MCR CRITICAL THINKING. SOLUTIONS DELIVERED.

Software Made Simple: Effort Adjustment Factors and the Accuracy of the Estimate

June 13, 2018

Jeremy Goucher
Space Solutions

Introduction

- Individual software development projects for DON can be greater than $1B each
  - GOA stated in a 2013 report that "IT projects too frequently incur cost overruns and schedule slippages…"
  - The same report stated that VistA-FM, GCSS-Army, GCCS-MC, and JWST reported a combined total of $1.3B in cost overruns largely due to poor cost controls (Powner, 2013)
- Nearly all modern development projects incorporate some software development requirements
- There is a need for better methods to estimate the cost of software development projects

Powner, D. A. (2013). *Information Technology: OMB and Agencies Need to More Effectively Implement Major Initiatives to Save Billions of Dollars.* District of Columbia: United States Government Accountability Office. Retrieved from GAO.gov: https://www.gao.gov/assets/660/656191.pdf

Data Set

- 33 Department of Defense programs spanning 14 development years from 2001 to 2014
- 212 total computer software configuration items (CSCIs)
- New, modified, reused, auto-generated code and total hours reported
- Both initial and final data reported
- Project sizes, measured in hours, between 22k and 3,000k required final hours
- Approximately 50% new and 50% upgrade efforts
- Includes all modern coding languages

Basic ESLOC Method

ESLOC Equation

- Equivalent source lines of code (ESLOC) are computed using effort adjustment factors (EAFs) in order to normalize the software size
- Normalization is needed because new code takes longer to develop than modified code which takes longer than reuse code
- The simplest ESLOC equation is of the form

$$EstESLOC = EAF_{New} * New\ SLOC + EAF_{Mod} * Modified\ SLOC + EAF_{Reuse} * Reuse\ SLOC$$

Growth and Productivity Rates

- Growth rates are needed to account for estimating error within the ESLOC estimate
- Productivity rates are needed to convert ESLOC to hours

$$GrowthRate = \frac{ESLOC_{final}}{ESLOC_{initial}}$$

$$Productivity = \frac{ESLOC_{final}}{Hours_{final}}$$

Computing Required Hours

- Mean growth and productivity rates are computed from a historic data set
- Estimated ESLOC is computed for the new project based on engineering assessment
- Estimated hours are computed as follows

$$EstHours = \frac{EstESLOC * E[GrowthRate]}{E[Productivity]}$$

EAF Error

Error Equation

- EAFs are generally subjective in nature and carry bias error
- Suppose there were two sets of EAFs denoted l and k
- Further, suppose l represents the true, but unknowable set of EAFs and k represents the set used by the estimator
- The error associated with the estimate is measured as follows

$$EAFerror = EstHours_l - EstHours_k$$

$$= \frac{EstESLOC_l * E[Growth_l]}{E[Productivity_l]} - \frac{EstESLOC_k * E[Growth_k]}{E[Productivity_k]}$$

Error Equation (Cont'd)

- Growth and productivity rates are quotients
- Expectation operator does not distribute into a quotient
- Result - the error equation cannot be simplified

However…

Assumptions and Inequalities

- With some assumptions, a set of inequalities can be developed
- For
  - $ModEAF_k > ModEAF_l$
  - $ReuseEAF_k > ReuseEAF_l$
  - $NewEAF = 1 \; for \; all \; cases$
- Assuming
  - $EstNewSLOC \cong E[NewSLOC]$
  - $EstModSLOC \cong E[ModSLOC]$
  - $EstReuseSLOC \cong E[ReuseSLOC]$
- Implying
  - $EstESLOC \cong E\left[ESLOC_{final}\right]$
- It can be shown that
  - $EstESLOC_k > EstESLOC_l$
  - $Productivity_k > Productivity_l$
  - $\frac{EstESLOC_k}{E[Productivity]_k} > \frac{EstESLOC_l}{E[Productivity]_l}$

