Some Indicators For Testing The Accuracy Of A Prediction Model And Codes In Matlab

Markela Muça*, Llukan Puka**
University of Tirana, Faculty of Natural Sciences, Dep. of Matematics
***markela.moutsa@yahoo.com, ** lpuka2001@yahoo.co.uk**

Abstract

The accurate software cost estimation of the software project is an important issue in the software engineering community. To date there are build different algorithmic models of effort (especially when there are some independent variables). After that, the work is focused upon selection of the most efficient prediction model. The goodness of the model is expressed on high percentage of spreading of dependent variables and lessening the number of independent variables. The accuracy of the model measure via F statistics, T statistics, or another indicator which are function of the model’s relative errors. Usually, the efficient model used in practice is whichever give the less mean error without testing if this model is in fact statistically significant. This can lead to unstable (erroneous) results (conclusions). There are statistics used in practice which are conditioned from the data’s probability distribution; when this is unknown, the test of hypothesis yields problems. In this paper, the application of non parametric criterion, like Wilcoxon criterion, seem to be more reasonable. In this work is used this technique in order to test the significance between two prediction models: linear regression and log-linear regression. A program is written, tested and executed in MATLAB 6.5 the calculations, implementation and testing of those algorithms are performed with mathematics package MATLAB 6.5.

Keywords: Software cost estimation, Condidence interval, Non-parametric test, Accuracy measure

Regression analysis, Prediction model

1. Introduction

An important aspect of any software programming project is to know how much it will cost. In most cases it has to be with manpower. For this reason estimation of effort for development programming is central to the management and control of a software project. A main questions that needs to be asked for any estimation method is how accurate are the predictions. Accuracy is usually defined as mean of absolute percentage errors **(MMRE, mean magnitude of relative error),** [2]. This measure is unbalanced and penalizes overestimates more than underestimates. Miyazaki**,** [3], propose another measure which is balanced mean magnitude of relative error **(BMMRE**). This approach has been criticised by Hjuz [4] as a computation that consist of two different measure that should not be combined. Median magnitude of relative errors (MdMRE) is another accuracy indicator. Others works have used the adjust coefficient of determination $\overline{R}^2$ or the coefficient of determination R^2 to indicate the percentage of variation in the independent variable that can be explained by the dependent variables. Unfortunaly, this is not always adequate indicator of prediction quality where there are extreme and outlier values. Others indicators are standard deviation **(SD)** or **Pred (25)** which is the percentage of prediction that fall within 25% of the actual value.

Clearly, the choice of accuracy indicators depends upon objectives of those using the prediction system. So, MMRE is fairy conservative with tendency against everestimates, while Pred (25) will identify those prediction systems which are generally accurate, but in some cases are inacurate. Most of the research works use validation measure based on MMRE and Pred (25).

In some of the most studies parametric and non-parametric procedures are applied to test the validity of the most prediction model. A well-known parametric test, which is suitable for comparisons and we are interested in is the paired sample t-test, while an alternative non- parametric test is the Wilcoxon signed rank test.

In these studies, there are build some scripts in MatLab, which are used to build two model and to compute the indicators and the test that we mention above.

2.Calculation of MRE for regression Log-Linear model

There are build two prediction regression model, linear and log linear regression, to predict effort. In this section we will show how the formula for computing MRE is derived when we apply the second model.

Let y be the actual cost and $\hat{y}$ be the prediction. The log-linear regression model is of the form

$$\ln y = \ln\alpha + \beta \ln X + \ln\varepsilon \qquad (1)$$

The sample model is

$$\ln \hat{y} = \sim + ¦ \ln X \qquad (2)$$

The residual is given by:

$$gabimi = \ln y - \ln \hat{y} = \ln(\frac{y}{\hat{y}}) \Leftrightarrow e^{-gabimi} = \frac{\hat{y}}{y} \Leftrightarrow \boxed{1 - e^{-gabimi} = \frac{y-\hat{y}}{y} \qquad (3)}$$

$$\boxed{MRE = \left|\frac{y-\hat{y}}{y}\right| \qquad (4)}$$

Thus from (3) and (4) we may restate MRE as, $MRE = |1 - e^{-gabimi}|$

3. Application to real data

The dataset Albrecht [1] contains from a infomation for n=24 projects. It contains of a dependent variable, the cost of a project, and 5 predictor variables. The main variable of this dataset is EFFORT which characterizes the cost of project. The main problem in this type of dataset is the cost prediction model.

