Empirical Software Engineering, 7, 377–381, 2002.

Reply to comments by M. Jorgensen, on the paper: 'A Simulation Tool for Efficient Analogy Based Cost Estimation' by L. Angelis and I. Stamelos, Published in Empirical Software Engineering, 5, 35–68 (2000)

L. ANGELIS
I. STAMELOS
Department of Informatics, Aristotle University of Thessaloniki, Greece

In our paper (Angelis and Stamelos, 2000) we tried to enhance the well-known method of estimation by analogy (EbA) for predicting the effort of a software project, by adding the notion of confidence interval (CI). Specifically, we used variations of the statistical simulation method known as bootstrap in order to attach CIs to the point estimation obtained by EbA. We also proposed the bootstrap method as a tool for calibrating the parameters of EbA.

In Section 7 of (Angelis and Stamelos, 2000) we presented a comparative study of the bootstrap CIs for EbA and the CIs obtained by the ordinary linear regression methods using two well-known small data sets. Our main purpose was to show that the bootstrap intervals for EbA are quite 'reasonable' compared to those commonly produced by regression, as a measure of accuracy and not to provide some new method which is better than regression. It is obvious that complete comparison of two methods can be achieved only by theoretical tools or extensive simulation studies.

In his recent correspondence, Jorgensen made some interesting comments, regarding a problem in the comparison of the CIs between the two methods, so we feel that it is necessary to clarify some issues and provide some more results in order to complement our previous study. The main point of our discussion is the discrimination of the terms 'confidence interval' and 'prediction interval' obtained in ordinary regression analysis.

First of all, we must explain the meaning of two types of intervals produced by the regression procedure. Consider a linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{or} \quad \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1k} \\ 1 & X_{21} & \cdots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{nk} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

where $\mathbf{Y}$ is the $n \times 1$ vector of observations from the dependent variable, $\mathbf{X}$ is the $n \times (k + 1)$ matrix containing observations from the k independent variables, $\boldsymbol{\beta}$ the $(k+1) \times 1$ vector of the unknown parameters and $\boldsymbol{\varepsilon}$ is the vector of errors. Using the least squares method we obtain the estimate for vector $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}$.

It is known (see for example (Draper and Smith, 1981; Montgomery and Peck, 1992)) that when a prediction $\hat{Y}_0 = \mathbf{X}_0'\hat{\boldsymbol{\beta}}$ is made at a specific observation $\mathbf{X}_\mathbf{0}' = (1, X_{01}, \ldots, X_{0k})$, applying the fitted regression equation, this is an unbiased point estimator of $E(Y/\mathbf{X}_0)$, i.e. of the true mean value of Y at $\mathbf{X}_0$ but at the same time it is also the predicted value of an individual observation. Therefore, from the same prediction we can obtain two different $(1-a)100\%$ CIs:

(a) for the true mean value of Y at $\mathbf{X}_0$:

$$\hat{Y}_0 - t_{a/2,n-k-1}\sqrt{\hat{\sigma}^2\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0} \leq E(Y/\mathbf{X}_0) \leq \hat{Y}_0 + t_{a/2,n-k-1}\sqrt{\hat{\sigma}^2\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0}$$

and

(b) for the individual observation:

$$\hat{Y}_0 - t_{a/2,n-k-1}\sqrt{\hat{\sigma}^2(1 + \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0)}$$

$$\leq Y_0 \leq \hat{Y}_0 + t_{a/2,n-k-1}\sqrt{\hat{\sigma}^2(1 + \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0)}$$

where $\hat{\sigma}^2 = (\mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y})/(n-k-1)$ is the estimator of the errors' variance and $t_{a/2,n-k-1}$ is the upper $a/2$ percentage point of the t_{n-k-1} distribution.

Unfortunately, there is no standard terminology for confidence intervals in the literature or in statistical software. For example, in (Draper and Smith, 1981) the two intervals are called 'confidence limits for the true mean value of Y' and 'confidence interval for a new observation' while in the statistical package SPSS the term 'prediction interval' is used for both intervals but they are distinguished as 'mean' and 'individual'.

In our study, since we were interested in estimating the cost (or effort) of a single software project, we used the term 'confidence intervals for the estimation of the effort' which we thought quite appropriate to represent the interval for the new individual observation. We also believe that it is clear that the bootstrap techniques applied to the EbA method provide intervals for new observations (see p. 47 and Figure 3).

