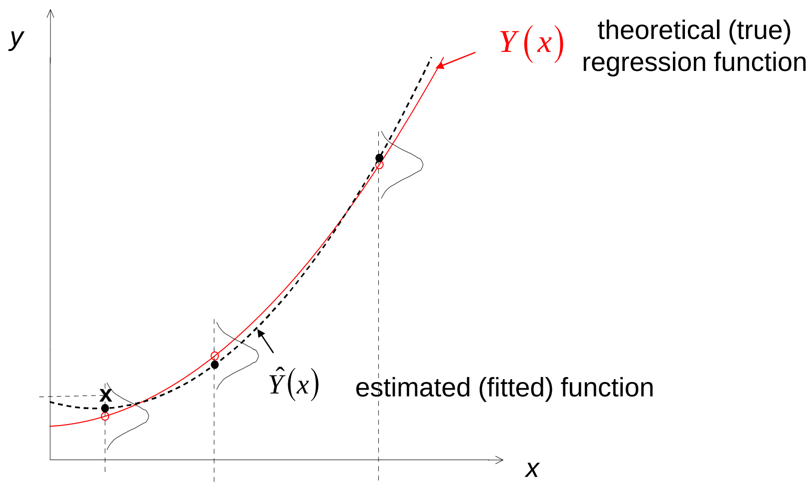
REGRESSION ANALYSIS

Linear regression – fitting a straight line

THEORETICAL AND ESTIMATED FUNCTION



TYPICAL TASKS OF THE REGRESSION ANALYSIS

- to estimate parameters of the function (model)
- to check the adequacy of the function fitted
- to test hypotheses on the parameters (e.g. if the theoretical line goes through the origin, if the slope of the line differs from zero)
- to give confidence interval (band) for the parameters (for the function)

TWO CONCEPTS FOR THE ROLE OF THE FUNCTION

1. curve (interpolating function) is to be fitted to represent properly the measured data

2. causalistic model is to be fitted, with physically sound parameters, thus the model have extrapolation ability as well

MODEL

x is the independent varable

Y is the true (theoretical or expected) value of the dependent variable

Y is a function of x $\quad Y(x) = \varphi(x; \alpha, \beta, \gamma, ...)$

E.g. for linear regression

$$Y(x) = \beta_0 + \beta x$$

$y = Y + \varepsilon \quad y$ is the measured dependent variable value

ε is the measurement error

$E(\varepsilon) = 0 \qquad Var(\varepsilon) = \sigma_y^2$

ASSUMPTIONS

- $Y(x) = \varphi(x;\ \alpha, \beta, \gamma, ...)$ is the known or assumed functional relationship with α, β, $\gamma, \ldots$ parameters
- $Var(\varepsilon) = \sigma_y^2$ is constant
- the ε_i experimental errors committed at different i measurement points are independent of each other
- y at all x values follows normal (Gauss) distribution that is the ε_i experimental errors $\sim N(0, \sigma^2)$
- x is free of error

FITTING A STRAIGHT LINE IF $\sigma_y^2 = \text{const}$

The least squares estimation criterion:

$$\phi = \sum_i \left(y_i - \hat{Y}_i\right)^2 = \min.$$

$$Y_i = \beta_0 + \beta x_i = \alpha + \beta\left(x_i - \bar{x}\right) \qquad \beta_0 = \alpha - \beta\bar{x}$$ intercept

$$\hat{Y}_i = b_0 + bx_i = a + b\left(x_i - \bar{x}\right) \qquad b_0 = a - b\bar{x}$$

$$\phi = \sum_i \left(y_i - b_0 - bx_i\right)^2 = \min.$$

FITTING A STRAIGHT LINE IF $\sigma_y^2 = \text{const}$

The normal equations:

$$\frac{\partial\phi}{\partial b_0} = -2\sum\left[y_i - b_0 - bx_i\right] = 0$$

$$\frac{\partial\phi}{\partial b} = -2\sum\left[y_i - b_0 - bx_i\right]x_i = 0$$

After rearranging

$$\sum y_i = nb_0 + b\sum x_i$$

$$\sum y_i x_i = b_0\sum x_i + b\sum x_i^2$$

As $\sum x_i \neq 0$

the b_0 and b estimators are not independent

Fitting the model in the form $Y_i = \alpha + \beta(x_i - \bar{x})$

$$\frac{\partial \phi}{\partial a} = -2\sum\left[y_i - a - b(x_i - \bar{x})\right] = 0$$

$$\frac{\partial \phi}{\partial b} = -2\sum\left[y_i - a - b(x_i - \bar{x})\right](x_i - \bar{x}) = 0$$

