Tutorials and worked examples for simulation, curve fitting, statistical analysis, and plotting.

https://simfit.org.uk
https://simfit.silverfrost.com

Comprehensive 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. This option provides procedures to check for goodness of fit which are not available with the simple linear regression option.

From the SIMFIT main menu choose [A/Z], open program **linfit**, choose advanced linear regression and inspect the default test file `line.tf2` which has the following data.

x	y
28.10	11.88
28.60	11.08
28.90	12.19
29.70	11.13
30.80	12.51
33.40	10.36
35.30	10.98
39.10	9.570
44.60	8.860
46.40	8.240
46.80	10.94
48.50	9.580
57.50	9.140
58.10	8.470
58.80	8.400
59.30	10.09
61.40	9.270
70.00	8.110
70.00	6.830
70.70	7.820
71.30	8.730
72.10	7.680
74.40	6.360
74.50	8.880
76.70	8.500

The two columns of data have the following meanings.

1. Column one is the independent x (with no error), the temperature in degrees Fahrenheit.
2. Column two is the dependent variable y (with error), in pounds of steam per month.

This options then fits a straight line in the form $y = mx + c$ leading to the following results.

Table 1: Parameter estimates

Parameter	Value	Std. Error	Lower95%cl	Upper95%cl	p
constant (c)	13.623	0.58146	12.420	14.826	0.0000
slope (m)	-0.079829	0.010524	-0.1016	-0.058059	0.0000
$r^2 = 0.7144, r = -0.8452, p = 0.0000$					

Table 2: Residuals

x	y	Theory	Residuals	
28.1	1.188	1.138	0.5002	
28.6	1.108	1.134	-0.2599	
28.9	1.219	1.132	0.8741	*
29.7	1.113	1.125	-0.1221	
30.8	1.251	1.116	1.3460	**
33.4	1.036	1.096	-0.5967	*
35.3	1.098	1.081	0.1750	
39.1	9.570	1.050	-0.9317	*
44.6	8.860	1.006	-1.2030	**
46.4	8.240	9.919	-1.6790	**
46.8	1.094	9.887	1.0530	**
48.5	9.580	9.751	-0.1713	
57.5	9.140	9.033	0.1072	*
58.1	8.470	8.985	-0.5149	*
58.8	8.400	8.929	-0.5291	*
59.3	1.009	8.889	1.2010	**
61.4	9.270	8.722	0.5485	*
70.0	8.110	8.035	-0.0750	
70.0	6.830	8.035	-1.2050	**
70.7	7.820	7.979	-0.1591	
71.3	8.730	7.931	0.7988	*
72.1	7.680	7.867	-0.1873	
74.4	6.360	7.684	-1.3240	**
74.5	8.880	7.676	1.2040	**
76.7	8.500	7.500	0.9999	**

Table 3: Analysis of residuals

Sum of squared residuals: SSQ	18.223
Estimated average % coefficient of variation	9.45%
R^2, correlation coefficient(theory,data)^2	0.7144
Largest Absolute relative residual	18.85%
Smallest Absolute relative residual	0.93%
Average Absolute relative residual	7.80%
Percentage of absolute relative residuals in range 0.1–0.2	36.00%
Percentage of absolute relative residuals in range 0.2–0.4	0%
Percentage of absolute relative residuals in range 0.4–0.8	0%
Percentage of absolute relative residuals > 0.8	0%
Number of residuals < 0 (m)	13
Number of residuals > 0 (n)	12
Number of runs observed (r)	17
$P(runs \leq r:$ given m and $n)$	0.9502
5% lower tail point	9
1% lower tail point	7
$P(runs \leq r:$ given m plus $n)$	0.9680
$P(signs \leq$ least number observed)	1.0000
Durbin-Watson test statistic	1.9930
Shapiro-Wilks W statistic	0.9596
Significance level of W	0.4064
Akaike AIC (Schwarz SC) statistics	-3.904 (-1.467)
Verdict on goodness of fit: *fantastic*	

Table 1
This illustrates that there was a strong linear correlation between x and y with well determined parameters, as all p values were less than 0.01.

