Application of Logistic Regression to Predict over Target Baseline of Software Projects

R. Chandrasekaran
Associate Professor
Department of Statistics
Madras Christian College
Tambaram, INDIA – 600 059

R.Venkatesh Kumar
Research Scholar
Department of Statistics
Madras Christian College
Tambaram, INDIA – 600 059

ABSTRACT
Earned Value Analysis (EVA) is a project management technique (Stephan and Mario, 2006). It is one of the most effective performance measurement tools for controlling and managing the development projects. EVAt assists, the project manager to cognize the project status and predicts future performance. The objective of this paper is to predict the *Over Target Baseline* (OTB) and *Estimate At Completion* (EAC) of the in-progress projects, based on the earned value analysis (EVA) using Logistic Regression techniques. For this purpose, we obtained ongoing data pertaining to projects from one of the major information technology (IT) company. The progressive estimates of projects, such as, *baseline cost, Planned Value*, *Earned Value* and *Actual Cost* are obtained for the real time data. This approach is applicable to have better forecast of the project cost and decreasing the risk of project cost overrun, and therefore it could be beneficial for planning preventive actions.

Keywords
Earned Duration Method, Earned Value Analysis, Estimate At Completion, Over Target Baseline, Logistic Regression.

1. INTRODUCTION
Software project management is closely related to the history of development of software packages. Software packages are developed for specific purposes on dedicated machines. Later, the concept of object-oriented programming concepts became popular in the 1960's, making *repeatable solutions* possible for the software industry. Dedicated systems could also be modified to other uses by adapting the component-based software engineering tools. Due to the simplicity of use of those software programming tools, the software industry grew very quickly in from 1970's onwards. In order to manage new development efforts, companies applied established project management methods, but project schedules failed during test runs, especially when confusion occurred between the user specifications and the delivered software. In order to shun these problems, software project management methods concentrated on matching user-requirements to delivered products, in a method known as the *Earned Value Management System* (EVMS) (Fleming and Koppelman, 2005). In this paper, the logistic regression technique is used to forecast the *Time Estimate at completion* (EAC) of an active development projects (Seyed and Zahra, 2008).

This paper is structured as follows: Section II explain the literature Review about the Earned Value Management and Logistic Regression methods. Section III includes a brief explanation about Logistic Regression. Section IV exhibits a method for forecasting Estimate at completion (EAC) of real time projects and finally, section V presents conclusion of the present research.

1.1 EARNED VALUE ANALYSIS (EVA)
Earned Value Analysis is a project management tools that tracks ongoing project status (Suketu Nagrecha, 2002). An introductory write up of EVA is available in *CDC Unified Process Practices Guide: Earned Value Management* (2006). The main objective is to identify any deviation as early as possible, so the project manager having enough time to assess the project deviation and take corrective actions, if necessary.

The EVM technique

- is useful to monitor the current project performance and forecast the future performance.
- integrates the scope, schedule and cost of a software project.
- answers a lot of questions to the stakeholders in a project relating to the performance of the project.
- is useful to demonstrate past performance, current performance and predict the future performance of the project through the use of Statistical techniques.
- when used with good planning will reduce a large amount of issues arising out of schedule and cost overruns.

In 1960's, EVA developed into a financial analysis field in US Government, but has become a subdivision of project management. During the period of late 1980s and early 1990s, it emerged as a project management methodology to be cognized and used by project managers and senior executives. EVA has become an essential part associated with the software industries. Various terms considered as input values for the EVA by many authors are given below.

1.2 INPUT PARAMETERS
Earned Value Analysis is used for monitoring, measuring, reporting and forecasting the software project performance. It can be used through all the phase of a project from Initiation to Closure.

- Define the Work Breakdown Structure (WBS)
- Assign values to each WBS Item, also called Planned Values (PV)
- Establish earning rules for each WBS Item

