Part VII

Miscellaneous

Miscellaneous (Cases 93–94)

CASE 93

Estimation of Working Hours in Software Development

Abstract: This study involves the estimation of total working hours in software development. The objective was to determine the coefficients of the items used for the calculation of total working time. This method was developed by Genichi Taguchi and is called *experimental regression*.

1. Introduction

In many cases, office computer sales companies receive an order for software as well as one for hardware. Since at this point in time, functional software is not well developed, the gap between the contract fee and actual expense quite often tends to be significant. Therefore, if the accuracy in rough estimation (also called the *initial estimation*) of working hours based on tangible information at the point of making a contract were to be improved, we could expect a considerable benefit from the viewpoint of sales.

On the other hand, once a contract is completed, detailed functional design is started. After the functional design is completed, actual development is under way. In this phase, to establish a development organization, development schedule, and productivity indexes, higher accuracy in labor-hour estimation than the one for the initial estimation is required. Thus, it is regarded as practically significant to review the estimated number of development working hours after completion of the contract. This is called working-hour estimation for actual development.

2. Estimation of Actual Development Working Hours

For approximately 200 software products that have been sold by manufacturers and suppliers of office computers, we requested development engineers to answer a questionnaire that contains the following items:

1. *Number of files:* only files for system design are included; work and program files are excluded
2. *Number of items:* total number of items in a file
3. *Number of data input programs:* number of programs for data input or voucher issuance
4. *Number of table creation programs:* number of programs for documentation after data processing
5. *Number of renewal programs:* number of programs for balance renewal, repeated data processing, and simple data conversion
6. *Number of calculation programs:* number of programs for master data renewal through daily transaction and calculation
7. *Number of COBOL-based programs:* number of COBOL-based programs
8. *Number of all files used:* number of files accessed from all programs
9. *Use of existing designs:* designs of existing similar systems are used or not
10. *Work experience as software engineer* (*SE*): SE in charge of system design has work experience in developing similar systems or not

11. *Familiarity with operating systems as SE:* SE in charge of system design has knowledge about or work experience in operating systems that developed system rests upon
12. *Number of subsystems:* transaction consisting of approximately 10 documents related to sales, purchase, and inventory, etc.
13. *Use of special devices:* use of devices other than printer, cathode-ray tube, keyboards, disk, and floppy disk
14. *Development manpower:* total time needed for system development

Assumed Functional Equation and Initial Values

In the case of estimating actual development working hours after understanding the entire system, items 1 to 13 are considered explanatory variables, whereas item 14 is regarded as an objective variable. That is, assuming that the development labor hours result in zero when all items are zero, the equation can be expressed as follows:

$$y = a_1x_1 + a_2x_2 + \cdots + a_{13}x_{13} \quad (1)$$

This is a linear proportional regression equation.

Although 200 questionnaires were collected, by excluding missing and incomplete answers or exceptions, such as "a development cycle was doubled because it was redesigned after its full operation," the number of effective questionnaire results was 54. After obtaining a rough calculation result and using actual data, we realized that some cases have large gaps between actual data and estimations. To determine whether there are special reasons for this gap, by asking those who answered questionnaires about development situations in more detail and excluding items with a clear reason for the large gap, we still had 37 useful questionnaires. Several of the major special reasons were incomplete data, doubled development time because of redesign after starting, and too-specific orders from customers. But the majority of the cause was due to legacy availability and use of existing systems, and inability to calculate actual development working hours due to lack of records to trace back overtime work, and so on. For

Table 1
Initial values of coefficients in equation of actual labor-hour estimation

		Initial Value of Coefficient a_i for Level:		
i	Variable x_i	1	2	3
1	Number of files	0	3	6
2	Number of items	0	1	2
3	Number of data input programs	0	10	20
4	Number of table creation programs	0	3	6
5	Number of renewal programs	0	2	4
6	Number of calculation programs	0	5	10
7	Number of COBOL-based programs	0	2	4
8	Number of all files used	0	1	2
9	Use of existing designs	−300	−150	0
10	Work experience as SE	0	25	50
11	Familiarity with operating system as SE	0	20	40
12	Number of subsystems	0	15	30
13	Use of special devices	0	30	60

these reasons, we estimated parameters with 37 cases. According to practical assumptions by relevant engineers, initial values were set as shown in Table 1.

Converged Value and Analysis

Table 2 shows the three levels of coefficients after the sixth convergence calculation. Table 3 illustrates a part of the deviations and sums of residuals squared after the sixth convergence calculation.

