Full citation: Kitchenham, B.A., Pickard, L.M., MacDonell, S.G., & Shepperd, M.J. (2001) What accuracy statistics really measure, IEE Proceedings - Software 148(3), pp.81-85.
http://dx.doi.org/10.1049/ip-sen:20010506

What accuracy statistics really measure

Barbara A. Kitchenham
Department of Computer Science
Keele University
Keele, ST5 5BG UK
+44 1782 583413
barbara.kitchenham@cs.keele.ac.uk

Lesley M. Pickard
Keele University
lesley@cs.keele.ac.uk

Stephen G. MacDonell
Department of Information Science
University of Otago
P.O. Box 56, Dunedin, New Zealand
+64 3 479 8142
stevemac@infoscience.otago.ac.nz

Martin J. Shepperd
Empirical Software Engineering Research Group
Department of Computing
Bournemouth University, Talbot Campus
Poole, BH12 5BB, UK
+44 1202 595503
mshepper@bmth.ac.uk

Abstract

This paper aims to provide the software estimation research community with a better understanding of the meaning of, and relationship between, two statistics that are often used to assess the accuracy of predictive models: the mean magnitude relative error, MMRE, and the number of predictions within 25% of the actuals, pred(25). We demonstrate that MMRE and pred(25) are, respectively, measures of the spread and the kurtosis of the variable z where z=estimate/actual. Thus, we consider z to be a measure of accuracy and statistics such as MMRE and pred(25) to be measures of properties of the distribution of z. We suggest that we need measures of the central location and skewness of z as well as measures of spread and kurtosis. Furthermore, since the distribution of z is non-Normal, we may need non-parametric measures of these properties. For this reason, boxplots of z are useful alternatives to simple summary metrics. We also note that the simple residuals are better behaved than the z variable and could also be used as the basis for comparing prediction systems.

Keywords: Prediction systems; software estimation; goodness-of-fit statistics; prediction accuracy; MMRE; pred(25).; residual analysis.

1. INTRODUCTION

A major challenge for managers of software projects is to be able to make accurate predictions. For example, how long will a project take, how much effort will it require and how many defects will a particular component contain? To answer this type of question has been a major goal of workers in the field of software metrics over the past 25 years. In general, the approach adopted has been to collect various measures that can then be used to construct a prediction system. For example, one might count the number of function points or perhaps count the number of reports that are to be generated and investigate the relationship between these measures and some other measure of interest such as the effort to complete a project.

In this paper we are not concerned with the methods used to construct a prediction system, we are interested in how researchers determine that one prediction system leads to better predictions than another. A large number of different prediction accuracy statistics have been used in the literature (see for example, Conte et al, 1986, Jorgensen, 1995, Lo and Gao, 1997, Miyazaki et al., 1991). However, in a given situation the different accuracy statistics often give contradictory results. This indicates that they are not measuring the same aspect of prediction accuracy. We believe that the lack of understanding of what different accuracy statistics actually measure is hindering progress in this important branch of software engineering.

In this paper we investigate the two most commonly used accuracy statistics: the mean magnitude relative error, MMRE, and the count of the number of predictions within m% of the actuals, pred(m), where m is usually taken to be 25. These are particularly important accuracy statistics because almost the entire software metrics research community has relied on MMRE and to a less extent pred(m) since Conte et al. publicised them. If these metrics are the basis of making comparisons between competing prediction systems, we need to be very sure what they mean.

2. THE MMRE AND PRED(25) PREDICTION ACCURACY STATISTICS

2.1 Mean Magnitude of Relative Error - MMRE

The Mean Magnitude Relative Error (MMRE) prediction accuracy statistic is the most widely used indicator in

recent years, particularly when assessing the performance of software effort estimation models. The MMRE is defined by Conte *et al*. (1986) as:

$$\frac{1}{n}\sum_{i=1}^{i=n}\left(\frac{\left|x_i-\hat{x}_i\right|}{x_i}\right)$$

where x_i is the actual value and $\hat{x}_i$ is the estimated value of a variable of interest.

In our view, however, this is not particularly meaningful for assessing predictions (as opposed to providing a goodness of fit statistic). If the aim is to generate an estimate of the effort for a new project, upper and lower bounds about the estimate are normally required, in order to present a range of values likely to contain the actual value. In other words interest is in the deviation relative to the *estimate* not relative to the *actual*. This is consistent with statistical residual analysis where the residuals (i.e. the estimate-actual) are plotted against the estimated values not the actual values.

