Simple least squares linear regression is used when there are two variables, X which is known accurately and can be regarded as an independent variable, and Y which is a linear function of X except that there is measurement error or random variation which is normally distributed with zero mean and constant variance.

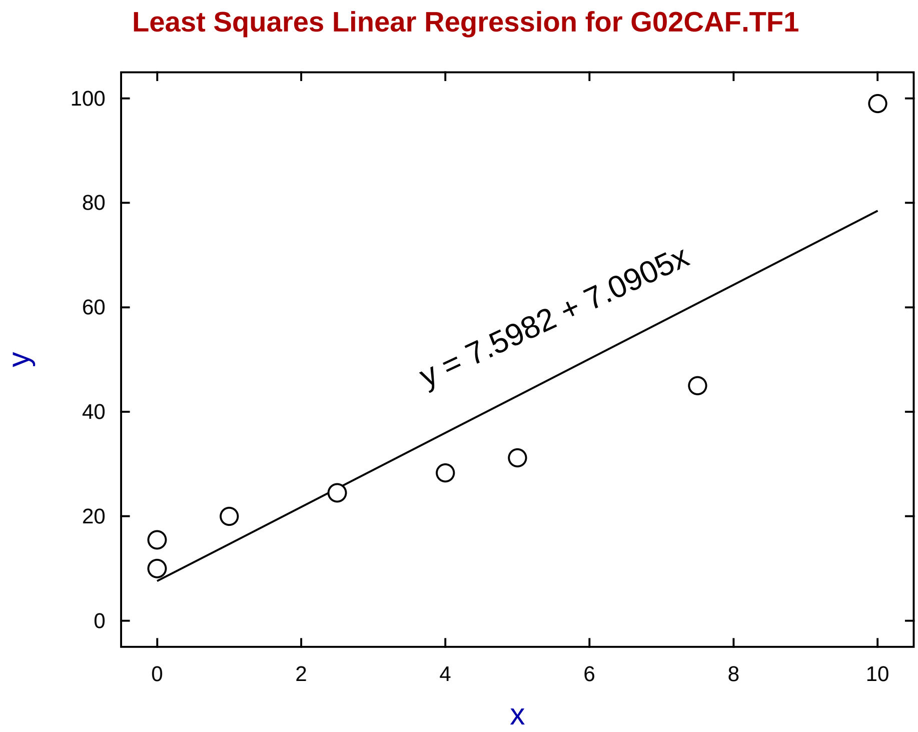
From the SIMFIT main menu choose [A/Z], open program **linfit**, choose simple linear regression and inspect the default test file `g02caf.tf1` which has the following data.

x	y
0.0	10.0
0.0	15.5
1.0	20.0
2.5	24.5
4.0	28.3
5.0	31.2
7.5	45.0
10.0	99.0

Analysis yields the following results table and plot for the least squares best-fit straight line.

Parameter	Value	Std. Error	Lower95%cl	Upper95%cl	p	
constant (c)	7.5982	6.6858	-8.7613	23.958	0.2991	**
slope (m)	7.0905	1.3224	3.8548	10.326	0.0017	

$(r^2 = 0.8273, r = 0.9096, p = 0.0017)$



The way to interpret this table is as follows.

Column 1 This indicates that the equation fitted is $y = mx + c$.

Column 2 Values for the estimated parameters ($\hat{m}$ and $\hat{c}$).

Column 3 The standard errors for the parameter estimates ($\hat{se}_m$ and $\hat{se}_c$).

Column 4 The lower 95% confidence limit for the true parameters.

Column 5 The upper 95% confidence limit for the true parameters.

Column 6 The significance level for the t variables $t_m = \hat{m}/\hat{se}_m$ and $t_c = \hat{c}/\hat{se}_c$.

Column 7 The stars indicate that the constant is not significantly different from zero.

Last line This records the Pearson product-moment correlation coefficient r, and the significance level p, indicating that the probability of these data resulting from a bivariate distribution with zero correlation parameter ρ is less than 1%.

Theory

The assumed model is that $y_i = mx_i + c + \epsilon_i$ for $n > 2$ observations, where ϵ_i is normally distributed with zero mean and variance σ^2, and the best fit parameters are those at the minimum value of SSQ defined as the sum of squared residuals, that is

$$\begin{aligned} SSQ &= \sum_{i=1}^{n} \epsilon_i^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{m}x_i - \hat{c})^2. \end{aligned}$$

The sample means $\bar{x}, \bar{y}$, standard deviations s_x, s_y, Pearson product-moment correlation coefficient r, and estimates $\hat{m}, \hat{c}$ are as follows.

$$\begin{aligned} \bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ \bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i \\ s_x &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ s_y &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} \\ r &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \\ \hat{m} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{c} &= \bar{y} - \hat{m}\bar{x} \end{aligned}$$

In order to perform an analysis of variance and estimate parameter standard errors further quantities are required. The total sum of squares SST with degrees of freedom $n - 1$, the sum of squares of deviations about the regression SSD with degrees of freedom $n - 2$, the sum of squares attributable to the regression SSR with degrees of freedom 1, and the mean square of deviations about the regression MSD are defined as follows.

$$\begin{aligned} SST &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\ SSD &= SSQ \\ SSR &= SST - SSD \\ MSD &= SSQ/(n-2) \end{aligned}$$

MSD is used as an estimate for the constant variance of y_i in order to estimate the standard errors of the slope and constant. Then the standard errors of the slope se_m and constant se_c are

$$\begin{aligned} \hat{se}_m &= \sqrt{\frac{MSD}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \\ \hat{se}_c &= \sqrt{\frac{MSD \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2}}. \end{aligned}$$

Another quantity that is sometimes required is the multiple correlation coefficient

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

where $\hat{y}_i$ is the best-fit value evaluated at x_i, and R is the correlation coefficient for y_i and $\hat{y}_i$. R^2 is said to measure the proportion of the total variation about $\hat{y}$ explained by the regression.

In the special case of fitting a straight line by least squares then we also have

$$R^2 = \frac{\left(\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2},$$

and so the multiple correlation coefficient equals the square of the Pearson product-moment correlation coefficient r between X and Y.

It should be emphasized that the equation

$$R^2 = r^2$$

is only true for the special situation where the best-fit equation is assumed to be the least squares line, that is

$$y(x) = \hat{m}x + \hat{c}.$$