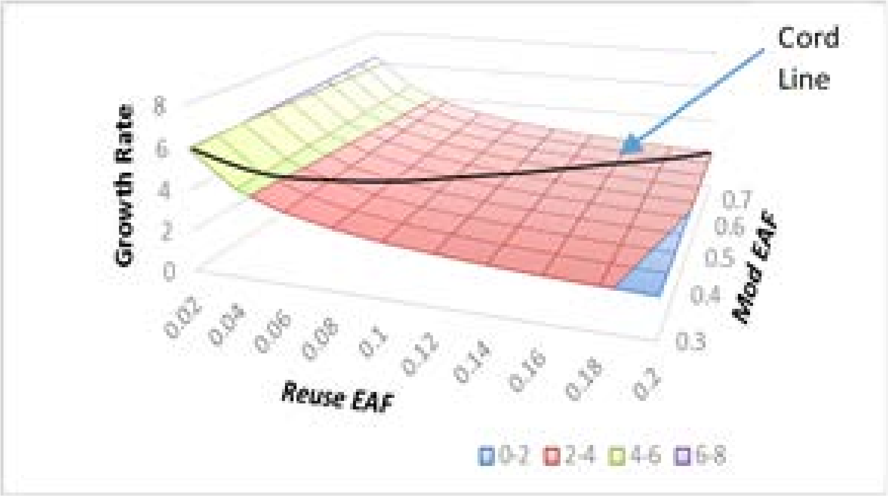
But…

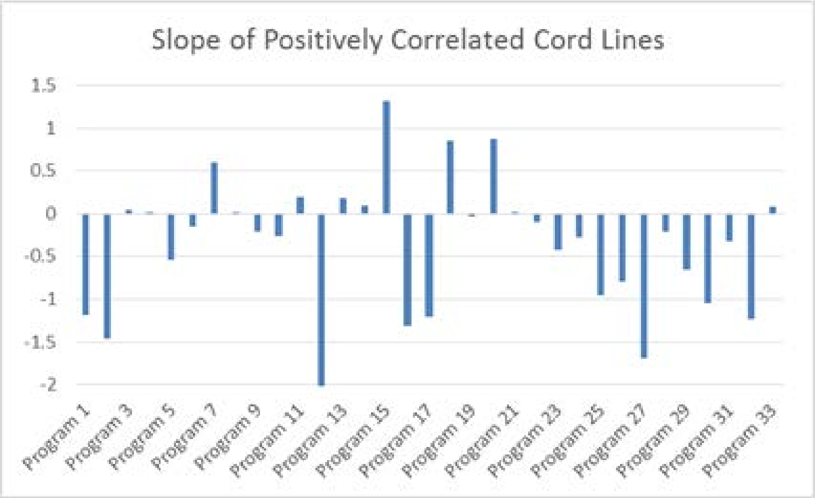
- The growth rate is the quotient of final ESLOC over initial ESLOC
- Final ESLOC and initial ESLOC are independent variables
- Therefor there is no way to know if the quotient is increasing or decreasing

Growth Rate Change

- To assess changes in the growth rate due to changes in the assumed EAFs, a series of surface plots was created for each program in the data set
- A “cord line” was drawn along the diagonal of positively correlated EAFs
- The slope of the cord line was measured for each program and the slopes plotted

Growth Rate Change (Cont'd)





