanation
$\bar{e} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)$	a residual is the actual value of y minus the predicted value, definition of the mean
$\hat{y}_i = r * \frac{\sigma_y}{\sigma_x} * x_i + \bar{y} - (r * \frac{\sigma_y}{\sigma_x} * \bar{x})$	this is the formula for the prediction of y based on x
$\hat{y}_i = (r * \frac{\sigma_y}{\sigma_x}) * (x_i - \bar{x}) + \bar{y}$	we can rewrite that formula and group the terms with the slope in common.
$\bar{e} = \frac{1}{n} \sum_{i=1}^n (y_i - ((r * \frac{\sigma_y}{\sigma_x}) * (x_i - \bar{x}) + \bar{y}))$	now let's plug that value back into the formula for the mean of the residuals.
$\bar{e} = \frac{1}{n} \sum_{i=1}^n (y_i) - \frac{1}{n} \sum_{i=1}^n ((r * \frac{\sigma_y}{\sigma_x}) * (x_i - \bar{x}) + \bar{y})$	we can break apart the sum into the sum of the actual values of y minus the sum of the predicted values of y.
$\bar{e} = \frac{1}{n} \sum_{i=1}^n (y_i) - [\frac{1}{n} \sum_{i=1}^n ((r * \frac{\sigma_y}{\sigma_x}) * (x_i - \bar{x})) + \frac{1}{n} \sum_{i=1}^n (\bar{y})]$	we can also break apart the sum of the predicted values into the sum of the values that involved the slope and the sum of the means of y.
$\bar{e} = \bar{y} - \frac{1}{n} \sum_{i=1}^n ((r * \frac{\sigma_y}{\sigma_x}) * (x_i - \bar{x})) - \frac{n*\bar{y}}{n}$	we can simplify the very first sum to the mean of y because the sum of all the values divided by the number of values is equal to the mean; we can simplify the last mean because the sum of a constant is equal to the number of values times that constant.
$\bar{e} = \bar{y} - \frac{1}{n} \sum_{i=1}^n ((r * \frac{\sigma_y}{\sigma_x}) * (x_i - \bar{x})) - \bar{y}$	simplify the last term because n is in both the numerator and the denominator.
$\bar{e} = -\frac{1}{n} \sum_{i=1}^n ((r * \frac{\sigma_y}{\sigma_x})(x_i - \bar{x}))$	cancel out the plus y mean and minus y mean.

$\bar{e} = -\left(\frac{1}{n} * r * \frac{\sigma_y}{\sigma_x}\right) * \sum_{i=1}^n (x_i - \bar{x})$	we can pull out the constant that's multiplying the values inside the sum.
$\bar{e} = -\left(\frac{1}{n} * r * \frac{\sigma_y}{\sigma_x}\right) * [\sum_{i=1}^n (x_i) - \sum_{i=1}^n (\bar{x})]$	we can then break apart that sum into the sum of x values and the sum of the means of x.
$\bar{e} = -\left(\frac{1}{n} * r * \frac{\sigma_y}{\sigma_x}\right) * (n * \bar{x} - n * \bar{x})$	the sum of all the values is equal to the mean of the values times the number of values; the sum of a constant is equal to the number of values times that constant.
$\bar{e} = -\left(\frac{1}{n} * r * \frac{\sigma_y}{\sigma_x}\right) * 0$	adding those two terms gives us zero
$\bar{e} = 0$	so the mean of the residuals equals zero.

Isn't math fun?

2) Residuals show no trend

Now let's see why the residuals cannot have a trend. Let's say our residuals graph looked like this-- clearly there's a trend.



