Residual Analysis in Regression

View Video

Because a linear regression model is not always appropriate for the data, you should assess the appropriateness of the model by defining residuals and examining residual plots.

Residuals

The difference between the observed value of the dependent variable (y) and the predicted value ($\hat{y}$) is called the **residual** (e). Each data point has one residual.

Residual = Observed value - Predicted value

$$e = y - \hat{y}$$

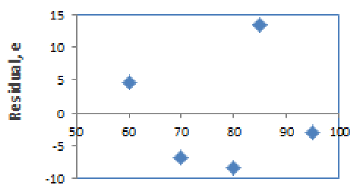
Both the sum and the mean of the residuals are equal to zero. That is, $\Sigma\, e = 0$ and $\bar{e} = 0$.

Residual Plots

A **residual plot** is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.

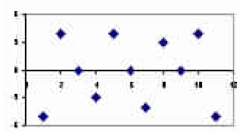
Below the table on the left shows inputs and outputs from a simple linear regression analysis, and the chart on the right displays the residual (e) and independent variable (X) as a residual plot.

x	60	70	80	85	95
y	70	65	70	95	85
ŷ	65.411	71.849	78.288	81.507	87.945
e	4.589	-6.849	-8.288	13.493	-2.945

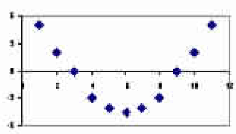


The residual plot shows a fairly random pattern - the first residual is positive, the next two are negative, the fourth is positive, and the last residual is negative. This random pattern indicates that a linear model provides a decent fit to the data.

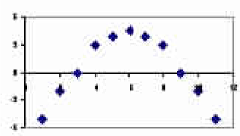
Below, the residual plots show three typical patterns. The first plot shows a random pattern, indicating a good fit for a linear model. The other plot patterns are non-random (U-shaped and inverted U), suggesting a better fit for a non-linear model.



Random pattern



Non-random: U-shaped



Non-random: Inverted U

In the next lesson, we will work on a problem, where the residual plot shows a non-random pattern. And we will show how to "transform" the data to use a linear model with nonlinear data.

Test Your Understanding

In the context of regression analysis, which of the following statements are true?

I. When the sum of the residuals is greater than zero, the data set is nonlinear.
II. A random pattern of residuals supports a linear model.
III. A random pattern of residuals supports a non-linear model.

(A) I only
(B) II only
(C) III only
(D) I and II
(E) I and III

Solution

The correct answer is (B). A random pattern of residuals supports a linear model; a non-random pattern supports a non-linear model. The sum of the residuals is always zero, whether the data set is linear or nonlinear.

< Previous lesson Next lesson >

About Us Contact Us Privacy Terms of Use Resources Advertising

View Mobile Version