Empirical Software Engineering, 7, 375–376, 2002.


# Comments on 'A Simulation Tool for Efficient Analogy Based Cost Estimation', by L. Angelis and I. Stamelos, published in Empirical Software Engineering, 5, 35–68 (2000)

MAGNE JØRGENSEN
*Simula Research Laboratory, Norway*

The paper 'A Simulation Tool for Efficient Analogy Based Cost Estimation' by L. Angelis and I. Stamelos describes a very interesting study on software development effort estimation. As far as we know, this is the very first paper empirically evaluating software development effort prediction interval models! Unfortunately, the regression model based software development effort minimum–maximum intervals seem to be identical to the *confidence intervals of the mean effort*, instead of the more appropriate *effort prediction intervals* (of new observations). This error leads to problems with the comparison of 'hit rates', i.e., proportion of actual effort inside the minimum–maximum intervals, of the regression model and the analogy + bootstrap model suggested by the authors. For example, the authors report that the 95% prediction interval, i.e., the interval that includes the actual effort with a probability of 95%, for the third smallest observation of the Albrecht data set (Albrecht and Gaffney, 1983), to be [8.3; 18.1]. When we replicated their analysis, with the same type of regression model and the same data, we found the prediction interval to be [3.2; 48]. The actual effort on that project was 3.6. We replicated all the prediction interval calculations of the regression model and found a hit rate of 96%, i.e., the same hit rate as the analogy + bootstrap model and close to the 'statistically correct' hit rate of 95%. The authors, however, reported a hit rate of the regression model based intervals of only 63%. As opposed to what the authors claim, we found that the regression model had the same prediction interval accuracy as the analogy + bootstrap model.

We find the use of the terms confidence and predictions intervals in many statistical tools, text-books and scientific papers rather confusing. There is a wide variety in use of different terms to denote the same type of interval, and same terms to denote different types of intervals. This lack of standardised terminology may have been one reason for the calculation error described above. It is, therefore, important to be very careful when interpreting and precise when describing output from statistical analyses on confidence and/or prediction intervals. For example, we were not sure what the authors' meant by 'confidence intervals for our estimation'. Was this meant to be the minimum–maximum interval of a new prediction (as it should have been), or the confidence interval of the mean effort (as it actually was)?

The following example illustrates the important difference between the two types of intervals: Assume that we have a single random sample ($n$ independent observations) from a normally distributed population. The term *confidence interval* (applying the terminology we prefer) may then denote the interval of possible $\mu$ (mean) values in which we are 'confident' that the true value of $\mu$ lies and is defined as:

$$y\_\text{mean} \pm t(\text{confidence level, degrees of freedom}) * s * (1/\sqrt{n});$$

where $y_$mean is the sample mean, $t()$ is the $t$-value and $s$ is the standard deviation of the sample. A high $n$ gives a narrow confidence interval around the $y_$mean. If $n$ becomes very large, the confidence interval of the mean will be close to 0.

The *prediction interval*, on the other hand, indicates the probable values of a new observation. This is a very different interval where the sources of variance are both the uncertainty of where the true mean value of the population lies and the standard deviation of the population. The definition of prediction interval in the one sample situation is:

$$y\_\text{mean} \pm t(\text{confidence level, degrees of freedom}) * s * \sqrt{(1 + 1/n)}.$$

Here, the $1/n$-term reflects the uncertainty of the mean value and the 1-term the standard deviation of the population. The prediction interval is always wider than the confidence interval. For regression models the difference between a prediction and confidence interval increases with the input's distance to the 'input mean', e.g., the difference is even more important for extreme input values. More on this topic can be found in, for example, (Christensen, 1998).

## References

Albrecht, A. J., and Gaffney J. E. 1983. Software function, source lines of code, and development effort prediction, *IEEE Transactions on Software Engineering* 9: 639–648.

Christensen, R. 1998. *Analysis of Variance, Design and Regression. Applied statistical methods*. Chapman & Hall.