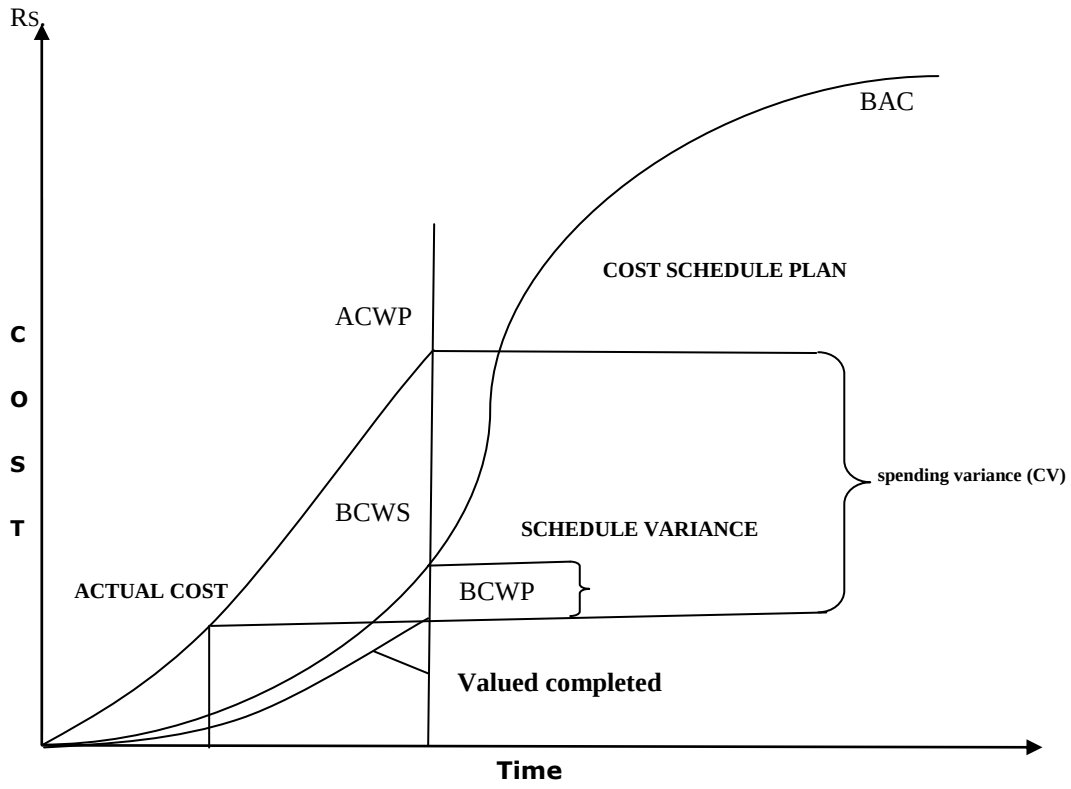

Fig. 1 Basic time and cost S-curve in EVA

On the completion of the above basic steps, the project plan can be created to track the project performance. Once these tasks have started, the earned value could be tracked. The basic data elements are shown in *Figure1* as time-cost curve with S shape in EVA literature (Seyed and Mansoureh, 2008).

Earned value analysis requires the following three key parameters to measure project performance:

Planned Value (PV) is the Budgeted Cost for Work Scheduled (BCWS)

Earned Value (EV) is the Budgeted Cost of Work Performed (BCWP)

Actual Cost (AC) is the total incurred in accomplishing the work scheduled (ACWP).

Once above three parameter are determined then they can be used to identify schedule and cost variances and to calculate overall project performance using the formulas given below.

Schedule Variance (SV) = EV – PV. Schedule Variance is a competence indicator, which reflects the schedule Performance of the project. It measures the difference between what was initially planned to complete the work and what has actually been completed. Once the project is completed, schedule variance will be become zero, because all of the planned values will have been earned.

Cost Variance (CV) = EV – AC. Cost Variance is an efficiency indicator that reflects the cost performance of the project. It measures the difference between the budget and the actual amount spent to complete the task.

Schedule Performance Index (SPI) = EV / PV. The Schedule Performance Index is used to predict the project's completion time and forecast the project's estimate at completion (EAC). SPI is defined as the ratio of the EV to PV. If the value of SPI is below 1.0, it indicates that the project is behind schedule. If SPI is equal to 1.0 then the project would be completed on time. When SPI is above 1.0, it shows that the project is ahead of schedule.

C**ost Performance Index (CPI)** = EV / AC. The Cost Performance Index is used to track the project cost and predict cost overruns. It is a commonly used to track cost – efficiency. CPI is simply the ratio of the EV to AV. The value of CPI below 1.0 indicates that the project's cost is over the planned cost for the work performed. If CPI is equal to 1, then the project would be completed on planned cost. When CPI above 1.0, the project's cost is under-planned cost for the work performed.

