

Linear Regression

Frequently a linear polynomial, or straight line, will fit the data, or we can transform the model into a straight line, as in the thermocouple dynamic calibration. In this case the process is called linear regression, or a linear least squares fit. There are only two equations, which can be solved explicitly for the slope and intercept, where $y = a_0 + a_1 x$.

$$a_0 = \frac{\sum x_i \sum x_i y_i - \sum x_i^2 \sum y_i}{(\sum x_i)^2 - N \sum x_i^2} \quad (4.26^1, 4.37^{2,3})$$
$$a_1 = \frac{\sum x_i \sum y_i - N \sum x_i y_i}{(\sum x_i)^2 - N \sum x_i^2}$$

We can estimate the precision of the slope and intercept of the fit by computing the standard errors:

$$S_{a_1} = S_{yx} \sqrt{\frac{N}{N \sum_{i=1}^N x_i^2 - \sum_{i=1}^N x_i^2}} \quad (4.28^1, 4.39^{2,3})$$

$$S_{a_0} = S_{yx} \sqrt{\frac{N \sum_{i=1}^N x_i^2}{N \left[N \sum_{i=1}^N x_i^2 - \left(\sum_{i=1}^N x_i \right)^2 \right]}} \quad (4.29^1, 4.40^{2,3})$$

See class web site
/notes/LINEST-ArrayFunction.xls

	LINEST Output		
slope	3.258271	1.871429	intercept
Standard error for the slope (Eqn. 4.39)	0.008693	0.096604	Standard error for the intercept (Eqn. 4.40)
r^2 (Square of Eqn. 4.38)	0.999872	0.224167	S_{yx} , Standard error of the fit (Eqn. 4.34)
The F statistic	140492.9	18	ν , Degrees of freedom [N-(m+1)]
regression sum of squares $\sum Y_i - \bar{Y}^2$	7059.858	0.904511	D, residual sum of squares (Eqn. 4.31)

Another frequently used estimator of the goodness of fit is the correlation coefficient

$$r = \sqrt{1 - \frac{S_{yx}^2}{S_y^2}}$$

where (4.27¹, 4.38^{2,3})

$$S_y^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{y})^2$$

Linear Regression

Here S_{yx}^2/S_y^2 is the ratio of the variance of the fit compared to the variance of the original data. If it is small, the data are linearly correlated. $r > .9$ is usually taken as a good linear correlation.

We can use the Student-t estimator to place confidence limits on the fit:

$$y = a_0 + a_1 x \pm t_{v,P} \frac{S_{yx}}{\sqrt{N}} \quad (4.36^4)$$

(Equations 4.35^{1,2,3} are all wrong)

indicates that the probability is P% that the fit lies in the given range, where n is the degree of freedom of the fit. See example 4.6¹, 4.9^{2,3} in the text.

LINEST Output

slope	3.258271	1.871429	intercept
Standard error for the slope (Eqn. 4.39)	0.008693	0.096604	Standard error for the intercept (Eqn. 4.40)
R Square (Square of Eqn. 4.38)	0.999872	0.224167	S_{yx} , Standard error of the fit (Eqn. 4.34)
The F statistic	140492.9	18	v , Degrees of freedom [N-(m+1)]
regression sum of squares $\sum Y_c - \bar{Y}^2$	7059.858	0.904511	D, residual sum of squares (Eqn. 4.31)

FIGURE 4.8 Results of regression analysis for Example 4.6.

