REGRESSION ANALYSIS

Linear regression – fitting a straight line

ASSUMPTIONS OF REGRESSION

1. The model is adequate

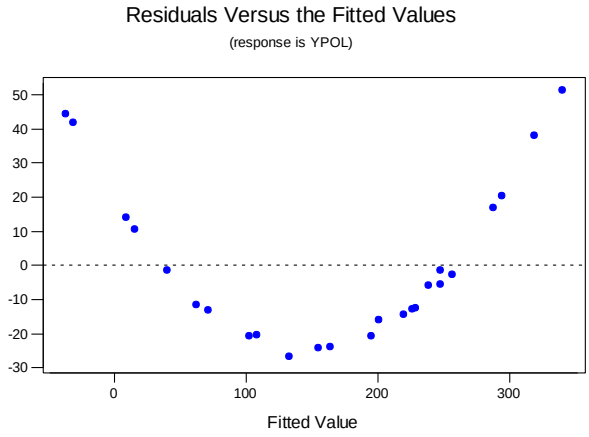
The applied model (e.g. linear, quadratic,...) describes the true relationship between the dependent (*Y*) and the independent (*x* or multiple x_i) variable(s).
Assumption check: plot residuals vs estimated *Y*

2. $Var(\varepsilon) = \sigma_y^2$

The error of the dependent variable (*y*) is constant.
Assumption check: plot residuals vs estimated *Y.*

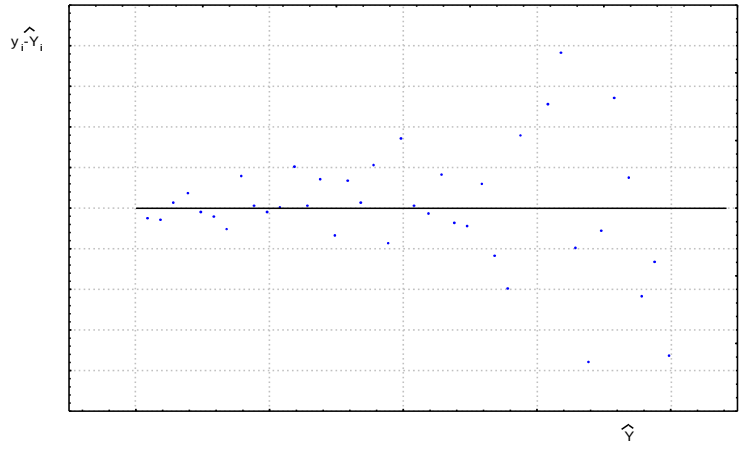
CHECKING THE ASSUMPTIONS

Residuals with respect to estimated Y
The model is not adequate



CHECKING THE ASSUMPTIONS

Residuals with respect to estimated Y
The variance of y is not constant



ASSUMPTIONS OF REGRESSION

3. The ε_i experimental errors committed at different *i* measurement points are independent

   Assumption check: plot residuals vs the order of measurements

4. *y* at all *x* values follows normal (Gauss) distribution $y \sim N(Y, \sigma^2)$

   (or, in other words: ε_i experimental errors $\sim N(0, \sigma^2)$)

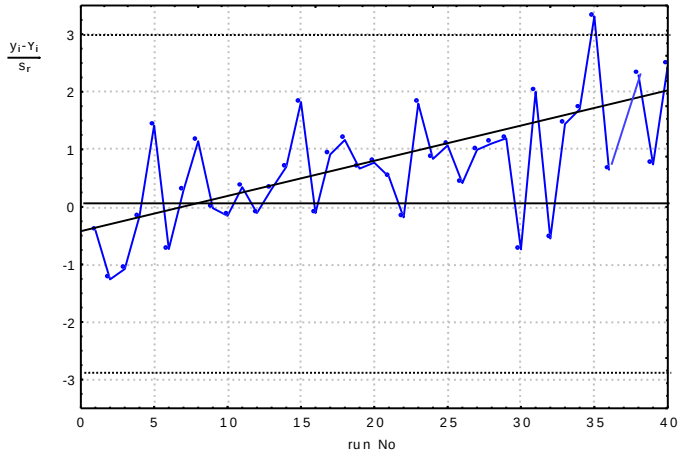
   Assumption check: check the normality of residuals with histogram or normal probability plot

5. *x* is free of error

CHECKING THE ASSUMPTIONS

Residuals vs. run number (order of measurements)

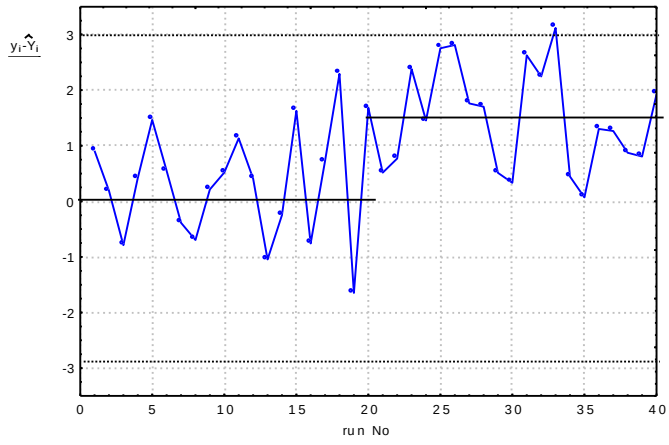
Trend → the measurement errors are not independent



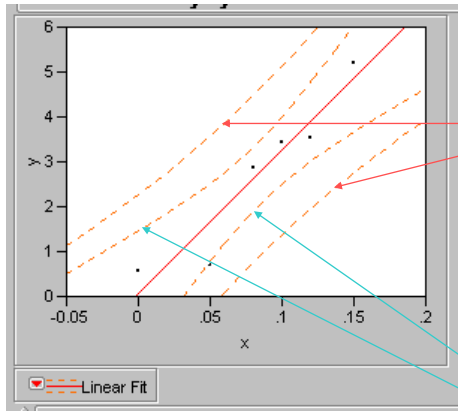
CHECKING THE ASSUMPTIONS

Residuals vs. run number

Jump → the measurement errors are not independent



Confidence and prediction band



The prediction band refers to the outcome of the dependent variable (*y*). (Expected value + error)

The confidence band refers to the expected value of the dependent variable (*Y*).

R-square

R-square: the proportion of variation explained by the model

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \bar{y})^2$$

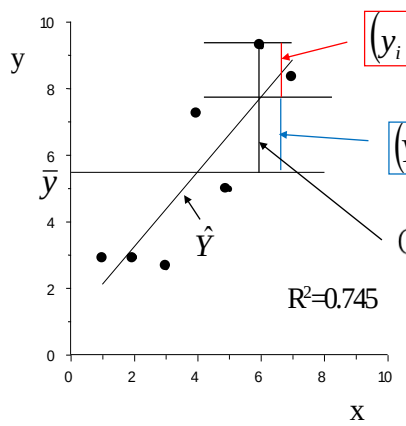
Model — Error (residual)

$$R^2 = \frac{\sum_i (\hat{Y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{\sum_i (y_i - \hat{Y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Total

R-square

Total $\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \bar{y})^2$



$(y_i - \hat{Y})$ from the total variation not explainedby the line

$(\hat{Y} - \bar{y})$ from the total variation explained by the line

$(y_i - \bar{y})$ total variation