Linear approximation used for nonlinear plots

- The slope of the cord lines varies in both magnitude and direction

Delta Equation Conclusion

$$EAFerror$$

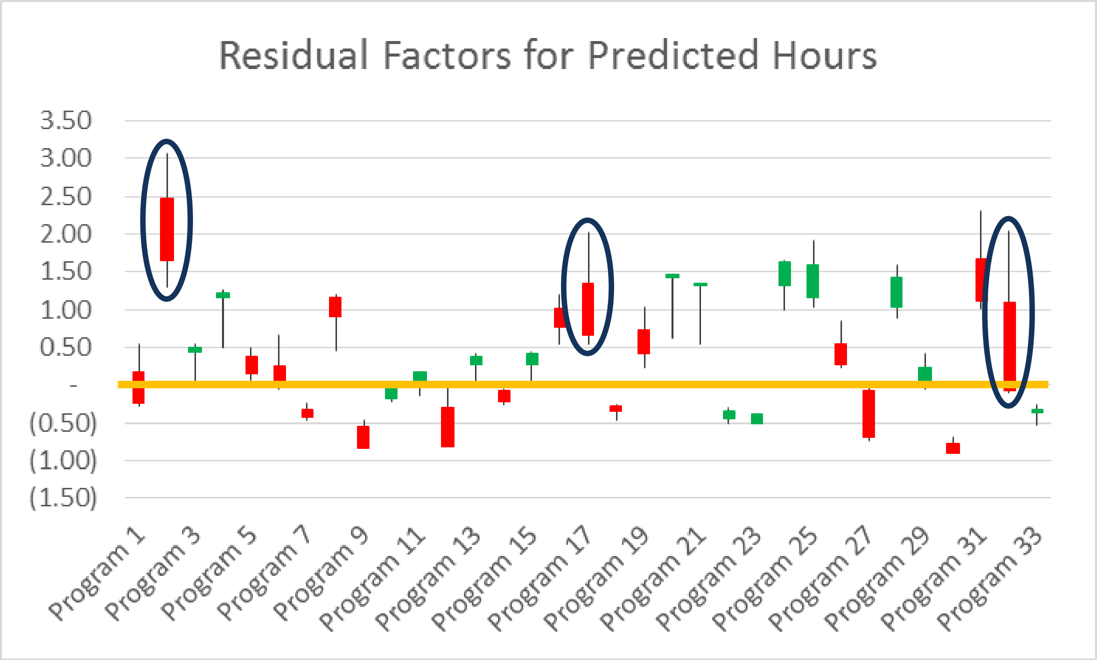
$$= \frac{EstESLOC_l * E[Growth_l]}{E[Productivity_l]} - \frac{EstESLOC_k * E[Growth_k]}{E[Productivity_k]}$$

- With a few assumptions, it can be shown that the ratio of EstESLOC to E[Productivity] is increasing when the EAFs are increasing
- There is no way to determine if the growth rate is increasing
- Therefore we cannot determine if the EAF error equation is positive or negative or attempt to assess the magnitude
- If the EAFs are moving in opposite directions, that is the mod EAF is understated while the reuse EAF is overstated, even less information is known about the error

EAF Error within the Sample

- Since the equations are indeterminate, residual error was measured for the data set described using five different sets of assumed EAFs
  - Low Mod, Low Reuse $\{ModEAF = 0.3, ReuseEAF = 0.02\}$
  - Medium Mod, Medium Reuse $\{0.5, 0.1\}$
  - High Mod, High Reuse $\{0.75, 0.2\}$
  - Low Mod, High Reuse $\{0.3, 0.2\}$
  - High Mod, Low Reuse $\{0.75, 0.02\}$
- The New EAF is equal to 1 in all cases
- Residual factors were computed by dividing the residual hours by the actual hours for each program
- The residual factors were graphed using a candlestick chart

Residual Factors Plot



- The highest and lowest points on the “wicks” represent the highest and lowest residual factors
- The highest and lowest points on the “bar” represent the second highest and second lowest residual factors
- Red bars imply the residual factor and the EAF values are negatively correlated; green bars imply they are positively correlated

Residual Factors Results

- The plot of residual factors demonstrates that the wrong choice of EAFs can cause estimates to be overstated or understated by as much as 300%
- More than half the programs assessed would likely have under or overstated the hours required by more than 50%
- The three circled programs on the previous plot have particularly large variance as a result of varying EAFs