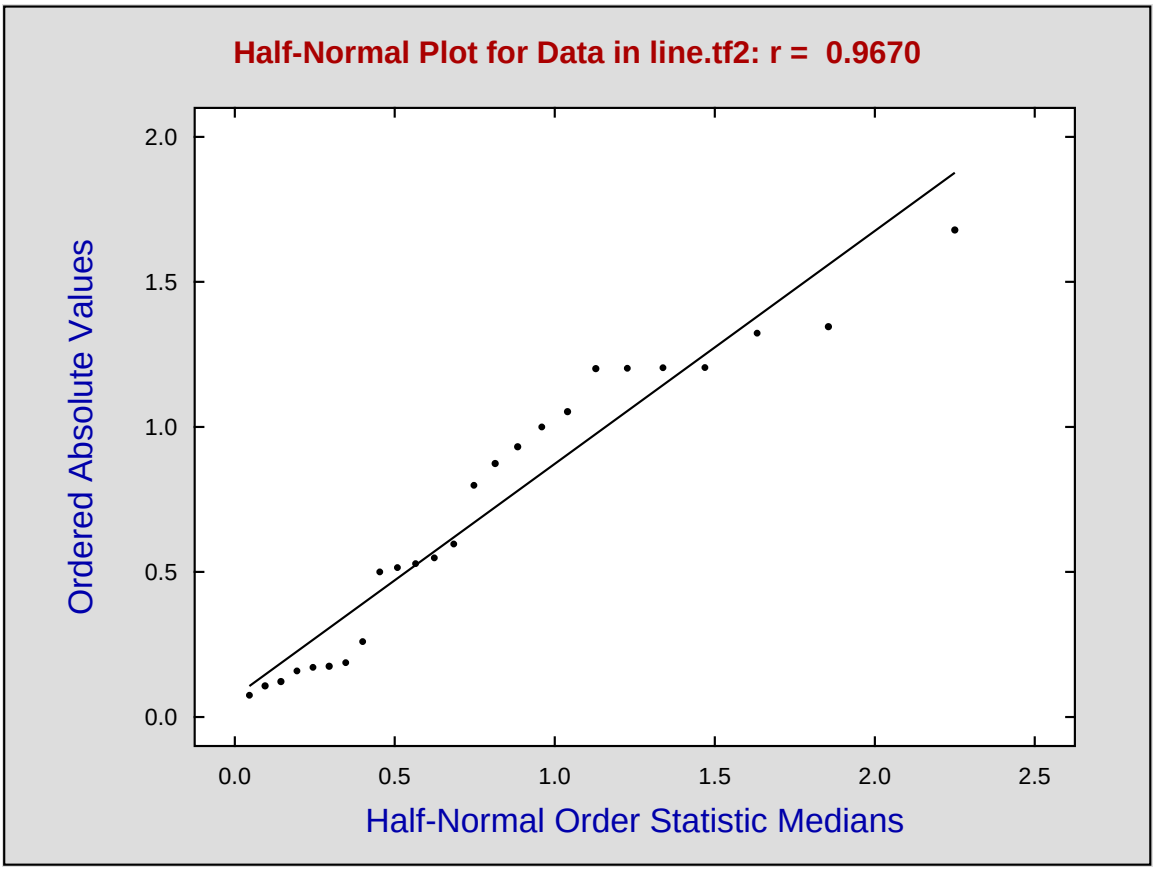
Table 2
This highlights large absolute relative residuals by the following scheme
$******> 160\%, *****> 80\%, ****> 40\%, ***> 20\%, **> 10\%, *> 5\%$
indicating that the fit is fairly reasonable, as there are only a few large values and no extremely large absolute relative residuals. Absolute relative residuals are the absolute values of the ratios of residuals to the average of experimental observations and best-fit values, that is

$$\frac{2|y_i - \hat{m}x_i - \hat{c}|}{\max(\epsilon, |y_i| + |\hat{m}x_i + \hat{c}|)}$$

where ϵ is machine precision. These are very useful because they summarize what, to most experimentalists, would be an indicator of how well a model fits the data, even though they do not have any standard statistical interpretation.

