Web Based Application Maintenance Time Estimation Modeling by Bayesian SEM

Somchai Prakancharoen
Department of Computer and Information Science Faculty of Applied Science
King Mongkut's University of Technology North Bangkok
spk@kmutnb.ac.th

***Abstract*— The objectives of this research were to find out the coefficient and other parameter estimation under unknown distribution and compare new model's coefficient accuracy with [19]. This new coefficient and parameter were estimated with Bayesian analysis instead of Maximum likelihood. The model used in Bayesian analysis was start from the ML-model result from [19]. The new values were replaced into the former model then MMRE was detected from 30 completed software maintenance projects. The result of cross validation was about 40.32% while the ML-model was 42.55%.**

Keywords- Web based application software maintenance, Factor analysis, Structural equation modeling, Bootstrap, Bayesian analysis

I. INTRODUCTION

Software Maintenance is effort consumption activity and may cause critical event if it cannot be delivered to user in a suitable time. If we can precisely estimate amount of software maintenance time then project planning of software maintenance could be easily defined.

The Software maintenance time estimation model [19] was especially designed for the public and private sector of Thailand web based application during 2006-2008. Structural equation modeling" and Bootstrap technique were used to find out factors relationship and good estimation respectively. Indeed, the small dataset might cause error estimation the Bayesian analysis technique was chosen to refine the estimation parameter and coefficient results.

II. RELATED THEORY AND RESEARCH

A. Factor analysis [1]

Factor analysis is technique of reducing some unimportance indicators and grouping some related indicators to be new latent variable (factor or component). There are many method used to compose indicators to new component such as Principle component analysis: PCA, Maximum likelihood: MLE.

B. Structural Equation Modeling: SEM [2]

Structural equation modeling is a method of confirmatory factor analysis. It can analyze whether a user's purposed model of factor relation is good or not.

Estimation method of model calculation are Maximum Likelihood Estimation: ML, Generalized Least Squares: GLS, Asymptotically distribution free: ADF, Unweighted least square :ULS …etc. Each method gave difference model fitting.

Result model of SEM must be checked with Goodness of fit by some statistics such as Chi-square (χ^2) (ought to non significance), Goodness of Fit Index: GFI (exceed > 0.9), Adjusted Goodness of Fit Index: AGFI (exceed > 0.9), Root Mean Square Error of Approximation: RMSEA (lower than <= 0.06) and Hoelter's N (ought to exceed 75).

C. Software sizing

Donald J.Reifer [11] present that "Web Objects", source code sizing method, which has a similar concept to "Function Point". "Web object" defined the web based software sizing with more reasonable size estimation to web based application than Function point. The attribute of web object covered Internal logical files, External interface files, External inputs, External outputs, External inquiries, # multi-media files, # web building blocks, # scripts (animation, audio, video, visual, etc.) , # of links (xml, html and query language lines). We have chosen to use web object in software sizing in this research.

D. Relevant research

Literature review from related papers [3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 15] were studying and collect some commonly used indicators in software maintenance time estimation.

E. Evaluation criterion

Cross validation of purposed model is an important activity to confirm of model reliability. The final best fitted model would be tested with completed software maintenance in actual time. Predicted time from the purposed model is then calculated and then analyzed

MMRE value (Magnitude of Relative Error–MRE) [16] as equation (1).

$$MMRE = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{|ActualTime_i - PredictedTime_i|}{ActualTime_i}\right]\times 100 \quad (1)$$

F. Bootstrap [14]

Bootstrap is a technique which increase sample dataset from original dataset. These new sample datasets should have the same size with the original dataset. Each dataset was created by re sampling (in random) with replacement any case from the original dataset. Normally, five hundred of new datasets was the minimum amount that will be suitable for using to calculation of specific parameters. The method of how to estimate are the same as structural equation model estimation (ML, GLS, SLS,ULS). Discrepancy value which produced by these methods should indicate the poor or better fit of the model distribution estimation. Normally, the small discrepancy value mean more fitted.