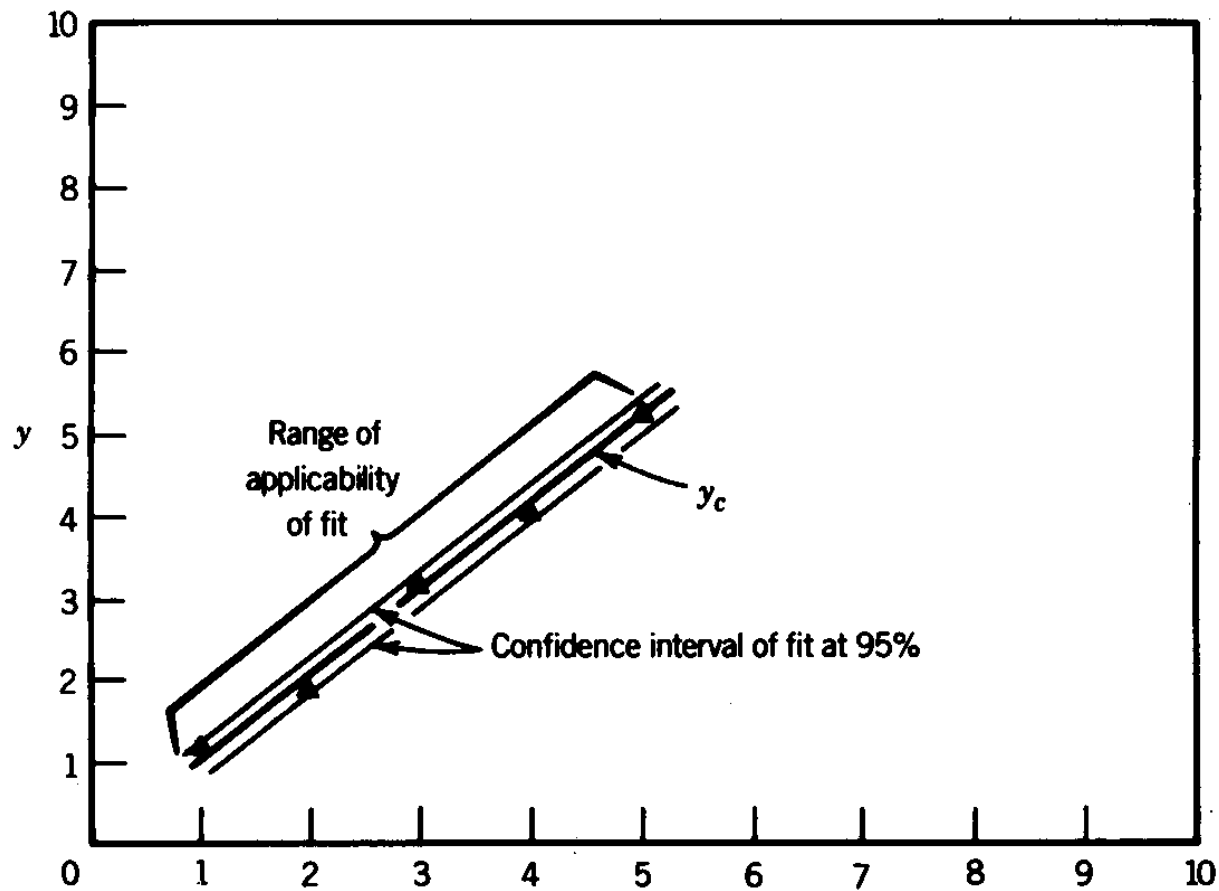


Figure 4.10 in 2nd and 3rd Edition

Number of Measurements

Suppose we want to estimate the value of a variable with a certain precision. How many times do we need to measure the quantity? The student-t estimator again provides the answer. The confidence interval CI is given by:

$$CI = x' - \bar{x} = \pm (t_{v,P}) \frac{S_x}{N^{1/2}}$$

Since the confidence interval is 2 sided, the number of necessary measurements is approximately:

$$N \approx \left[\frac{t_{v,P} S_x}{CI / 2} \right]^2 \quad (4.31^1, 4.44^{2,3})$$

Generally we don't know S_x in advance. We can make a few measurements to get an estimate of the standard error, then compute the necessary N . Once we have made the N measurements, we need to go back and see if our estimate was correct.

See Examples 4.9¹, 4.12^{2,3} and 4.10¹, 4.13^{2,3} in the text.