A formulation of the MMRE where the absolute residuals are divided by the estimate can be referred to as the EMMRE (Estimation MMRE).

In order to understand what the MMRE measures, consider a random variable x distributed normally with mean μ and variance σ^2. Iglewicz (1983) demonstrated that for a sample of size n, where $\bar{x}$ is the average of the n observations:

$$d_n=\frac{1}{n}\sum_{i=1}^{i=n}\left|x_i-\bar{x}\right|\rightarrow\sigma\sqrt{\frac{\pi}{2}}\ \text{as}\ n\rightarrow\infty$$

If we rewrite the MMRE as follows:

$$\frac{1}{n}\sum_{i=1}^{i=n}\left|\frac{\hat{x}_i}{x_i}-1\right|=\frac{1}{n}\sum_{i=1}^{i=n}\left|z_i-1\right|$$

it is clear that if $\hat{x}_i$ is an unbiased estimator of x_i, the expected value of $z_i=\dfrac{\hat{x}_i}{x_i}$ is 1.

Furthermore, if z_i is distributed Normally with mean 1 and variance σ_z, the MMRE tends to the value $\sigma_z\times\sqrt{\dfrac{\pi}{2}}$. This demonstrates that the MMRE is an estimate of the *spread* of the variable *z* that will not be so vulnerable to large outliers as the conventional root mean square estimate. In addition, the *median* magnitude relative error would be an even more robust measure of spread. Since MMRE is a measure of spread it is incorrect to refer to it as a measure of prediction accuracy. The variable *z* is a better indicator of prediction accuracy since it has a defined optimum value (i.e. 1) which indicates clearly whether or not the prediction system under- or overestimates.

Using the above argument, the EMMRE will be an estimate of spread of the variable $q=\dfrac{1}{z}$.

This discussion indicates that the quality of a prediction system can be reported in terms of the average or median value of the prediction accuracy variables *z* or *q*, and the MMRE or EMMRE can be used to assess the variability of *z* and *q* respectively.

2.2 Pred(m)

Another widely used prediction quality indicator is pred(m), which is simply the percentage of estimates that are within m% of the actual value. Typically m is set to 25 so the indicator reveals what proportion of estimates are within a tolerance of 25%. Clearly, pred(m) is insensitive to the degree of inaccuracy of estimates outside the specified tolerance level. For example, a pred(25) indicator will not distinguish between a prediction system for which predictions deviate by 26% and one for which predictions deviate by 260%.

As with MMRE, it is preferable to formulate pred(m) for estimating by considering the percentage of actuals within m% of the estimate.

Based on the discussion of MMRE above, it is clear that if the prediction accuracy (i.e. z=estimate/actual) is approximately Normal, pred(m) has (asymptotically) a functional relationship with MMRE. If $z_i=\dfrac{\hat{x}_i}{x_i}$ is distributed normally with mean $\mu=1$ and variance σ_z^2, then the proportion of actuals within m% of the estimate depends on the size of the variance compared with a Standard Normal variate which has mean of zero and a variance of 1. The MMRE provides an estimate of the variance of *z*. Recalling that the mean of *z* is 1, the proportion of actuals within m% of the estimate can be calculated using the tables of the standard normal variate and the ratio:

$$\frac{m}{100}\Big/\sigma_z$$

For example, if m=25% and MMRE=0.5, an estimate of σ_z, is $0.5/\sqrt{\dfrac{\pi}{2}}$ which is approximately 0.5/1.2533=0.3989. The proportion of actuals within 25% of the estimate corresponds to the number of actuals in the range 0.75 to 1.25. This depends on how the variance of *z* compares with the proportion m/100. In this case an upper and lower bound of 0.25 around the mean and a standard deviation of 0.3989 corresponds to plus or minus .25/.3989=0.627 standard deviations about the mean. From tables of the standard normal variate, this range corresponds to a probability of 0.46. Thus if a sample comprises 100 estimate-actual pairs, 46 of the actuals should be within 25% of the estimate.

However, pred(25) is *not* a measure of the spread of *z*. To understand what it measures, consider what happens if a

distribution is more peaked than a Normal distribution. A sample from a more peaked distribution would have more values within 25% of the mean than normal. Similarly a sample from a flatter distribution would have less values within 25% of the distribution. Thus, pred(25) is related to the *shape* of the distribution *z*. Shape has two dimensions: *skewness* which describes whether or not the distribution is symmetrical about a central value and *kurtosis* which describes the extent to which the distribution peaks around its central value. Pred(25) is therefore a measure of kurtosis.