G. Bayesian analysis [20]

Bayesian analysis was the technique to estimation of those parameters from a given samples. Maximum likelihood in SEM assume that probability distribution of Uniform distribution (prior) while Bayesian analysis try out the probability distribution from empirical environment and combined it with Bayesian's theorem to gain posterior distribution. The unknown parameters should be estimated from this distribution without former assunmption about distribution.

III. RESEARCH FRAMEWORK

A. Independence variables collection.

Literature review from II.D were considered and summary that 17 indicators were mostly referenced in software maintenance time estimation as illustrated in table I.

TABLE I. MOST COMMON REFERENCED INDICATORS

Indicator	Description	DataType
W_O	Web Object	0-∞
App_Req	Application Requirement	1-5
App_Reli	Application Reliability	1-5
App_Comp	Application Complexity	1-5
App_Modu	Application Modularity	1-5
MT-Cap	Maintenance Team Capability	1-5
MT_Exp	Maintenance Team Experience	1-5
MT_Coh	Maintenance Team Cohesive	1-5
MT_Stab	Maintenance Team Stability	1-5
MT_App-Exp	Maintenance Team Application Experience	1-5
App_Plt_Diff	Application Platform Difficulty	1-5
App_Lang_Diff	Application Language Difficulty	1-5
App_Aging	Application Aging	1-5
App_Lang_Old	Application Language Oldie	1-5
App_Rel_T_Org	Application Related to Organization	1-5
CMM_Lev	Capability Maturity Model	1-5
MT_T	Maintenance Team Tool	1-5

(level 1-5 is Likert rating scale : 1 =very low, 2 =low, 3 =nominal, 4 =high, 5 =very high)

B. Data collection

Indicators from III.A were designed to be a form fill-in questionnaire. This questionnaires were then submitted to target sample (thirty five public and private sector software department). One hundred and forty completed software maintenance projects were sent back. Data from questionnaire was prepared and cleaning before statistical processing. Some Indicator (Such as MT_Time, W_O) have more skew ness than criterion (not exceed +/- 1) thus Log_{10} transformation for both indicator (represent with L_MT_Time, L_W_O) would reduce their skew ness to become a normal distribution.

C. Factor Analysis

One hundred and ten completed software projects were (training case) then transformed to be standardized value for preventing from sizing domination of each indicator. PCA method was then used to factor extraction. KMO value was "0.775", Bartlett's test of Sphericity was non significant *(α=0.001)* and cumulative variance explained = "71.081 %" (appropriate large). This could conclude that factor analysis met good criterion. Variamax rotation method could depict more clarity factor and their indicators as presenting in table II.

TABLE II. FACTOR AND THEIR COMPOSED INDICATORS

Factor	Indicator
App_Attribute	zApp-Req, zApp_Aging, zApp_Lang_Old, zApp_Rel_T_Org, zCMM_Lev, zL_W_O
App_Diffculty	zApp_Comp, zApp_Plt_Diff, zApp_Lang-diff
App_Reliability_Modularity	zApp_Reli, zApp_Modu
MT_Attribute	zMT_Cap, zMT_Exp, zMT_App_Exp, zMT_Coh, zMT_Stab, zMT_T

D. Structural equation modeling (SEM)

- Four extracted factors and their indicator were constructed to be a purposed model subject to concept of researcher as represented in figure 1. Latent variable maintenance time (MT_Time) was a target latent variable which composed of one manifest variable zL_MT_Time. MT_Time was endogenous latent variable while another factor were exogenous latent variables (dominator factor).