Rearranging (normal equations):

$$\sum y_i = na + b\sum(x_i - \bar{x})$$

$$\sum y_i(x_i - \bar{x}) = a\sum(x_i - \bar{x}) + b\sum(x_i - \bar{x})^2$$

Here the *a* and *b* estimators are independent as

$$\sum(x_i - \bar{x}) = 0$$

$$\bar{x} = \frac{\sum x_i}{n}$$

$$\sum y_i = na \qquad \sum y_i(x_i - \bar{x}) = b\sum(x_i - \bar{x})^2$$

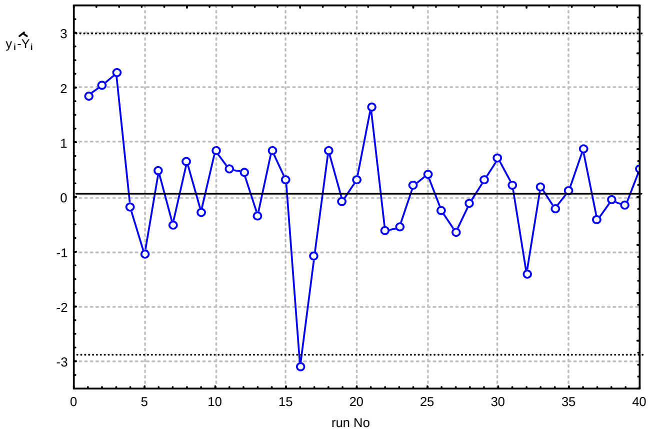
a and *b* are obtained independently of each other from the two normal equations

$$a = \frac{\sum_i y_i}{n} \qquad b = \frac{\sum_i y_i(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

$$\hat{Y} = a + b(x_i - \bar{x}) \qquad E(\hat{Y}_i) = Y_i = \alpha + \beta(x_i - \bar{x})$$

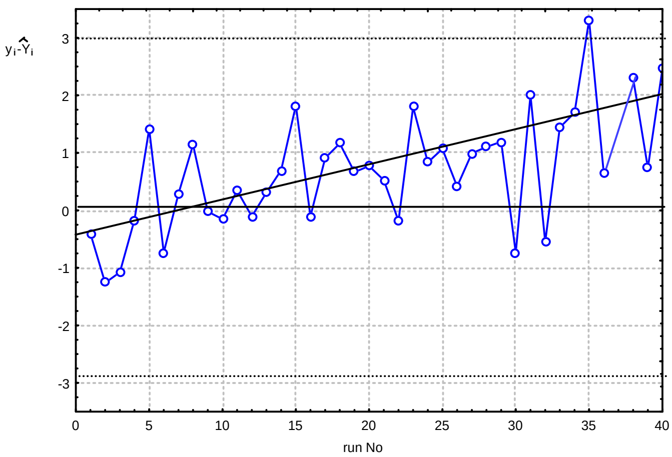
Checking the assumptions

Residuals vs. run number: outlier



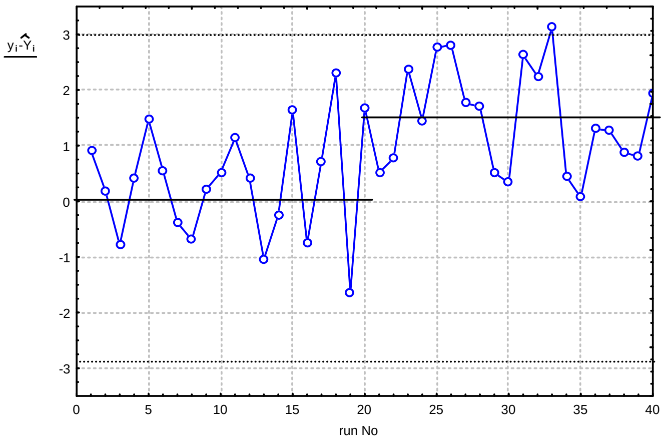
Checking the assumptions

Residuals vs. run number: trend



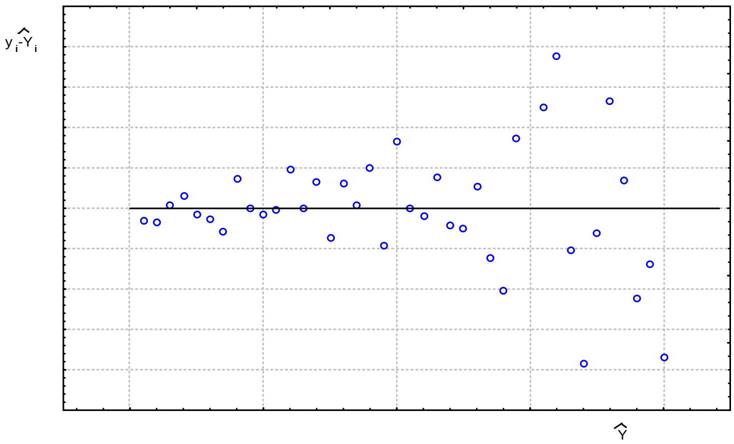
Checking the assumptions

Residuals vs. run number: jump



Checking the assumptions

Residuals with respect to Y (or estimated Y or x): change of the variance



CHECKING THE ASSUMPTIONS

Residuals with respect to Y (or estimated Y or x): adequacy of the model