We have build a sample with 24 units and each one consists of p=2 variables which are: EFFORT and SLOC. A scatter plot of the data is represented in figure 3.1, left. This graph is build in order to represent the relationship between two variables.

The **linear model** is of a form (A) :

$$y\_EFFORT = \beta_0 + \beta_1 * x\_SLOC + \varepsilon \qquad (A)$$

If we apply (A) to data set [1], we obtain the following linear regression model (using least square methods).

$$y\_EFFORT = -1.703 + 0.386 * x\_SLOC \quad (A_1)$$

In figure 3.1, left, we observe that there is a linear relationship between EFFORT and SLOC, but exhibits a pronounced heteroscedasticity (increasing variance). OLS regression analysis assumes that the data set are homoscedastic. Therefore, we propose to transform the data in an attempt to eliminate the heteroscedasticity. There exist several alternatives for transforming the data. One alternative is to calculate the logarithm of y_EFFORT and x_SLOC values.

The log-linear model is of a form (B):

$$y\_EFFORT = e^{\beta 0} * (x\_SLOC)^{\beta 1} * I \quad (B)$$

where I is lognormal with mean equal 1.Thus, $I=e^{\varepsilon}$, ε with normal distribution and then we have:

$$y\_EFFORT = e^{\beta 0} * (x\_SLOC)^{\beta 1} * e^{\varepsilon}$$

Taking the logarithm of (B), we obtain the log-linear regression model:

$$\ln(y\_EFFORT) = \beta_0 + \beta_1 * \ln(x\_SLOC) + \varepsilon$$

Taking the logarithm of the original variables, we have transformed the data. A plot of the transformed data is presented in figure 3.1, right. Appling the model (B) to the data set, we get the following regression model (B1) or (B2).

$$\ln(y\_EFFORT) = -1.917 + 1.172 \cdot \ln(x\_SLOC) \quad (B_1)$$

Backtransforming (B1), we get

$$y\_EFFORT = 0.147 \cdot (x\_SLOC)^{1.172} \quad (B_2)$$

Comparing (A) and (B), we can state that in (A) we believe in a linear relationship between EFFORT and SLOC whereas in (B), we believe in an exponential relationship between these two variables. If we consider the parameter's value, which is 1.172, close to 1 then we will observe that model (B2) is not exponential. This observation also is presented in figure 3.1, right.

\>>**Diagram_1** **%** we are in the command window, MatLab.

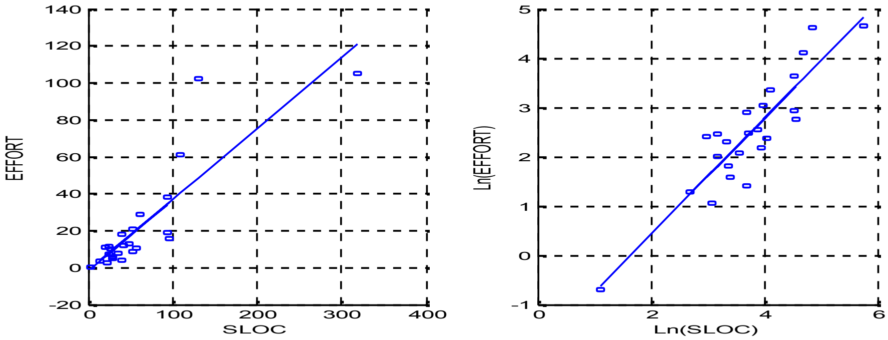

Figure 3.1. Left, SLOC versus EFFORT. Right, ln(SLOC) versus ln(EFFORT).

diagrama_1.m, this is a script, built in MatLab. This script build two graphs in figure 3.1 and calculate the prediction values from model (A) and (B) respectively.

4. The calculation of accuracy measure in two regression models using MatLab

Suppose that we wish to compare two cost prediction models (Model (A) and (B)) on the same data set. We estimate the cost of each one of the projects. Then we calculate the diffrence between the observed and real values from model (A) and (B) respectivily. We have draw two error samples (gabimi_a dhe gabimi_b).

Also, we compute confidence intervals for mean error using the above two samples (vectors with error). From each sample, estimates of the mean and the standard deviation $\bar{x}_{gab}$, s_{gab} respectively are calculated. These values are used to calculate a 100(1-*a*)% confidence interval and the results are presented in table 4.1. The results for the comparison of accuracy measure, applying two model (A) and (B) to the data set [1], are presentend in table 4.2.