2.3 Inconsistencies evaluations using MMRE and pred(25)

Since MMRE and pred(25) measure different properties of the distribution of *z* it is not surprising that the two statistics may appear to give inconsistent results if they are used to evaluate alternative prediction systems. For example, using the Desharnais data set (Desharnais, 1989), we can predict effort from size (measured in raw function points) in three ways:

1. OL: Using ordinary least squares on the raw data.
2. MR: Using a median regression technique on the raw data (as implemented in the STATA statistical analysis tool).
3. LNOLS: Using ordinary least squares regression on the data after applying the natural logarithmic transformation to the effort and size variables.

Using the complete 81 project data set to generate the models, and then using each of the models to make a prediction for each of the projects, we can generate the MMRE and pred(25) values for each of the prediction systems as shown in Table 1 (where the statistics for the logarithmic model are calculated after the predictions have been transformed back to the raw data scale).

Table 1. MMRE and pred(25) for Desharnais data set

Prediction System	pred(25)	MMRE
Ordinary Least Squares	42	0.697
Median Regression	42	0.652
Logarithmic Transformation	37	0.599

Based on MMRE we would conclude that the logarithmic transformation produced the best prediction system whereas the pred(25) values suggest that it produced the worst prediction system.

3. DISCUSSION

3.1 Summary statistics for *z*

We have shown that MMRE is a measure of the spread (i.e. standard deviation) of the variable *z* where $z_i = \frac{\hat{x}_i}{x_i}$ and that pred(25) is a measure of how peaked the distribution of *z* is. Thus, the two accuracy statistics measure two different properties of the distribution of *z*. This explains why they may appear to give contradictory results when they are used to assess different prediction systems. There is no reason why the distribution of *z* obtained from one system should not have a smaller variance than that of another system while also having a flatter distribution.

We have also noted that the distribution of a random variable has two other important properties: central location and skewness. The central location of the variable *z* can be assessed by the mean or median of *z*.

Skewness is conventionally measured as:

$$a_3 = \frac{m_3}{s^3}$$

where s^3 is the cube of the standard deviation and m_3 is the third moment about the mean. That is

$$m_3 = \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^3$$

where *n* is the number of values of the variable *z* in a data set and $\bar{z}$ is the mean of the *n* values. Note s^2 (the variance) is the second moment about the mean. This measure of skewness has a theoretical value of 0 for a Normal distribution, since the Normal distribution is perfectly symmetric.

The conventional measure of peakedness (kurtosis) is:

$$a_4 = \frac{m_4}{s^4}$$

where s^4 is the standard deviation taken to the fourth power and m_4 is the fourth moment about the mean. That is

$$m_4 = \frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})^4$$

Thus, pred(25) is related to kurtosis but is not the standard way of measuring it.

A problem with the use of conventional measures of central location, spread, skewness and kurtosis is that they were derived from consideration of the Normal distribution. They are not very good measures of the properties of non-Normal distributions. Since the variable *z* is defined on the range 0 to ∞ with a theoretical mean of 1, *z* must, by definition, be skewed and hence non-Normal. Thus, to understand the accuracy of a prediction system we need to understand the distribution of *z*. If we can determine the functional form of the distribution of z we can identify appropriate summary statistics to measure properties of interest. However, if we cannot identify the functional form of the distribution we need non-parametric measures of properties of the distribution.

3.2 Robust distribution statistics

Pickard et al. (1999) recommend inspecting boxplots of the residuals to compare models. This gives a good indication of the distribution of the residuals and can help

explain the behaviour of the summary statistics. In a similar way, boxplots can display the distribution of *z*. Boxplots are based on non-parametric statistics. They show the median value as the central location for the distribution. If the median is close to 1, the predictions are unbiased. The length of the box from lower tail to upper tail gives an indication of the spread of the distribution. The position of the median in the box and the length of the boxplot tails show the skewness of the distribution. If the upper and lower tails are approximately equal and the median is in the centre of the box the distribution is symmetric. If the distribution is not symmetric the relative lengths of the tails and the position of the median in the box indicate the nature of the skewness. The length of the box relative to the length of the tails gives an indication of the shape of the distribution. A boxplot with a small box and long tails represents a very peaked distribution, a boxplot with a long box represents a flatter distribution.