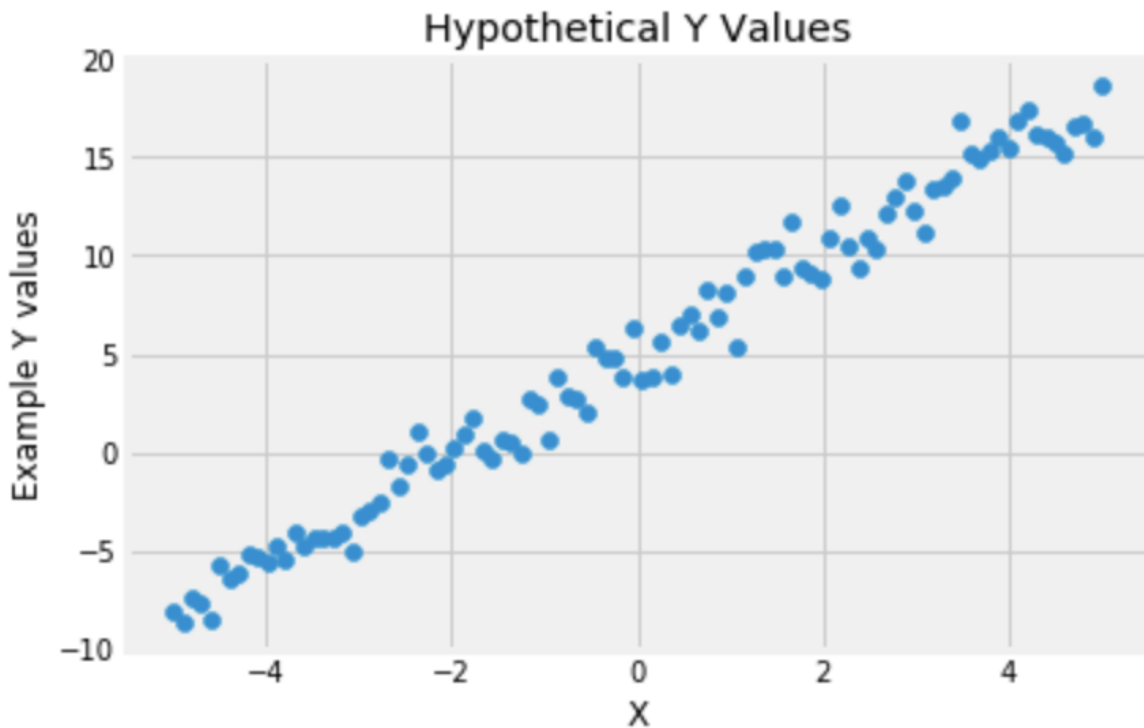
What would a line of best fit look like that had these residuals? Let's arbitrarily say our line of best fit is:

$$y = 2x + 5$$

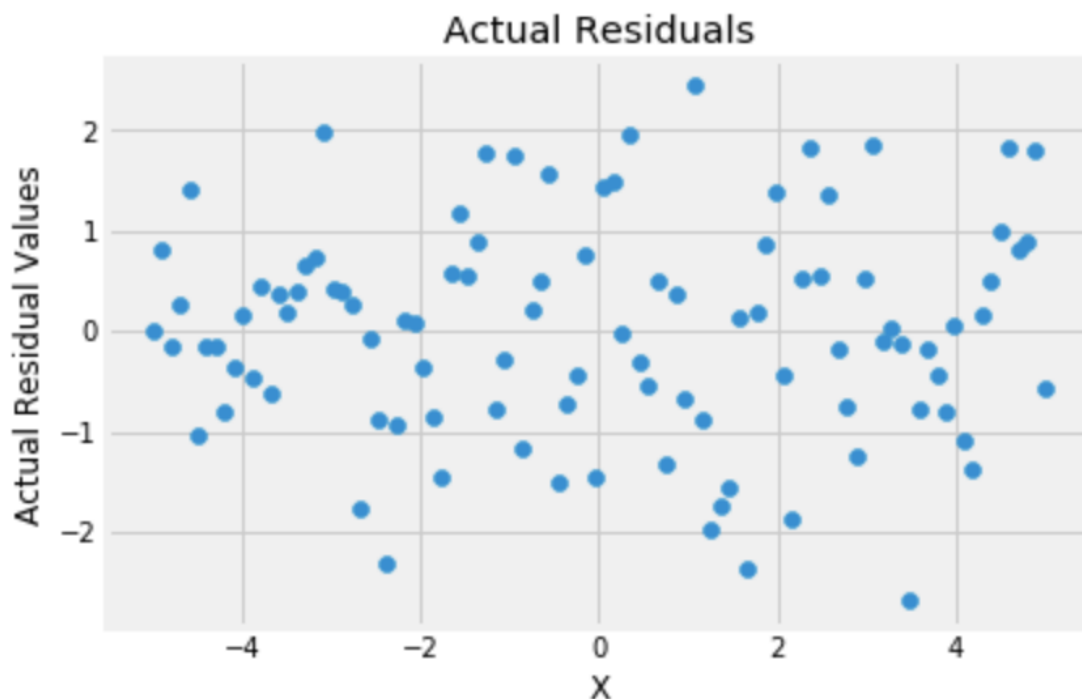
Our line of best fit could be any line, but we've chosen these numbers to make the example more concrete. To get our original values back we can calculate that line and subtract the residuals.