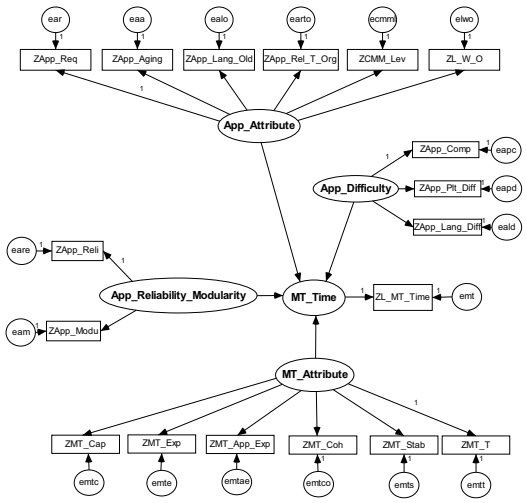

Figure 1. User's purposed model

- One thousand bootstrap datasets were generated and used to search for the best estimation methods. The discrepancy between the population moments and the sample moments were presented in the table III. The result of bootstrap calculation show that ML(small value of discrepancy-χ^2) was the best estimation methods in this research domain.

TABLE III. DISCREPANCY VALUE

		Population	Discrepancy	
		C-ML	C-SLS	C_ULS
Sample discrepancy	C-ML	478.246(0.965)	765.988(1.664)	752.491(1.634)
	C-SLS	499.963(0.922)	724.652(1.415)	711.883(1.390)
	C_ULS	499.142(0.858)	725.059(1.409)	712.283(1.385)

- SEM with ML estimation and Bootstrap calculation (ML-1,000 datasets) for all remaining significance factors and Indicators were then considered modified their relationship by adding, cutting. Goodness of fit indices were immediately observed whether their statistics were closely to the best criterion. After many trials were tried out, the met criterion model (goodness of fit) was illustrated in figure 2. The passed statistics criterion was shown in table IV.

TABLE IV. IMPORTANT GOODNESS OF FIT STATISTICS

Statistics	**Value**
Chi_ square (χ^2)	23.337 (p=0.055)
GFI, AGFI	0.956, 0.887
RMSEA	0.077
HOELTER 's N	140(0.01)

Chi_square (χ_2), GFI, AGFI, RMSEA and Hoelter's N were all passed lower criterion of best fit model.

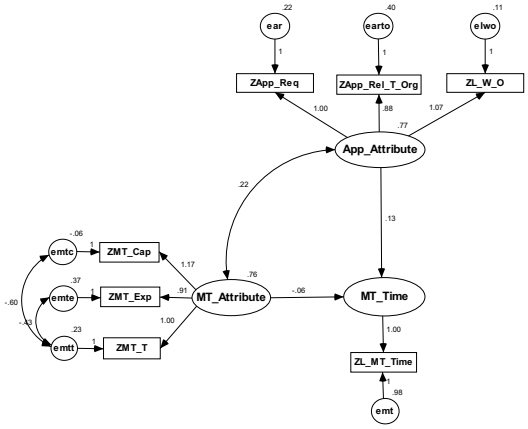

Figure 2. Final Model-best fitted

Final model in figure 2 illustrated significant factor and their indicators with MT_Time. Their relation ship quantity and direction were shown as well. MT_Time could be calculated from equation (unstandardized) (2)

$$\text{MT\_Time} = 0.13*\text{App\_Attribute} - 0.06*\text{MT\_Attribute} \quad (2)$$

$$\text{ZL\_MT\_Time} = \text{MT\_Time} + 0.98*\text{emt} \quad (3)$$

$$\text{ZMT\_Time} = \text{anti Log 10 (ZL\_MT\_Time)} \quad (4)$$

$$\text{Time} = \text{ZMT\_Time} * \sigma_{time} + \overline{time} \quad (5)$$

while "emt" estimation value was equal to "0.979", σ_{time} = "6.039" and $\overline{time}$ = "5.584"

- Bayesian analysis

The SEM model from Figure 2 was refined their coefficient and parameter with Bayesian analysis which ignore the prior distribution (Uniform distribution). After perform MCMC simulation (Monte Carlo Markov chain) about round 70,000 round the partial example of coefficient as presented in table v.

