**linear regression**

A simple form of *regression analysis* that attempts to find a linear relationship between two *data sets* or *random variables* $X$ and $Y$. Linear regression attempts to find the optimal parameters $a$ and $b$ of the function $y = ax + b$ describing the relation between $X$ and $Y$, i.e. the values for $a$ and $b$ that result in the least *residual sum of squares* (RSS). Visually, linear regression is a line through a *scatter plot* that is as close as possible to all *data points*. The parameter $a$ determines the *slope* of the *regression line*, and $b$ determines its *intercept*. The slope $b$ is calculated by the formula

$$a = \sum \frac{(x - \bar{x})(y - \bar{y})}{(x - \bar{x})^2}$$

where each $x$ is a data point of $X$, each $y$ is a data point of $Y$, and $\bar{x}$ and $\bar{y}$ are the *sample means* of the corresponding data sets or variables. Because $\bar{y} = a\bar{x} + b$, the intercept can then be derived from the slope:

$$b = \bar{y} - a\bar{x}$$

Given the *model* $y = ax + b$, a prediction for a value of $Y$ can be made based on a value of $X$, where $X$ is called the regressor (also: explanatory variable or *independent variable*) and $Y$ the regressand (also: explained variable, *dependent variable*). The accuracy of the predictions made by a linear regression model can be measured by the *coefficient of determination* ($r^2$), the *mean squared error* (MSE), or the *cross validation error* (CVE).