$$\text{residuals} = y_{\text{actual}} - y_{\text{predicted}}$$

$$y_{\text{predicted}} + \text{residuals} = y_{\text{actual}}$$



Now that we have our original y values back, let's calculate the true line of best fit.



$$y = 2.6x + 5.0$$

So those hypothetical residuals with a trend are impossible! Remember that residuals are the errors for the best fit line- clearly the best fit line we had was not the actual best fit line, so those hypothetical residuals were not possible!

3) The SD of the residuals is equal to the RMSE

After proving that the residuals have no trend and are centered at zero we can look at some other properties as well! The RMSE is equal to the SD of the residuals:

To make this proof easier to read we are going to use some fancy math terminology.

The mean of x can be written as \bar{x} , read as "x bar".

The standard deviation of x can be written as σ_x read as "SD x" or "sigma x".

We can write the prediction for y_i as \hat{y}_i , read as "y hat sub i".

For this proof instead of writing "residuals" we will use the variable e .

You read $\sum_{i=1}^n (x_i)$ as "the sum from i equals 1 to i equals n of the ith element of x"

Note that $\frac{1}{n} \sum_{i=1}^n (x_i)$ represents the mean of x .

Math	Explanation
$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$	The RMSE is the square root of the mean of the squared errors, an error is the actual value of y minus the predicted value of y.
$\sigma_e = \sqrt{\frac{1}{n} \sum_{i=1}^n (e_i - \bar{e})^2}$	This is the standard deviation formula
$\sigma_e = \sqrt{\frac{1}{n} \sum_{i=1}^n (e_i - 0)^2}$	the mean of residuals is 0, plug that in
$\sigma_e = \sqrt{\frac{1}{n} \sum_{i=1}^n (e_i)^2}$	simplify
$\sigma_e = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$	a residual is the error (for the best fit line), plug in the definition of an error.
$\sigma_e = RMSE$	You can see that the two values are identical.

4) Variance Decomposition

Residuals have some other properties that have excessively complicated proofs:

$$SD_{\text{Residuals}} = \sqrt{1 - r^2} SD_y$$

$$\frac{SD_{\text{Fitted Values}}}{SD_Y} = |r|$$

In summary:

- The mean of residuals is zero and they show no trend
- The standard deviation of residuals is equal to the RMSE
- $SD_{\text{Residuals}} = \sqrt{1 - r^2} SD_y$, $\frac{SD_{\text{Fitted Values}}}{SD_Y} = |r|$

[guides](#)

Updated 20 hours ago by Yanay Rosen

followup discussions for lingering questions and comments

☒ Resolved ☐ Unresolved



Anonymous Atom 29 seconds ago

is there also a pdf version?