Estimate At Completion (EAC) = Budget At Complete (BAC) / CPI. Estimate at completion is projected total cost of the planned work at completion.

Estimate To Complete (ETC) = EAC – AC. Estimate to Complete is the expected cost needed to complete the remaining planned work.

To-complete performance index (TCPI) To-Complete Performance Index (TCPI) provides a projection of the expected performance required to achieve either the BAC or the EAC.
For the TCPI based on BAC (The performance required to meet the original BAC budgeted total):

$$\textbf{TCPI}_{\textbf{BAC}} = \frac{BAC - EV}{BAC - AC}$$

Or for the TCPI based on EAC (The performance required to meet a new or revised budget total EAC):

$$\textbf{TCPI}_{\textbf{EAC}} = \frac{BAC - EV}{EAC - AC}$$

Independent estimate at completion (IEAC) The IEAC is a metric, it predict total cost for present performance and overall performance.

$$\textbf{IEAC} = \sum AC + \frac{(BAC - \sum EV)}{CPI}$$

2. LITERATURE REVIEW

Earned Value Analysis (EVA) is a management method for integrating scope, schedule, and resources. It is used for measuring project performance and progress (Anbari, 2003). Initially started as cost/schedule control system criteria (C/SCSC) by the U.S. Department of Defense (DOD) in the 1960s, it is now mandated for many U.S. government programs and projects (Christensen, 1994; Kim, Wells and Duffey, 2003). The interest in demand for applying and implementing EVM has increased in recent years in government agencies. Organizations and auditors are required to report on the adequacy of the organization's internal control over financial reporting (Fleming and Koppelman, 2003; 2010).

There are three methods in the literature that have been proposed to measure schedule performance: The planned value method and the earned duration method explain the well known SV and SPI indicators from monetary units to time units. The earned scheduled method determines two alternative schedule performance measures (referred to as SV(t) and SPI(t)) that are directly stated in time units (Anbari, 2003).

In the past, different studies estimated that a significant number of Information Technology projects were giving undue results. For example, 2004 CHAOS report by the Standish Group reported that as many as 74% of software development project did not meet the schedule, cost or scope constraints. Other studies estimated that as many as 90% of all IT projects failed to deliver on time and within budget.

The Air Force Cost Analysis Handbook (2007) pronounces the general purpose of Earned Value Management (EVM). It is a measurement tool that provides Government and contractor Program Managers (PMs) to have discernibility into the technical and cost aspects, and to schedule performance of their projects. They should also have the capability to mitigate the risks of a project not meeting its time, budget, and performance goals.

Christensen et al. (1995) reviewed as many as 25 EAC studies. In this review, he suggested two types of studies: (1) Studies that provided new techniques for developing EACs and (2) Studies that compared a variety of techniques to determine which techniques provided better EACs. Their review incorporates index-based methods, time series techniques, performance factors, and regression approaches. Based on their comparison studies, it was concluded that the accuracy of regression-based models over index-based formulas has not been established, and therefore additional research exploring the potential of regression analysis as a forecasting tool is needed (Christensen et al., 1995).

Following Christensen et. al.'s review of EAC research, several studies used regression models. In 2005, Tracy used multiple regressions to develop EACs at five different points throughout the life of a contract (Tracy, 2005). He developed five different regression models by using three to six predictors each, in forecasting the EAC. His results indicated that the regression models usually control the performance with the early models, 25 and 35 percent complete, and begin to trade best performance with the index based models at the 50 and 65 percent complete points. Therefore, Tracy's thesis indicates that regression models might be able to outperform index methods, but only at certain times, agreeing with Christensen et. al.'s work. An Over Target Baseline (OTB) associated with the accusation of contract is extensively dealt by Kristine (2010) and Bembers et al. (2003). An OTB occurs when the contractor cannot complete the remaining amount of work within the original budget when the work scope has not changed (Cukr, 2001). An OTB is a contract budget base rescheduled to include additional performance management budget. ANSI/EIA-748-1998 defines it as a recovery plan, a new baseline for management when the original objectives cannot be met and new goals are needed for management purposes (DAU, 2003).