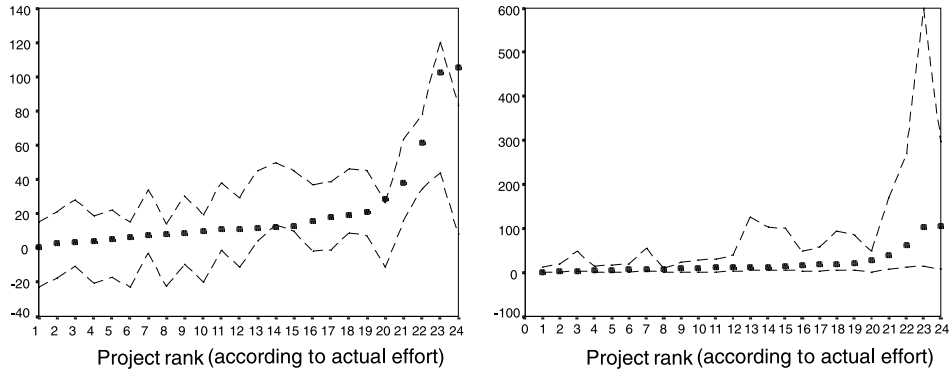
Regarding now the regression models we used in our analysis, the point estimations were accompanied by CIs computed by the S-plus statistical language (in fact all statistical analyses in the paper were developed in S-plus). Specifically, we used the S-plus function 'pointwise( )' with the description 'This function computes pointwise confidence limits for predictions…' assuming, falsely, that the intervals

Table 5a. Estimation and confidence intervals for the Albrecht data set.

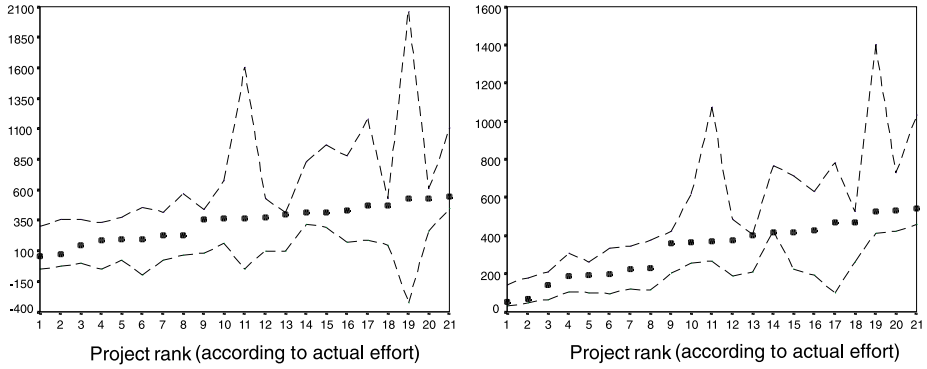
Rank	Actual effort	Estimation by analogy	95% CI non-parametric bootstrap	95% CI parametric bootstrap	Estimation by regression (original variables)	95% CI regression (original variables)	Estimation by regression (transformed variables)	95% CI by regression (transformed variables)
1	0.5	3.9	2.9–8.0	1.6–10.9	−3.8	−23.02–15.40	4.0	1.24–12.64
2	2.9	4.3	0.5–8.0	1.2–14.3	1.7	−17.45–20.92	4.4	0.98–19.39
3	3.6	8.1	5.1–10.6	3.4–34.9	8.8	−10.62–28.21	12.2	3.12–47.85
4	4.1	2.7	0.5–7.5	1.1–14.3	−1.1	−20.64–18.53	2.7	0.51–14.31
5	4.9	2.3	0.5–8.9	1.2–14.6	2.6	−16.95–22.06	3.9	0.85–17.41
6	6.1	6.8	2.3–10.0	1.5–16.6	−3.8	−22.81–15.13	3.9	0.80–19.05
7	7.5	18.5	13.6–21.1	3.2–34.4	15.6	−2.92–34.16	13.2	3.21–54.04
8	8.0	1.7	0.5–7.0	1.0–11.2	−4.2	−22.54–14.20	2.4	0.56–10.22
9	8.9	7.9	4.1–15.8	1.6–21.3	10.6	−9.42–30.52	5.0	1.13–22.39
10	10.0	4.9	3.6–7.1	2.0–23.5	−0.2	−19.86–19.50	5.3	1.06–27.01
11	10.8	12.4	4.9–15.8	2.2–25.5	18.3	−1.58–38.13	6.6	1.44–30.59
12	11.1	5.5	2.9–11.8	2.4–25.7	9.1	−11.13–29.36	7.6	1.49–39.07
13	11.8	15.6	12.9–28.8	4.8–52.8	24.8	4.21–45.29	23.4	4.39–124.68
14	12.0	20.9	3.6–38.1	5.6–65.0	31.7	13.50–49.82	23.0	5.16–102.87
15	12.9	20.1	9.7–21.1	5.6–54.5	27.5	9.63–45.30	24.2	5.85–100.27
16	15.8	8.2	7.5–19.0	2.6–33.6	17.7	−1.71–37.07	11.2	2.62–47.76
17	18.3	14.3	7.5–21.1	3.3–38.3	18.6	−1.33–38.53	12.0	2.49–58.09
18	19.0	17.0	7.5–21.1	4.9–55.1	27.4	8.75–45.95	21.9	5.14–92.93
19	21.1	16.0	7.5–19.0	5.0–48.4	26.2	7.09–45.26	19.8	4.61–85.32
20	28.8	15.1	9.7–18.3	2.6–29.5	7.6	−11.22–26.43	7.5	1.16–48.69
21	38.1	36.6	3.6–61.2	8.1–93.6	39.4	15.72–63.13	36.1	7.70–169.33
22	61.2	25.1	12.0–38.1	12.2–129.4	56.2	34.52–77.94	58.3	12.59–269.78
23	102.4	49.7	16.0–105.2	12.9–184.4	82.1	44.03–120.23	89.7	13.41–599.50
24	105.2	81.8	12.0–102.4	10.6–144.1	45.6	7.84–83.41	48.8	8.00–297.85
		MMRE: 73% Pred(25): 33%	INC: 58%	INC: 96%	MMRE: 103% Pred(25): 33%	INC: 87.5%	MMRE: 79% Pred(25): 25%	INC: 96%