Boxplots provide a simple means of comparing the predictions from alternative prediction systems. For example, using the Desharnais data set and three prediction systems, we can make a prediction for each of the projects based on each of the models. Then we can generate three different sets of *z* values: z_{OLS}, z_{MR} and z_{LNOLS} (where z_{LNOLS} is calculated after the predictions have been transformed back to the raw data scale). The boxplots for the three different *z* distributions are shown in Figure 1. Figure 1 suggests that the logarithmic model gives marginally better predictions than the other models: the box length and tails are slightly smaller than the box length and tails for the other models. Furthermore the outliers from the logarithmic model are slightly less extreme than the outliers from the other models. However, the logarithmic model appears to be more susceptible to under-estimation than the other models.

3.3 Statistical tests to compare alternative prediction systems

Although boxplots allow a simple graphical method of comparing predictions from alternative prediction systems, they cannot confirm whether one predictions system is significantly better than another. Stensrud and Myrtveit (1998) suggested using a paired *t* test to test whether the absolute relative error (i.e. the MRE) for each data point obtained using one prediction system is significantly different from the absolute relative error obtained using another system. If we view the MMRE as a measure of spread, Stensrud and Myrtveit's procedure can be interpreted as testing whether or not one prediction system is more variable than another. This seems a sensible approach to assessing whether one prediction system is better than another, but it is worth considering whether other tests of the distribution of z would also be useful.

Initially it would seem that we could test for bias in our prediction system by confirming whether or not the central location of the distribution is significantly different from 1. However, for skewed distributions it is not always clear what measure of central location should be tested, since the mean, median and mode will not be equal. Furthermore, the method used to construct the prediction system can directly influence the value of central location measures. A median regression will lead to a prediction system where the median of *z*=1 (where the values of z are based on the measures used to construct the prediction system). An ordinary least squares regression applied to the logarithmically transformed data will lead to a prediction system where the geometric mean of *z* on the raw data scale equals 1. Ordinary least squares regression applied to the raw data (i.e. size and effort measures) will not result in a prediction system for which the mean of *z* is always equal to one. Ordinary least squares regression results in a prediction system where the mean of the residuals (i.e. estimate-actual) always equal 0, but if the data is skewed the mean of *z* is not guaranteed to equal 1.

For example, the central location values of the three *z* distributions shown in Figure 1 are shown in Table 2.

Figure 1. Boxplots of'the z values for each prediction system
(i) Ordinary LS regression, (ii) Median regression, (iii) Log regression

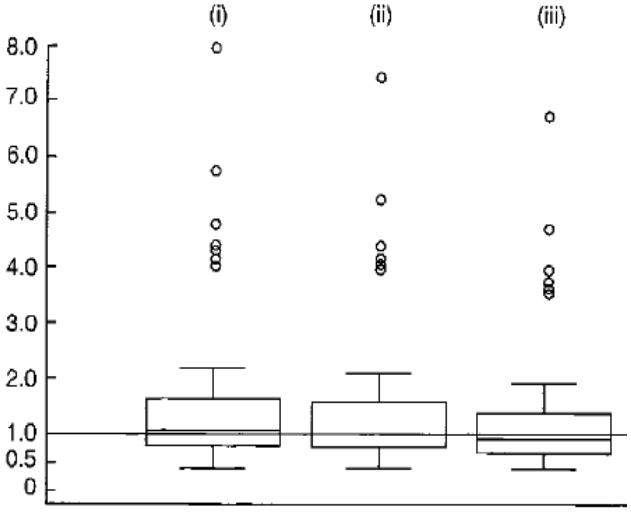


Table 2. Central Location statistics for three prediction systems

Central Location statistics	z_{OLS}	z_{MR}	z_{LNOLS}
Arithmetic average	1.463	1.384	1.251
Geometric mean	1.166	0.945	1.000
Median	1.045	1.000	0.904

It is possible to test whether the distribution of the predictions from the different prediction systems are equal or not using a non-parametric test such as the Wilcoxon matched pairs signed rank test (Siegal and Castellan, 1988) on each pair of prediction systems. In this case, the results of the Wilcoxon tests confirm that all the prediction systems have significantly different distributions ($p<0.01$). But it is not clear whether the median regression model is best because its median value is 1, or the logarithmic model is best because its mean value is closest to 1.