3. FORECASTING WITH LOGISTIC REGRESSION

Logistic regression is a variation of ordinary regression which is used when the dependent (response) variable is a dichotomous variable and the independent (input) variables are continuous, categorical, or both (Hair et. al. 1998). Unlike ordinary linear regression, logistic regression does not assume that the relationship between the independent variables and the dependent variable is a linear one. Nor does it assume that the dependent variable or the error terms are distributed normally. The form of the model is

$$\text{Log}\ (\frac{p}{1-p}) = \beta_0 + \beta_1 X_{1+} \beta_2 X_2 + \ldots. + \beta_k X_k$$

Fig 2. Logistic Regression Equation

In this model, p is the probability that the dependent variable Y=1 and X_1, X_2,.. .,X_k are the independent variables (predictors). β_0 is a constant and β_1, β_2, ... β_k are known as the regression coefficients, which have to be estimated from the data. Logistic regression estimates the probability of a certain event occurring. Logistic regression thus forms a predictor variable (log (p/(1-p)) which is a linear combination of the explanatory variables. The values of this predictor variable are then transformed into probabilities by a logistic function. Such a function has the shape of an S. On the horizontal axis we have the values of the predictor variable, and on the vertical axis we have the probabilities.

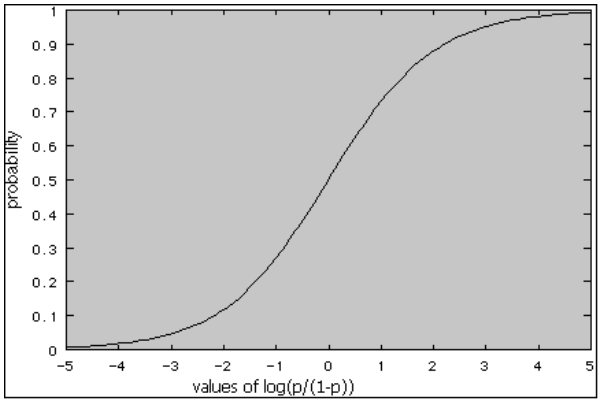

Fig 3. Odds Ratio S – Curve

Logistic regression also produces Odds Ratios (OR) associated with each predictor value. The odds of an event are defined as the probability of the outcome event occurring divided by the probability of the event not occurring. In general, the odds ratio is one set of odds divided by another. The odds ratio for a predictor is defined as the relative amount by which the odds of the outcome increase (OR greater than 1.0) or decrease (OR less than 1.0) when the value of the predictor variable is increased by 1.0 units, that is, (odds for PV+1)/(odds for PV) where PV is the value of the predictor variable.

4. METHODOLOGY

Our main objective of this section is to predict the Over Target Baseline (OTB) and Estimate At Completion-EAC of the in-progress development projects, based on the earned value analysis (EVA) using Logistic Regression techniques. For this purpose, the data is obtained from one of the major information technology company. It consists of various parameters relating to 928 on-going development projects. A variety of *potential predictor variables* are included in our model, such as *Budgeted Cost of Work Scheduled* (BCWS), *Budgeted Cost of Work Performed* (BCWP), *Actual Cost of Work Performed* (ACWP) and *Estimate At Completion* (EAC). An attempt is made to capture the variables that best explain the occurrence/non-occurrence of OTB. Using the available data on the development projects, the best possible logistic regression model is obtained to predict OTBs. These models could assist the decision makers in identifying OTBs and lead to the development of better EACs for such OTBs. Furthermore, these models explain why OTBs occur. Since an OTB generally recognizes a cost overrun, these models explain why cost overruns occur. Therefore, by building models to predict OTBs, we learn which variable has the greatest influence on the occurrence of cost overruns. The statistical software package IBM SPSS 19.0 is used for this research.

If we already know which projects are exceed target base line, then simply create the model to develop the better Estimate at Completion (EAC). If we do not know which projects will become over target baseline, we need to predict whether a project will become OTB in future. In this paper, Logistic regression technique is used to predict *Over Target Baseline* (OTB) of project duration in terms of *Time Estimate At Completion*-EAC. The models that we develop attempt to predict a dichotomous response with two possible outcomes which are given below.

- The project will become an OTB or
- The project will not become an OTB.

