Assess the Uncertainty of COCOMO Estimation

Lang Xie, Da Yang, Ye Yang
Institute of Software, Chinese Academy of Sciences, Beijing, China
{ xielang, yangda, ye }@itechs.iscas.ac.cn

1. Backgrounds

There is always possibility that the estimation value of software cost may be different from the actual value by small or large. This uncertainty of software cost estimation is a challenging issue, and it may prevent people from proper understanding and use of the estimation result.

Traditional ways of assessing the uncertainty of COCOMO estimation include: using empirical distribution of previous estimation accuracy to build confidence intervals [1, 2], building new models to predict the estimation error [3, 4], and assessing the uncertainty by expert judgments [5].

When using empirical distribution of previous estimation accuracy to build confidence intervals, the assessed uncertainty is static and does not depend on the new project under estimation. For example, when using the empirical interval “70% possibility the magnitude of relative estimation error within 30%”, the ratio 30% is same for all new projects. When assuming RE follows the normal distribution and using [Estimated Value*(1-1.045*Std RE), Estimated Value(1+1.045* Std RE) ] as the interval for 70% confidence, the ratio ‘1.045’ is same for all new projects.

Using new model to estimate the estimation accuracy and then building confidence interval based on the previous estimation accuracy also has some limitations. One limitation is that since the estimated estimation error may results from the original cost model, the new accuracy estimation model, or the combination of these two models, it is hard to differentiate the sources of error. Besides, building another model on the residual or relative error may increase the over-fitting threat.

The above figure shows the model training and estimation process of parametric models. When making COCOMO estimation, we use the training data to get the trained model, and then make estimation on the test data with the trained model. The uncertainty of estimation may result from the model function, the training data, or the test data. Our assumption is that the training data represents the experience of estimation, and the uncertainty of the estimation on the test data will depend on how the experience from training data may work on the test data. For example, if the test data varies too much from the training data, the uncertainty of the estimation should be high. So, the uncertainty of estimation should not only depend on the trained model, but also on the test data and the relationship between the test data and the training data.

In this paper we proposed a new method to assess the uncertainty of COCOMO estimation. The new method

This work is supported by the National Natural Science Foundation of China under Grant Nos. 90718042, 60873072, and 60903050; the National Hi-Tech R&D Plan of China under Grant Nos. 2007AA010303, 2007AA01Z186, and 2007AA01Z179; the National Basic Research Program under Grant No. 2007CB310802.
can assess the uncertainty results from the model function, training data, and test data. We used simulation to validate our assumption that “the uncertainty of estimation also depends on the test data and their relationship with the training data”. We also conducted experiment on the NASA COCOMO data, and showed that the new method can better assess the uncertainty of estimation than the traditional way of using empirical distribution of previous estimation accuracy.

2. Method

COCOMO and many parametric cost estimation models are based on regression. So, we can use two metrics $\sigma_y$ and $\sigma_{y-y}$ [6, 7] to measure the estimation uncertainty. These two metrics are first introduced in regression analysis.

Let $\hat{y}$ represent the estimated value, $y$ represent the actual value, and $x$ represent the test data, then:

$$
\sigma_{\hat{y}} = \sqrt{x^T \text{Cov}(\hat{\beta})x} \quad \text{and} \quad \sigma_{y-y} = \sqrt{\sigma^2 + \hat{\delta}_y^2}
$$

$\sigma_{\hat{y}}$ measures the variance of the estimated value, and $\sigma_{y-y}$ measures the variance of the relative estimation error.

$\text{Cov}(\hat{\beta})$ represents the covariance matrix. It is approximately equal to the reciprocal of hessian matrix of the likelihood function [7]. $\hat{\delta}^2$ represents MSE(mean standard error) on the training set, which is the MLE(max-likehood estimate) the variance of $y$ in linear regression assumption [7].

Because $\hat{\delta}_{y-y}$ measures the variance of the relative error, given certain confidence level $(1-\alpha)$ we can derive the confidence interval as: $(\hat{y} - t_{d,1-\alpha} \hat{\delta}_{y-y}, \hat{y} + t_{d,1-\alpha} \hat{\delta}_{y-y})$. Where $d$ is the freedom of training set.