The spread statistics for the z values are shown in Table 3. Table 3 suggests that the logarithmic model is superior. Both the standard deviation of z and the MMRE of z_{LNOLS} are smaller than the standard deviation and the MMRE of the other *z* variables, suggesting that the predictions from

the logarithmic model are less variable than the predictions from other models.

Table 3. Spread statistics for the three prediction systems

Distribution	Standard deviation	MMRE
z_{OLS}	1.2889	0.697
z_{MR}	1.2064	0.652
z_{LNOLS}	1.0901	0.599

Paired 't' tests of the absolute relative error for each data point confirm that the logarithmic model predictions are significantly less variable than the predictions from the other models ($p<0.01$). In addition, the predictions from the median regression are significantly better than the predictions from the ordinary least squares model ($p<0.01$). Since, the boxplots in figure 1 are skewed, it is preferable to use the Wilcoxon matched-pairs signed rank test on the absolute relative error values from each pair of predictions systems. In this case, the Wilcoxon tests give results that are the same as those obtained from the paired "t" tests. However, simple sign tests, as proposed by Pickard et al. (1999), are not powerful enough to detect a statistically significant difference between the prediction systems.

This discussion seems to suggest that presenting the mean of z and using a paired test of the absolute relative deviation is all that is necessary to compare alternative prediction systems. However, there are situations where these summary statistics are misleading. The standard deviation is based on the squared deviation from the observed mean of *z*, while the MMRE is based on the absolute deviation from the theoretical mean of *z* (i.e. 1). Thus, if a prediction system consistently predicted values much larger or much smaller than the real values, is would be possible to have a very large MMRE and a mean value of z far from 1, accompanied by a very small standard deviation. Such systematic bias is much easier to observe using a boxplot than using only summary statistics. Furthermore models can be adjusted to remove the effect of systematic bias, so a model that would be rejected on the basis of the summary statistics might be recognised as potentially superior from inspection of its boxplot.

3.4 Additional benefits of the z variable

We believe that identifying the variable *z* as a measure of prediction accuracy and other statistics such as MMRE and pred(25) as measures of properties of the distribution of *z* has additional benefits beyond merely increasing our understanding of what the statistics actually measure. Currently, prediction systems are assessed as good or bad against arbitrary values of MMRE and pred(25). That is, by custom, we regard an MMRE≤0.25 and a pred(25) ≥75 as indicative of a good predictive system. However, neither of these values allow us to make simple probability statements about the accuracy of future estimates. If we consider the distribution of *z*, we can estimate confidence limits about the central value of the distribution either using the boxplot for robust limits, or, if we can identify the functional form of the distribution of z, we can construct 95% or 99% confidence limits for our predictions.

Furthermore, in an effort to compare alternative prediction systems some researchers use summary statistics based on the MMRE such as the maximum MRE and the standard deviation of the MRE (Myrtveit and Stensrud, 1999). We believe such complications are unnecessary if researchers agree that:

- Accuracy is measured in terms of *z*.
- Comparison of the alternative prediction systems are based on comparisons of the boxplots of z from the competing predictions systems together with formal tests of properties such as the bias and variability of the prediction systems.

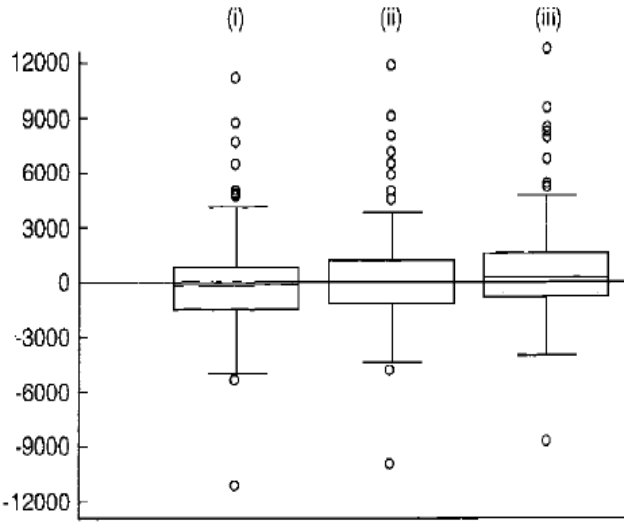
3.5 Limitations of the *z* variable

We have discussed the *z* variable at some length because it is the basis of MMRE and pred(n) but it is clear from Figure 1 that it has some undesirable properties including asymmetry. An implication of that asymmetry is that if we base our choice of prediction system on summary statistics of the z variable, we will tend to favour prediction systems that minimise overestimates rather than prediction systems that minimise underestimates. Since in most cases overestimates are less serious than underestimates, this may not lead to an appropriate choice of prediction system.