Table 6a. Estimation and confidence intervals for the Abran–Robillard data set.

Rank	Actual effort	Estimation by analogy	95% CI non-parametric bootstrap	95% CI parametric bootstrap	Estimation by regression (original variables)	95% CI regression (original variables)	Estimation by regression (transformed variables)	95% CI by regression (transformed variables)
1	52	206	106–225	88–357	127	−49.10–303.94	76	30.66–142.04
2	69	98	52–298	85–333	165	−27.19–357.10	103	47.52–179.76
3	143	147	52–278	67–331	176	−3.33–354.65	126	63.90–207.69
4	187	124	52–313	68–342	140	−50.02–330.00	193	105.49–306.61
5	195	294	69–400	74–338	199	22.51–375.96	172	100.52–262.77
6	198	298	169–400	111–401	180	−96.26–456.13	198	96.23–334.90
7	225	98	52–191	61–332	220	21.26–417.84	217	118.75–344.57
8	229	397	286–417	113–459	317	67.55–565.73	225	113.12–374.24
9	360	272	132–400	114–426	260	82.41–437.66	303	202.24–423.64
10	363	286	195–400	114–444	419	160.10–678.18	418	256.62–617.67
11	369	380	134–400	158–539	775	−50.12–1600.06	603	266.02–1075.00
12	377	323	195–416	85–409	315	99.79–530.42	318	187.13–482.36
13	400	278	165–360	88–357	254	96.28–410.62	301	40.98–406.28
14	416	350	229–471	125–489	575	316.65–832.51	584	426.71–766.69
15	418	446	143–531	130–487	631	295.72–966.65	434	221.75–716.14
16	428	501	388–531	185–609	523	171.38–875.43	378	190.81 – 628.06
17	468	365	198–531	124–498	682	187.91–1176.87	359	98.93–780.57
18	471	388	251–531	118–441	339	149.45–527.93	382	262.44–524.14
19	525	445	307–471	140–486	866	−325.65–2058.29	834	413.70–1401.25
20	531	444	359–471	170–563	434	260.83 – 607.78	566	424.55–727.29
21	544	417	198–501	192–600	778	446.20–1109.42	716	457.46–1031.67
		MMRE: 40% Pred(25): 62%	INC: 67%	INC: 76%	MMRE: 43% Pred(25): 38%	INC: 100%	MMRE: 22% Pred(25): 71.4%	INC: 100%



Figures 10a and 11a. 95% Confidence zones for individual observations obtained by regression with the original variables (left) and the transformed variables (right) for the Albrecht data set.



Figures 15a and 16a. 95% Confidence zones for individual observations obtained by regression with the original variables (left) and the transformed variables (right) for the Abran–Robillard data set.

computed were referring to individual observations while they were actually referring to mean value, as correctly M. Jorgensen pointed out.

In the following tables and figures (Tables 5a and 6a, and Figures 10a, 11a, 15a and 16a which correspond to Tables 5 and 6 and Figures 10, 11, 15, 16 of the original paper (Angelis and Stamelos, 2000)) we give the CIs for new observations obtained by the regression models.

References

Angelis, L., and Stamelos, I. 2000. A simulation tool for efficient analogy based cost estimation, *Empirical Software Engineering* 5: 35–68.

Draper, N. R., and Smith, H. 1981. *Applied Regression Analysis*. 2nd Ed. John Wiley & Sons.

Montgomery, D. C., and Peck, E. A. 1992. *Introduction to Linear Regression Analysis*. 2nd Ed. John Wiley & Sons.