The Regression Method

Regression Method

- The basic ESLOC method presented has high residuals, significant subjectivity, and cannot mathematically assess sensitivity or risk
- A regression method will
  - eliminate much of the estimate subjectivity
  - have the same or better residuals
  - have mathematically determined sensitivity parameters
- The proposed model is a multivariate linear regression of hours on SLOC types as described below

$$Hours = \beta_0 + \beta_1 * NewSLOC + \beta_2 * ModSLOC + \beta_3 * ReuseSLOC + \varepsilon$$

Regression Method (Cont'd)

- Regression uses final SLOC and final hours data only
- Growth rate is calculated based on initial versus final hours, rather than initial versus final ESLOC
- Complete model for total estimated hours is

$$\widehat{Hours} = \widehat{\beta_0} + \widehat{\beta_1} * NewSLOC + \widehat{\beta_2} * ModSLOC + \widehat{\beta_3} * ReuseSLOC$$

$$Growth = \gamma = E\left[\frac{Hours_{initial}}{Hours_{final}}\right] \sim N(\mu_H, \sigma_H^2)$$

$$FinalHours = \widehat{Hours} * \gamma$$

Regression Results

Model Form:	Weighted Linear model
Number of Observations Used:	33
Equation in Unit Space:	Hours = 2.255e+004 + 1.173 * New + 0.3617 * Mod + (-0.03106) * Reuse
Error Term:	MUPE (Minimum-Unbiased-Percentage Error)

Coefficient Statistics Summary

Variable	Coefficient	Std Dev of Coef	Beta Value	T-Statistic (Coef/SD)	P-Value	Prob Not Zero
Intercept	22547.7943	15474.2763		1.4571	0.1558	0.8442
New	1.1731	0.1573	0.8035	7.4574	0.0000	1.0000
Mod	0.3617	0.1877	0.2068	1.9270	0.0637	0.9363
Reuse	-0.0311	0.0190	-0.1787	-1.6376	0.1122	0.8878

Analysis of Variance

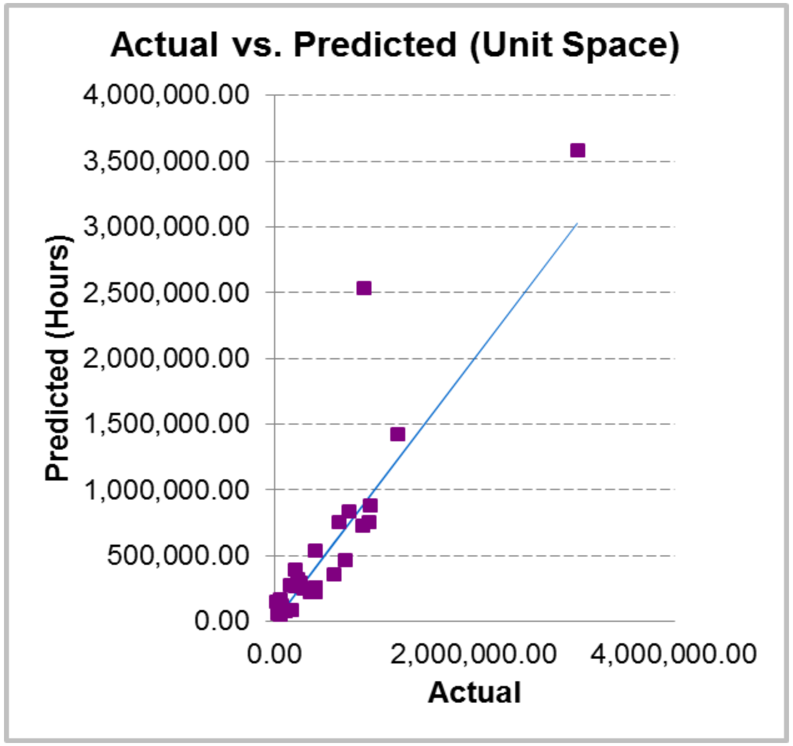
Due To	DF	Sum of Sqr (SS)	Mean SQ = SS/DF	F-Stat	P-Value	Prob Not Zero
Regression	3	14.8991	4.9664	21.4987	0.0000	1.0000
Residual (Error)	29	6.6992	0.2310			
Total	32	21.5983				