In our model, we have recoded the outcome variables as "1", if a project becomes an OTB; otherwise, "0", if it does not become an OTB, based on the Cost Performance Index. For a given projects, OTB is set as "1" if the Cost Performance Index (CPI) is greater than or equal to 1, else set as "0" if the CPI is less than 1. The outcome of the logistic regression functions is a probability, which represents the likelihood that a project will become an OTB. *Figures 2* and *3* provide the typical functional form and the graphical form of a logistic regression equation. In this equation, the $log[p/(1-p)] = \pi(x)$ is the likelihood of a project becoming an OTB, each X_n is a predictor variable and each β_n is the associated coefficient that fit for each predictor variable in the model. From the multiple models arrived at, the good models are selected based on three important criteria:

- The model's overall significance, given by a p-value associated with the chi-square statistic.
- The model's $r^2(U)$, a measure of the proportion of the total uncertainty that is attributed to the model fit for a logistic regression model (higher values are better).
- The area under the Receiver Operating Curve (ROC), which provides a measure of the model's ability to discriminate between those subjects who experience the outcome of interest versus those who do not (Hosmer and Lemeshow, 2000). It is to be noted that higher values of the corresponding statistic are better.

Table 1 presents Omnibus Tests of Model Coefficients and since the p-value is less than 0.05, the model fits the data.

Table 2 shows Model Summary provide the *best* model for ongoing development projects as the computed values of the statistics are high. The model captures the reasons why some of the project became an OTB.

Table 1. Omnibus Test of Model Coefficients

		Chi-square	df	Sig.(p-value)
Step 1	Step	1267.458	4	0
	Block	1267.458	4	0
	Model	1267.458	4	0

Table 2. Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	11.768[a]	0.745	0.996

The Classification results are shown in *Table 3*. The table shows that the model does an excellent job of correctly classifying Over Target Baseline (OTB) projects into two groups; those that development project will become an OTB and those that development project will not become an OTB, based on BCWS, ACWP, BCWP and EAC. Overall correct classification rate of the model is ***99.8%.***

Table 3. Classification Table

	Observed		Predicted		
			OTB		Percentage Correct
			0	1	
Step 1	OTB	0	505	0	100
		1	2	421	99.5
	Overall Percentage				***99.8***

Note:The cut value is.500

Table 4. Variables in the Equation

	Variables	B	S.E.	Wald	df	Sig. (p-value)	Exp(B)
Step 1[a]	BCWS	-.007	.017	.162	1	.687	.993
	ACWP	.140	.057	5.929	1	.015	1.150
	BCWP	-.140	.057	5.929	1	.015	.870
	EAC	.007	.017	.162	1	.687	1.007
	Constant	-2.385	.875	7.425	1	.006	.092

Variable(s) entered: BCWS, ACWP, BCWP, EAC.

The *Table 4* shows the output of all the predictor variables in the equation. The significance of each variable is measured using Wald statistic (Hair et. al 1998). By using p-value greater than or equal to 0.05 (5% level of significance) as a cutoff criterion for not including variables in the equation, it can be seen that BCWS (p=0.687) and EAC (p=0.687) do not appear to be important predictor variables. ACWP (p=0.015) and BCWP (p=0.015) significantly contribute to the dependent variable. The ACWP log odds ratio is 1.150 and this implies that the estimated odd of the development projects has to be improved by 15% of Actual cost work performed and BCWP log odds ratio is decreased by 0.870 for every additional hours. The results obtained for the development model show that when we predict an OTB, we are accurate the most of the time.

5. CONCLUSIONS

Estimate time and cost are important factors associated with the software development projects in the of Information Technology (IT) Companies. These two factors lead to the success or failure of a project. Providing a project to a customer on time and within budgeted cost are important factors to manage and control the project. In dealing with a complex task of project management, modification of the baseline project schedule, during execution, becomes mandatory. *Earned Value Analysis (EVA)* is one of the well-known methods of *performance measurement* of a project during its development. This paper attempts to develop a model to predict *Over Target Baseline* and *Estimate at completion* of software projects.

Based on our analysis, the development model indicates that the projects with high Budgeted Cost of Work Performed (BCWP) and low Actual Cost of Work Performed (ACWP) are more likely to influence Over Target Baseline (OTB). The present results indicated that the development projects have to improve by 15% of work process from the current level of development stage. In doing so, it is possible to complete the deliverables on time and within budgeted cost. It is to be noted that, currently the some of the development projects are taking more time to complete the actual work.

6. REFERENCES

[1] Air Force Cost Analysis Agency, 2007, Air Force Cost Analysis Handbook, Washington.

