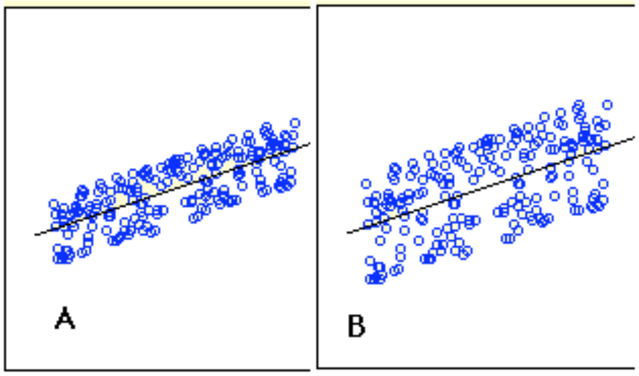
Standard Error of the Estimate

Prerequisites
prerequisites

Figure 1 shows two regression examples. You can see that in graph A, the points are closer to the line then they are in graph B. Therefore, the predictions in Graph A are more accurate than in Graph B.

Figure 1. Regressions differing in accuracy of prediction.



The standard error of the estimate is a measure of the accuracy of predictions. Recall that the regression line is the line that minimizes the sum of squared deviations of prediction (also called the *sum of squares error*). The standard error of the estimate is closely related to this quantity and is defined below:

$$\sigma_{est} = \sqrt{\frac{\sum (Y - Y')^2}{N}}$$

where s_{est} is the standard error of the estimate, Y is an actual score, Y' is a predicted score, and N is the number of pairs of scores. The numerator is the sum of squared differences between the actual scores and the predicted scores. Assume the data in Table 1 are the data from a population of five X-Y pairs.

Table 1. Example data.

	X	Y	Y'	Y-Y'	(Y-Y')^2
	1.00	1.00	1.210	-0.210	0.044
	2.00	2.00	1.635	0.365	0.133
	3.00	1.30	2.060	-0.760	0.578
	4.00	3.75	2.485	1.265	1.600
	5.00	2.25	2.910	-0.660	0.436
Sum	15.00	10.30	10.30	0.000	2.791

The last column shows that the sum of the squared errors of prediction is 2.791. Therefore, the standard error of the estimate is

$$\sigma_{est} = \sqrt{\frac{2.791}{5}} = 0.747$$

There is a version of the formula for the standard error in terms of Pearson's correlation:

$$\sigma_{est} = \sqrt{\frac{(1-\rho^2)SSY}{N}}$$

where r is the population value of Pearson's correlation and SSY is

$$SSY = \sum (Y - \mu_Y)^2$$

For the data in Table 1, m_Y = 10.30, SSY = 4.597 and r = 0.6268. Therefore,

$$\sigma_{est} = \sqrt{\frac{(1-0.6268^2)(4.597)}{5}} = \sqrt{\frac{2.791}{5}} = 0.747$$

which is the same value computed previously.

Similar formulas are used when the standard error of the estimate is computed from a sample rather than a population. The only difference is that the denominator is N-2 rather than N. The reason N-2 rather than N-1 is used is that two parameters (the slope and the intercept) were estimated in order to estimate the sum of squares. Formulas comparable to the ones for the population are shown below.

$$\mathrm{s}_{est} = \sqrt{\frac{\sum (Y - Y')^2}{N - 2}}$$

$$\mathrm{s}_{est} = \sqrt{\frac{2.791}{3}} = 0.964$$

$$s_{est} = \sqrt{\frac{(1 - r^2)SSY}{N - 2}}$$