TABLE V. PARTIAL EXAMPLE OF COEFFICIENT BAYESIAN ESTIMATION

Regression weights	Mean	S.E.
ZMT_Cap<--MT_Attribute	1.780	0.071
ZMT_Exp<--MT_Attribute	1.362	0.051
MT_Time<--MT_Attribute	-0.107	0.009
MT_Time<-- App_Attribute	0.134	0.001
ZApp_Rel_T_Org<-- App_Attribute	0.884	0.000
ZL_W_O<--App_Attribute	1.091	0.001

Coefficients in equation were then value replaced with new value of coefficient which were estimated by Bayesian analysis above. The new equations were presented in (6,7,8,9).

MT_Time = 0.134*App_Attribute
-1.07*MT_Attribute (6)

ZL_MT_Time = MT_Time + 0.98*emt (7)

ZMT_Time = anti Log 10 (ZL_MT_Time) (8)

Time = ZMT_Time * σ_{time} + $\overline{time}$ (9)

while "emt" estimation value was equal to "1.043", σ_{time} = "6.039" and $\overline{time}$ = "5.584"

- Cross validation.

Software maintenance time predicted value from equation (2) and (6) were then compared to thirty known software maintenance time of completed software maintenance project. MMRE of SEM equation (2) and (6) as presented in table VI.

TABLE VI. MMRE VALUE OF MT_Time

Item	Equation (2)	Equation (6)
MMRE	42.55%	40.32%
Accuracy	57.45%	59.68%

IV. CONCLUSION AND SUGGESTION

A. Conclusion

- Maintenance time estimation in equation (2) could be used to predict software maintenance time with accuracy percentage 57.45%.
- Maintenance time estimation in equation(6) refined coefficient with Bayesian estimation, could predict software maintenance time with more accuracy percentage 59.68%.

B. Suggestion for further research

This model was small dataset then Bayesian analysis gave more accuracy estimation cause of not defined Uniform distribution but if dataset increase cases the ML should be more accurate than Bayesian analysis. In some situation, user may fixed some coefficient. What we have to do in our model in case of preserve prediction accuracy.

REFERENCES

[1] R.J. Rummel. Factor analysis . Hawaii University : USA. 2002.
[2] Garson David. Structural equation modeling. North-Carolina state university: USA. 2007.
[3] Crosby Philip. Practical software engineering. University of Calgary:USA, 2005.
[4] Mukhija Arun. Estimating software maintenance. Institute of informatik. Zurich University: SWISS, 2008.
[5] Banker D. Rajiv et El. Factor affecting software maintenance. Carnegie mellon university: USA. 1988.
[6] Stark E. George. Measurement to manage software maintenance. Mitre corporation. Colorado: USA.1997.
[7] Mukhija Arun and Glinz M.. Estimating software maintenance. Requirement research group. Zurich university: SWISS. 2003.
[8] Mustafa K.. Quality metric development framework. Al husain bin talal university: JORDAN. 2005.
[9] Cristine W. Thackaberry. Estimating metrics for course ware maintenance effort. Washington state university: USA. 1998.
[10] Krishnan S. Mayuram. The role of team factors in software cost and quality. University of Michigan : USA. 1998.
[11] Riefer J. Donald. Web object counting convention. Reifer institute. CA: USA. 2006.
[12] David F. Rico. Software process improvement. J.Ross Publishing: USA. 2007.
[13] DACS. Software Maintenance Metrics. : USA. 2006.
[14] Trevor Hastie. Generalized additive models.Stanford University, USA,1995.
[15] Pankatt b. Hatt, Inflencing Factors In Outsource software maintenance, ACM SIGSOFT,2006.
[16] Martin Shepperd, Software project effort using analogies, IEEE Vol 23,1997.
[17] Chernick, Michael R. (1999). *Bootstrap Methods, A practitioner's guide*. Wiley Series in Probability and Statistics. ISBN 0471349127.
[18] Schermelleh Karin. Nonlinear SEM is Partial least squares and alternatives. Goethe university. Germany. 2009.
[19] Somchai Prakancharoen, Web Based Application Maintenance Time Estimation Modeling using Bootstrap SEM, ACTIS 1st journal Freb-August, RMITBK, 2011.
[20] Sik-Yum Lee, Structural Equation Modeling: A Bayesian Approach, Wiley press, USA, 2011.