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.

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>10.0</td>
</tr>
<tr>
<td>0.0</td>
<td>15.5</td>
</tr>
<tr>
<td>1.0</td>
<td>20.0</td>
</tr>
<tr>
<td>2.5</td>
<td>24.5</td>
</tr>
<tr>
<td>4.0</td>
<td>28.3</td>
</tr>
<tr>
<td>5.0</td>
<td>31.2</td>
</tr>
<tr>
<td>7.5</td>
<td>45.0</td>
</tr>
<tr>
<td>10.0</td>
<td>99.0</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Std. Error</th>
<th>Lower95%cl</th>
<th>Upper95%cl</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>constant ($c$)</td>
<td>7.5982</td>
<td>6.6858</td>
<td>-8.7613</td>
<td>23.958</td>
<td>0.2991 **</td>
</tr>
<tr>
<td>slope ($m$)</td>
<td>7.0905</td>
<td>1.3224</td>
<td>3.8548</td>
<td>10.326</td>
<td>0.0017</td>
</tr>
</tbody>
</table>

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

Least Squares Linear Regression for G02CAF.TF1
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 (\( \hat{s}_m \) and \( \hat{s}_c \)).

**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 \( t_m = \hat{m} / \hat{s}_m \) and \( t_c = \hat{c} / \hat{s}_c \).

**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 \( \rho \) is less than 1%.

**Theory**

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

\[
SSQ = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{m}x_i - \hat{c})^2.
\]

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:

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \\
\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \\
s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \\
s_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2} \\
r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}} \\
\hat{m} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\
\hat{c} = \bar{y} - \hat{m}\bar{x}
\]
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.

\[
SST = \sum_{i=1}^{n} (y_i - \bar{y})^2
\]

\[
SSD = SSQ
\]

\[
SSR = SST - SSD
\]

\[
MSD = SSQ / (n - 2)
\]

$MSD$ 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

\[
se_m = \sqrt{\frac{MSD}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}
\]

\[
se_c = \sqrt{\frac{MSD \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2}}
\]

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

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