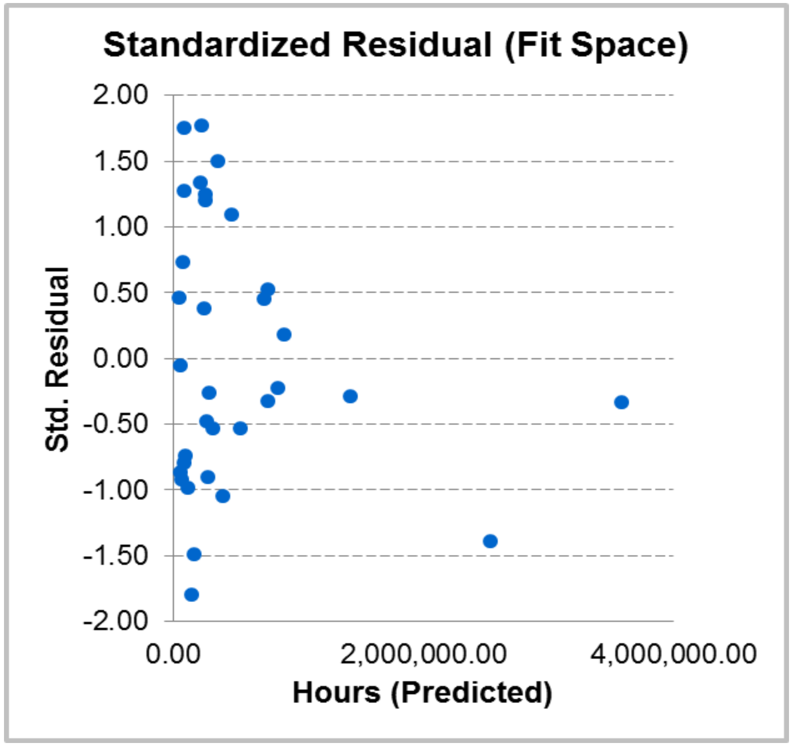
Goodness-of-Fit Statistics

Std Error (SE)	R-Squared	R-Squared (Adj)	Pearson's Corr Coef
0.4806	68.98%	65.77%	0.8306

Growth Rate (Hours Basis)		
Number of Observations	33	
μ_{γ}	σ_{γ}^{2}	$C.V._{\gamma}$
1.3699	.4671	.34

Regression Plots





ESLOC to Regression

- Regression method is actually a simplification of the ESLOC method

  - ○ $\widehat{Hours} = \widehat{\beta_0} + \widehat{\beta_1} * NewSLOC + \widehat{\beta_2} * ModSLOC + \widehat{\beta_3} * ReuseSLOC$
  - ○ Multiply by gamma to get final hours
  - ○ $FinalHours = \gamma * \widehat{Hours} = \gamma * \left(\widehat{\beta_0} + \widehat{\beta_1} * NewSLOC + \widehat{\beta_2} * ModSLOC + \widehat{\beta_3} * ReuseSLOC\right)$
  - ○ Multiply both sides by expected Productivity
  - ○ $\gamma * \widehat{Hours} * E[Productivity] = \gamma * \left(\widehat{\beta_0} + \widehat{\beta_1} * NewSLOC + \widehat{\beta_2} * ModSLOC + \widehat{\beta_3} * ReuseSLOC\right) * E[Productivity]$
  - ○ $E[Productivity] = E\left[\frac{ESLOC_{final}}{Hours_{final}}\right] = E\left[\frac{NewCode}{Hours_{final}}\right] = \frac{1}{\widehat{\beta_1}}$
    - Recall that the purpose of the EAFs are to normalize code to represent the same effort required to develop a single line of new code