An alternative to the use of the *z* variable, is to consider the distribution of the residuals (i.e. actual-estimate). Figure 2 shows the boxplots of the residuals, which are clearly better behaved than the *z* variable in terms of symmetry. It is interesting to note that paired "t" tests of the difference between the absolute residuals suggests there is no significant difference between the three prediction systems. The Wilcoxon matched pair rank test leads to the same conclusion.

Figure 2. Boxplols of residuals for each prediction system

(i) ordinary LS regression, (ii) median regression, (iii) log regression



4. CONCLUSION

Our analysis and results suggest that the two statistics most frequently used to assess the quality of prediction systems, MMRE and pred(25), are respectively measures of the spread (standard deviation) and peakedness (kurtosis) of the variable *z* (where *z*=estimate/actual). We believe that it is necessary to understand the distribution of *z* in order to assess the accuracy of a prediction system. We suggest that boxplots of the *z* values or the residuals give a better assessment of prediction quality than one or two summary statistics. The use of boxplots is particularly appropriate since boxplots are based on non-parametric summary statistics and the variable *z* is skewed and hence non-Normal. Boxplots are also suitable for showing the distribution of residuals even though residuals are better behaved in terms of symmetry than the z variable.

Whilst the arguments in this paper may appear arcane to the non-statistician, it is essential that we understand how to make comparisons between competing prediction systems. Researchers have employed a wide range of different accuracy indicators, some of which appear to give contradictory results. Without understanding what the various indicators are describing, meaningful comparison is not possible. Furthermore if we cannot make meaningful comparisons we cannot make progress. We have argued that the indicators are statistics describing the distribution of the variable *z* and that a number of different properties of the distribution need to be described. We also note that the simple residuals are better behaved than the *z* variable. For this reason we urge researchers to present boxplots of the residuals or the *z* variable values of competing prediction systems in addition to performing appropriate statistical tests.

ACKNOWLEDGMENTS

Dr Lesley Pickard's work is supported by the EPRSC Project GR/M 33709. We thank Magne Jorgensen for his perceptive criticisms of a previous draft of this paper.

REFERENCES

Conte, S.D., Dunsmore, H.E., and Shen, V.Y.(1986). *Software Engineering Metrics and Models*, Benjamin/Cummings, Menlo Park CA.

Desharnais, J-M. Analyse statistique de la productivitie des project de developpment en informatique a partir de la technique des point des fonction. Université du Quebec a Montreal, Masters thesis, 1989.

Iglewicz, B. (1983) Robust scale estimators and confidence intervals for location. In Understanding Robust and exploratory data analysis. Hoaglin, D.C., Mosteller, F. and Tukey, J.W. (eds), John Wiley & Sons Inc.

Jorgensen, M., (1995). Experience with the accuracy of software maintenance task effort prediction models, *IEEE Trans. Soft. Eng.* 21, 674-681.

Lo, B.W.N., and Gao, X. (1997). Assessing software cost estimation models: criteria for accuracy, consistency and regression. *Australian J. of Information Systems*, 5(1), 30-44.

Pickard, L.M., B.A. Kitchenham and S.J. Linkman. (1999c) "An investigation of analysis techniques for software data sets," *Proceedings of the Sixth International Symposium on Software Metrics* (Metrics 99), IEEE Computer Society Press, Los Alamitos, California.

Miyazaki, Y., Takanou, A., Nozaki, H., Nakagawa, N., and Okada, K, (1991). Method to estimate parameter values in software prediction models, *Information & Softw. Technol.*, 33(3), 239-243

Myrtveit, I. and Stensrud, E. (1999) Does history add value to Project Cost estimation? An empirical validation of a claim in CMM. Proceedings of the combined 10th European Software Control and Metrics Conference and the 2nd SCOPE Conference on Software Product Evaluation. Hestmonceux, England, April 1999, 71-79.

Siegal S. and Castellan. J. Jr. (1988) Non-parametric statistics for the behavioural sciences. McGraw-Hill, 2nd Edition.

Stensrud, E. and I. Myrtveit. (1998) Human performance estimating with analogy and regression models: An empirical validation, Proceedings of the Fifth International Software Metrics Symposium IEEE Computer Society Press.

[Stata97] Intercooled Stata for Windows 95, STATA Corporation. http://www.stata.com