\>> **treguesi_saktesise** **%** we are in the command window in MatLab

Table 4.1 Confidence interval for mean error in model (A) and (B)

Alpha	**Inteval_besimi[;]**	**Inteval_besimi[;]**
0.01	-8.1711 ; 8.1711	-6.2398 ; 10.8817
0.05	-6.0211 ; 6.0211	-3.9873 ; 8.6292
0.09	-5.1515 ; 5.1515	-3.0763 ; 7.7181

treguesi_saktesise.m this script, build in MatLab-i,calculate the accuracy measure.

Table 4.2 The accuracy measure of two model respectively

Regresi i thjeshte				
mmre_a	**bmmre_a**	**median_a**	**SD_a**	**pred25_a**
0.6270	0.5233	0.4086	0.0000	0.7917
Regresi Log linear				
mmre_b	**bmmre_b**	**median_b**	**SD_b**	**pred25_b**
0.4363	0.5585	0.3761	1.5562	0.6667

From the results of table 4.1, we can see that both models (A) and (B) are significant as the 99%, 95% and 91% confidence intervals contain the zero, whereas from the statistics of table 4.2, we can see that the measure MMRE, MEDIAN and PRED(25) are lower for model (B), which in this case can consider better than model (A).

The paired samples t-test is a parametric procedure that compares the means of two samples. We draw a sample of size n from the difference from all pairs of the samples (gabimi_a and gabimi_b). In order to test if the mean error difference is significantly different from zero we calculate a T statistics. There is a significant difference between two models if the value of T statistics is greater than $t_{n-1;1-a/2}$, where $t_{n-1;1-a/2}$ represents the 1-α/2 quantile of Student's t-distribution with n-1 degrees of freedom. The same test can be performed by computing a paired difference confidence interval which is obtained by the formula [5].

The Wilcoxon test is a non-parametric procedure, which tests if there is there is a significance difference between the medians of two paired samples. Model (A) is consider better than model (B), thus d<0, if the value of T statistics is lower than $-z_{\alpha}$.

Testi_wilcoxon_albrecht.m, this script, build in MatLab, calculate the confidence interval for the difference mean of paired ttest, paired t-test and wilcoxon test in oreder to test the two predicted model.

From the obtained results, we see that both 95% dhe 90% condidence intervals do not contain the zero, so we assess a significant difference between two models, but we can not say which model is better. The same result is obtained from the t-test. The Wilcoxon test shows that from the two comparative models, the model (A) is better than model (B).

5. Conclusions

Our prime aim is to exhibited how simple we can do comparisons between two prediction models. There are several models which can be fitted to a data set. We are interested in selected the best model using statistics procedures. For this reason we have build two prediction models based on a data set.

In this paper, we examined different criterian that are used to select the best prediction model. The best prediction model will be selected from the plot of the dataset, from the accuracy measure that are function of the model's relative errors, from the parametric Paired t-test and from the Wilxocon test which is a non-parametric analogue of the paired t-test.

The model (A) is better than the model (B) based on the these graph. We saw that the measure MMRE, MEDIAN and PRED(25) are lower for model (B), which in this case can consider better then model (A). In both cases we can not accept with accuracy for the best model. The t-test criterian shows that there is a significant difference between two model whereas the Wilcoxon criterian shows that model (A) is better than model (B).

References

[1] Albrecht, A.J. and Gaffney, J.E. 1983. Softëare function, source lines of code, and development effort prediction: A softëare science validation. *IEEE Trans.vol.9, no. 6:pp 639-648.*

[2] S.Conte, H. Dunsmore, and V.Y. Shen 1986. Software Engineering Metrics and Models. Mento Park, Calif:Benjamin Cummings.

[3] Y.Miyazaki et al.,1991. ''Method to Estimate Parameter Values in Software Prediction Models'' Information and software Technology vol.33,no3, pp. 239-243.

[4] R.T. Hughes,1996 ''Expert Judgement as an Estimating Method''. Information and Software ed. Technology, vol38,no.2,pp.67-75.

[5] Sheskin, D.J, 2004. Handbook of parametrikc and nonpametric Statistical Procedures, third ed. Chapman & Hall/CRC.