ESLOC to Regression (Cont’d)

- Substitute for productivity on right hand side
- $\gamma * \widehat{Hours} * E[Productivity] = \gamma * \frac{(\widehat{\beta_0}+\widehat{\beta_1}*NewSLOC+\widehat{\beta_2}*ModSLOC+\widehat{\beta_3}*ReuseSLOC)}{\widehat{\beta_1}}$
- Substitute the EAFs on the right hand side for each of the normalized coefficients being divided by the B1 coefficient
- $= \gamma * \left(\frac{\widehat{\beta_0}}{\widehat{\beta_1}} + NewEAF * NewSLOC + ModEAF * ModSLOC + ReuseEAF * ReuseSLOC\right)$
- Substitute EstESLOC on the right hand side
- $= \gamma * \left(\frac{\widehat{\beta_0}}{\widehat{\beta_1}} + EstESLOC\right)$
- Bring the left hand side back and divide both sides by productivity
- $\gamma * \widehat{Hours} = \frac{\gamma*\left(\frac{\widehat{\beta_0}}{\widehat{\beta_1}}+EstESLOC\right)}{E[Productivity]}$

Finally, consider the case where the regression model is forced through the origin, as the case with the ESLOC method. Then $\frac{\widehat{\beta_0}}{\widehat{\beta_1}} = 0$ and the desired result is achieved.

Summary

- The cost of software development is historically difficult to estimate
- Traditional ESLOC methods rely on subjective EAFs
- Regression methods remove a significant portion of subjectivity from the estimate
- Regression results are statistically significant
- With more data, specific regression models can be easily developed which focus on coding language type, development type (waterfall or agile), families of software products, or any other grouping of interest

Backup

Assumptions and Inequalities

- Suppose the k set is larger than the l set for mod and reuse EAFs

$$ModEAF_k > ModEAF_l$$
$$ReuseEAF_k > ReuseEAF_l$$
$$NewEAF = 1\ for\ all\ cases$$

- Suppose computed historic ESLOC is normally distributed with a mean very close the estimated ESLOC

$$EstESLOC \cong E\left[ESLOC_{final}\right]$$

Assumptions and Inequalities (Cont'd)

- Then $EstESLOC_k > EstESLOC_l$
  - All SLOC values are positive so a SLOC value times a larger factor produces a larger ESLOC value
- Also $Productivity_k > Productivity_l$
  - $E[ESLOC_k] > E[ESLOC_l]$ for the same reason $EstESLOC_k > EstESLOC_l$
  - Numerator is increasing while denominator, Hours, is staying the same

Ratio of EstESLOC to Productivity

- $EstHours = \frac{\uparrow EstESLOC * E[Growth]}{E[Productivity]\uparrow}$
- If EstESLOC is increasing at a rate faster than E[Productivity], then the ratio of EstESLOC to E[Productivity] is also increasing
- This assessment requires another assumption
  - $EstNewSLOC \cong E[NewSLOC]$
  - $EstModSLOC \cong E[ModSLOC]$
  - $EstReuseSLOC \cong E[ReuseSLOC]$
- Under the above assumption, not only will the estimated ESLOC and the mean ESLOC from the historic data set be approximately the same value, but they will behave similarly when EAFs are adjusted
- The result of the assumption is EstESLOC is increasing at a faster rate than productivity because productivity is the quotient of a number moving at about the same rate as EstESLOC and the denominator is fixed, so the quotient is moving at a fraction of the rate of EstESLOC

Growth Rate Change

- So far, under fairly restrictive assumptions, a case can be made for the ratio of EstESLOC to E[Productivity] to be increasing if $k > l$

- $EstHours = \frac{EstESLOC * E[Growth]}{E[Productivity]} = \boxed{\uparrow \frac{EstESLOC}{E[Productivity]}} * E[Growth]$

- If the growth rate is also increasing, then the delta equation is positive, implying an overstatement of the EAFs also overstates the estimated hours