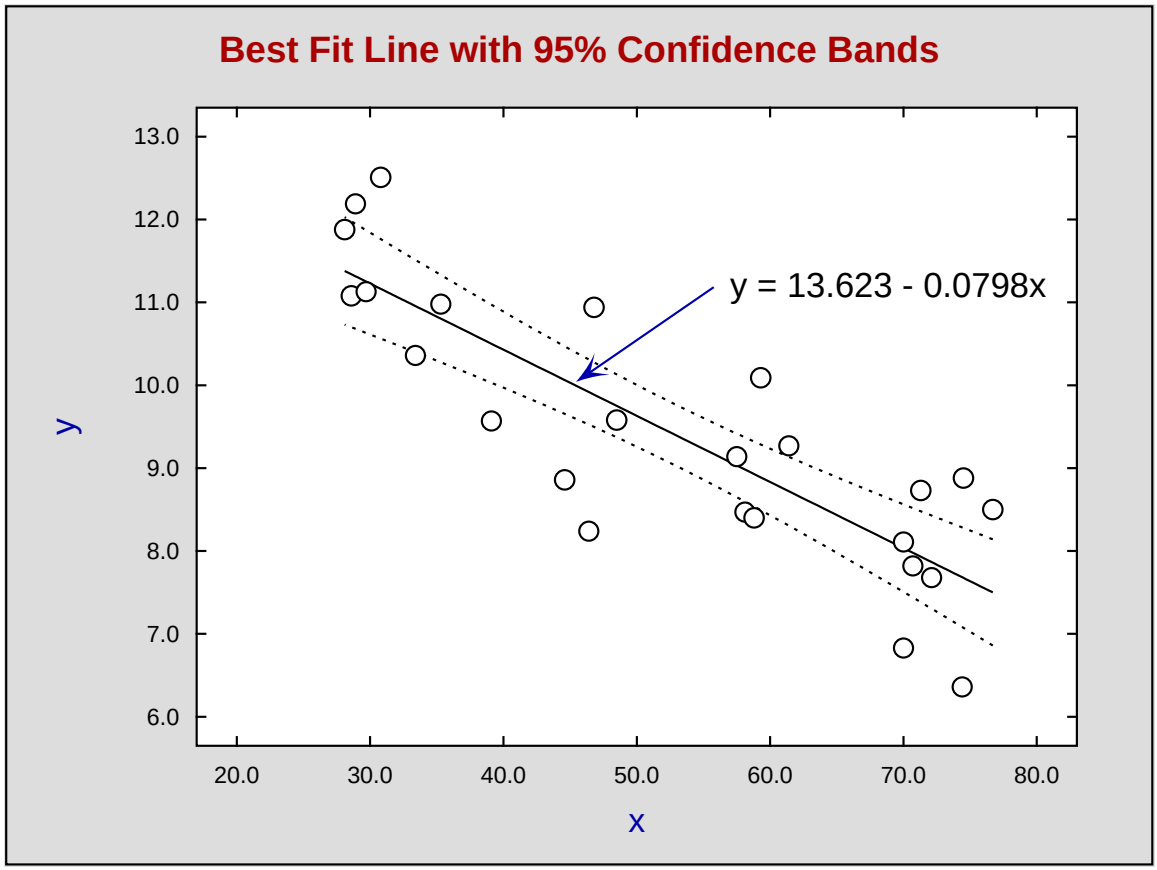
Table 3
This presents all the statistics that SimFIT uses to characterize goodness of fit leading to the qualitative, but probably over-enthusiastic, conclusion of a fantastic fit.

The Half-Normal plot



This shows a typical result with data winding around the best-fit line, and no sign of systematic deviation.

The next plot shows the data and best fit line $y = \hat{m}x + \hat{c}$ together with the 95% confidence envelope.



A $100(1-\alpha/2)\%$ confidence envelope can be added using the advanced line fitting and calibrating procedure in **linfit**, or by reading the data file into **polnom** and fitting a polynomial of degree one then requesting the addition of confidence limit curves. The confidence envelope is created using the two-valued function

$$f(x) = \hat{m}x + \hat{c} \pm t(n-2, 1-\alpha/2)\left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)^{1/2} s$$

where t is the upper 0.975 point of a distribution with $n-2$ degrees of freedom and $\alpha = 0.05$, while s is the variance estimate $SSQ/(n-2)$.

The confidence curves are used by **polnom** to estimate confidence limits for predicting x from y when a best-fit curve us used as a calibration curve.