Because of the monotone of exponent function, we can get the interval of COCOMO estimation as $(e^{\hat{y} - t_{d,1-\alpha} \hat{\delta}_{y-y}}, e^{\hat{y} + t_{d,1-\alpha} \hat{\delta}_{y-y}})$. Consider the rate of exponent growth, the up-bound of our PI tends to be too high. So, we propose to use the following interval: $(e^{\hat{y} - t_{d,1-\alpha} \hat{\delta}_{y-y}}, e^{\hat{y} + (e^{\hat{y} - t_{d,1-\alpha} \hat{\delta}_{y-y}} - e^{\hat{y} - t_{d,1-\alpha} \hat{\delta}_{y-y}})})$

3. Experiments with simulation

In this simulation experiment, we generate synthetic data based on the function $y_i = \beta x_i + \epsilon_i \ (i = 1 \cdots 10000)$ and $\epsilon \sim N(0,1)$. We sample the $x$ of training data with a uniform distribution on the range $(3,5)$, and sample the $x$ of test data with a uniform distribution on range $(1,7)$.

![Figure 2. The variance of relative estimation error](image-url)

Figure 2 shows that $\sigma_{y-y}$ (the variance of relative estimation error) increases as the test data deviate from the center of the training data. This validates our assumption that the uncertainty of OLS estimation depends on the test
data and its relationship with the training data. While, the confidence intervals derived from the empirical distribution of previous estimation accuracy do not depend on the relationship between the test data and the training data. So, \( \sigma_{y-g} \) may be a better metrics to assess the uncertainty of estimation and derive better confidence intervals than that derived from the empirical distribution of previous estimation accuracy.

4. Experiments on NASA COCOMO data

In this section we conduct experiments to compare the confidence intervals derived from \( \sigma_{y-g} \) with the confidence intervals derived from empirical distribution of previous estimation accuracy. These experiments are based on the NASA COCOMO data [8]. We split the NASA COCOMO data set by the time of the data. We use the 77 software systems before 1985 as the training set, and the next 16 software systems from 1985 to 1987 as the test set for evaluations and comparisons. We fix the confidence level at 70%, that the actual value may drop within the estimated confidence interval with at least 70% possibility.

In figure 3, the triangles represent the confidence intervals built with our method. The squares represent the confidence intervals derived from Std RE of previous estimations. Assuming RE is normal distribution, then [Estimated Value*(1-1.045*Std RE), Estimated Value(1+1.045* Std RE) ] is the interval for 70% confidence. Figure 3 show that the confidence intervals based on Std RE is much larger than that of ours. While, the hit rate of our confidence is as high as 14/16.

In Figure 4, the triangles represent the confidence intervals built with our method, and the squares represent the confidence intervals built on empirical distribution of RE without assuming normal distribution of RE [1].

![Figure 3. Compare with confidence intervals built on Std RE](image1)

![Figure 4. Compare with confidence intervals built on empirical distribution of RE](image2)
Figure 4 shows that for 13 out of 16 projects the empirical distribution based intervals are too narrow and outperformed by our intervals, only for 2 out of 16 projects the empirical distribution based interval hit and outperformed our interval.

Figure 5. Compare with empirical confidence intervals

In the previous estimation of the training data, for 70% projects the actual effort is within interval \([\hat{y} \times (1 - 0.86), \hat{y} \times (1 + 0.86)]\). The squares in Figure 5 represent such empirical confidence interval. Figure 5 shows that the empirical confidence intervals are always larger than that of ours. The confidence intervals of our method provide more accurate assessment of the uncertainty.

5. Conclusions

In this paper we proposed new method to assess the uncertainty of COCOMO estimation. We propose that “the uncertainty of estimation also depends on the test data and their relationship with the training data”. We conducted simulation experiment and demonstrated this phenomenon. Our new methods can reflect this phenomenon, while the traditional methods of building confidence interval based on the previous estimation accuracy do not. We conducted experiments on NASA COCOMO data and showed that the confidence intervals given by our method outperformed the confidence intervals built on the previous estimation accuracy.