model

A function that estimates (→ *estimate*) the distribution of *data points*. More precisely, the existence of a *data generating function* is assumed, which models the generated data perfectly (but causes some *variance* due to the *irreducible error*). The *model* is then a function that resembles the data generating function as closely as possible. The data generating function is generally unknown. *Probability distributions* can be considered to be statistical models.

The *error* of a model is the fraction by which the model fails to predict *observations*. The error is measured by the *mean squared error*, which is in turn composed of the *bias* and *variance* and, in case of statistical modeling, a normally distributed irreducible error, ε:

$$MSE = E[(\hat{Y} - Y)^2] + \varepsilon$$

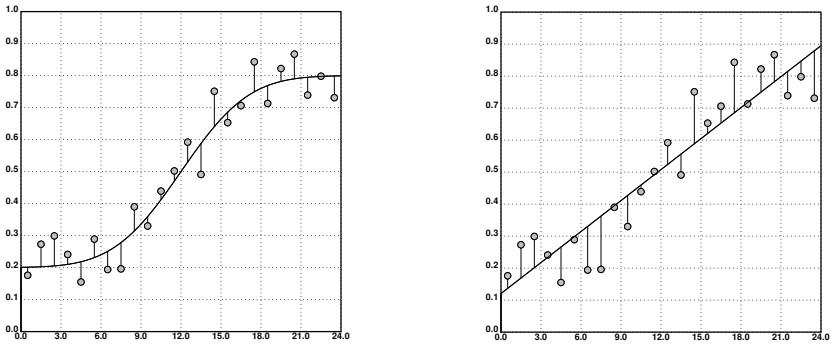

Figure **MOD:** left: data generating function, error bars show the irreducible error; right: linear model, the error bars show the total (reducible and irreducible) error; actual data points are shown as dots

The process that creates statistical models is called *regression analysis*. A simple form of statistical modeling is the *linear regression*. Figure MOD shows a data generating function and a linear model describing the same data set. The accuracy of a model can be examined using the *coefficient of determination*, the mean squared error (MSE), or the *cross validation error* (CVE).