[2] Anbari, F. T., 2003, Earned value project management method and extensions, Project Management Journal, Vol.34, No.4, pp.12–23.

[3] Bembers, I., Boord, M., Byrnes, T. A., Finefield, T., Gran, W., Haupt, E., 2003, Over Target Baseline and Over Target Schedule Handbook, Fort Belvoir: Defense Acquisition University.

[4] CDC Unified Process Practices Guide: Earned Value Management, 2006, http://www2.cdc.gov/cdcup/ library/ practices_guides/CDC_UP_Earned_Value_Practices_Guide.pdf.

[5] Christensen, D. S., 1994, A review of C/SCSC literature, Project Management Journal, Vol.25, No.3, pp.32–39.

[6] Christensen, D. S., Antolini, R. D., and McKinney, J. W., 1995, A Review of Estimate at Completion Research, Journal of Cost Analysis, Issue 2, pp.41-62.

[7] Cukr, A., 2001, When is an Over Target Baseline (OTB) Necessary?, The Measurable News, March 2001.

[8] Defense Federal Acquisition Regulation Supplement: Earned Value Management System, 2009, Subpart 234.2, Department of Defense, Washington.

[9] Over Target Baseline and Over Target Schedule Handbook, 2003, Defense Acquisition University.

[10] Fleming, Q.W., and Koppelman, J. M., 2003, What's your project's real price tag? Harvard Business Review, Vol.81, No.9, pp.20–22.

[11] Fleming, Q.W., and Koppelman, J.M., 2005. Earned Value Project Management, 3rd edition. Newton Square, PA, Project Management Institute (PMI). ISBN 1-93069989-1.

[12] Fleming, Q.W., & Koppelman, J. M., 2010. Earned value project management (4th ed.), Newtown Square, PA, Project Management Institute.

[13] GSA FAR Secretariat, 2009, Federal Acquisition Regulation: Earned Value Management System, Subpart 34.2, Washington.

[14] Hair, J.F., Tatham, R.L., Anderson, R.E. and Black, W.C., 1998, Multivariate Data Analysis, 5th Edition, Prentice Hall, New Jersey.

[15] Hosmer, D.W., and Lemeshow, S., 2000, Applied Logistic Regression, John Wiley & Sons, New York.

[16] Kim, E. H.,Wells, W. G., Jr., and Duffey, M. R., 2003, A model for effective implementation of earned value management methodology. International Journal of Project Management, Vol.21, No.5, pp.375–382.

[17] Kristine, E. T., 2010, Predicting Over Target Baseline (OTB) Acquisition Contracts, Air Force Institute of Technology, Wright-Patterson Air Force Base, Ohio.

[18] Kristine, E. T., and Edward D. W., 2010, Predicting Over Target Baselines (OTBs), The Measurable News, The Magazine of the PMI, Issue 4.

[19] Ricardo,V.V., 2003, Earned Value Analysis in the Control of Projects: Success or Failure?, AACE Transactions, CSC.21.1.

[20] Seyed H.I., and Mansoureh, Z., 2008, Application of Artificial Neural Network to Forecast Actual Cost of a

Project to Improve Earned Value Management System, World Academy of Science, Engineering and Technology, Vol.42, pp.210-213.

[21] Seyed, H.I., and Zahra, M., 2008, Application of Data Mining Tools to Predicate Completion Time of a Project, World Academy of Science, Engineering and Technology, Vol.42, pp.204-209.

[22] Stephan V., Mario, V., 2006, A comparison of different project duration forecasting methods using earned value metrics, International Journal of Project Management, Vol.24, pp.289-302

[23] Suketu Nagrecha, 2002, An Introduction to Earned Value Analysis, A Report in http://www.pmiglc.org/COMM/ Articles/0410_nagrecha_eva-3.pdf.

[24] Tracy, S.P., 2005, Estimate at Completion: A Regression Approach to Earned Value. MS Thesis, AFIT/GCA/ENC/05-04, Graduate School of Engineering and Management, Air Force Institute of Technology (AU), Wright Patterson AFB OH.

[25] Trahan, E. 2009, An Evaluation of Growth Models as Predictive Tools for Estimates at Completion (EAC) MS Thesis, AFIT/GFA/ENC/09-01, Graduate School of Engineering and Management, Air Force Institute of Technology, Wright Patterson AFB OH.