TOLERANCE INTERVALS FOR LINEAR REGRESSION

W. ALLEN WALLIS
UNIVERSITY OF CHICAGO

1. Introduction

Elementary textbooks frequently give the impression that lines drawn parallel to a least squares linear regression at a distance, measured in the direction of the dependent variable, equal to the standard error of estimate will include about 68 per cent of future observations from the same population, that lines at a distance equal to three times the standard error of estimate will include 99.7 per cent, and so forth.

More specifically, let y be a normally distributed random variable whose variance is σ^2 and whose mean ψ is a linear function of a second variable, x:

$$\psi = \alpha + \beta x . \tag{1}$$

From a sample of N independent observations, (x_i, y_i), maximum likelihood estimates of α and β are:

$$a = \bar{y} - b\bar{x} , \tag{2}$$

$$b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} \tag{3}$$

where $\bar{y} = \sum y/N$, $\bar{x} = \sum x/N$, and the summations, like all that follow in this paper, run over all N values of x or y. Then the estimated mean Y of y for any value of x is given by the regression line

$$Y = a + bx . \tag{4}$$

The estimate of σ, called the standard error of estimate and sometimes denoted by $s_{y.x}$, is given by

$$s^2 = \frac{\sum (y - Y)^2}{N - 2} = \frac{\sum y^2 - N\bar{y}^2 - b\sum (x - \bar{x})(y - \bar{y})}{N - 2} . \tag{5}$$

In these terms, the implication often given by elementary textbooks is that, whatever Y and s may be,

$$A = Pr(Y + K_\epsilon s > y > Y - K_\epsilon s) = \epsilon \tag{6}$$

where K_ϵ is that number which a unit normal deviate exceeds in absolute value with probability $1 - \epsilon$; that is, K_ϵ is defined by

$$\frac{1}{\sqrt{2\pi}} \int_{-K_\epsilon}^{+K_\epsilon} e^{-t^2/2} dt = \epsilon , \tag{7}$$