If the values at level 2 after the sixth calculation are used, the resulting regression equation is expressed as

$$\begin{aligned} y = {} & 3.188x_1 + 0.312x_2 - 5.625x_3 + 3.188x_4 \\ & + 1.75x_5 + 5.312x_6 + 2.125x_7 + 0.469x_8 \\ & - 159.375x_9 + 25.0x_{10} + 20.0x_{11} + 15.938x_{12} \\ & + 58.125x_{13} \end{aligned} \tag{2}$$

Setting the total of squares of actual working hours to S_T, we obtain the following value:

$$S_T = 12{,}734{,}085 \tag{3}$$

On the other hand, according to Table 3, the sum of the squared residuals, S_e, turns out to be

$$S_e = 681{,}703 \tag{4}$$

Therefore, the present contribution ρ (%) is

$$\rho = \left(1 - \frac{S_e}{S_T}\right)(100) = 94.6\% \tag{5}$$

Now, since a linear proportional regression equation was used, we defined a ratio to the total sum of squares as the contribution. In addition, the standard deviation of errors, σ, is

$$\sigma = \sqrt{\frac{S_e}{n}} = 135.73 \tag{6}$$

As a reference, we show the estimation equation based on a least squares method as

Table 2
Converged values after sixth calculation

	Level		
Coefficient	**1**	**2**	**3**
a_1	3.0	3.1875	3.375
a_2	0.28125	0.3125	0.34375
a_3	5.0	5.625	6.25
a_4	3.0	3.1875	3.375
a_5	1.5	1.75	2.0
a_6	5.0	5.3125	5.625
a_7	2.0	2.125	2.25
a_8	0.4375	0.46875	0.5
a_9	−168.75	−159.375	−150.0
a_{10}	23.4375	25.0	26.5625
a_{11}	18.75	20.0	21.25
a_{12}	15.0	15.9375	16.875
a_{13}	56.25	58.125	60.0

Table 3
Part of the deviations and sum of residuals squared

No.	Deviation	Sum of Residuals Squared	Combination												
1	43.3114871	734,657.989	1	1	1	1	1	1	1	1	1	1	1	1	1
2	0.460980390	681,703.051	2	2	2	2	2	2	2	2	2	2	2	2	1
3	−42.3895264	780,436.115	3	3	3	3	3	3	3	3	3	3	3	3	1
4	15.9880074	694,443.380	1	1	1	1	2	2	2	2	3	3	3	3	1
5	−3.93175610	667,459.864	2	2	2	2	3	3	3	3	1	1	1	1	1
⋮	⋮	⋮							⋮						
36	−8.90810745	678,690.578	3	2	3	1	2	1	2	3	1	1	2	3	3

$$y = 7.919x_1 + 0.00234x_2 - 7.654x_3 - 5.712x_4 - 4.1021x_5 + 0.984x_6 + 4.346x_7 + 2.8637x_8 - 224.3517x_9 + 92.6744x_{10} - 140.3562x_{11} + 9.1052x_{12} + 130.9149x_{13} \quad (7)$$

Although the error variation, contribution, and standard deviation of errors are 311,112.77, 97.6%, and 113.72, respectively, we considered this equation impractical because it contains unrealistic coefficients.

Figure 1 illustrates the correspondence between the estimation and actual result when using equation (2).

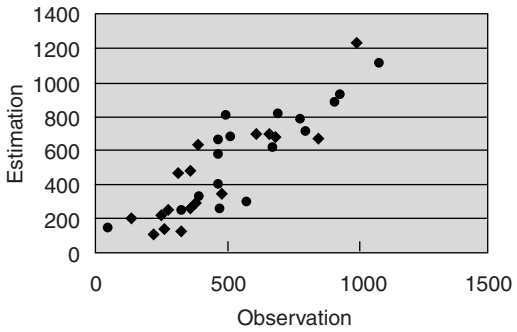

Figure 1
Estimation by observation and result using equation (2)

3. Rough Estimation

Comparing with the actual development working-hour estimation, the initial estimation at the time of receiving order is discussed below.

Selection of Items

As we stated at the beginning, we needed to determine a contract fee when receiving an order for software. However, most items were not determined at this point. More specifically, the only items that could be clarified were the (1) number of files, (2) number of table creation programs, (3) number of subsystems, and (4) number of special devices. Next, we studied whether we could roughly estimate the development working hours using only these four items.

Setup of Initial Values

A linear proportional regression equation based on the four items selected is assumed. Since all factors other than these four are represented by only these four, the initial values of the coefficients are set with wide intervals (Table 4). Now, the suffixes of the coefficients follow those in the estimation equation of actual development working hours.

Convergence Result and Evaluation

Because the number of unknowns is four, we used an L_{18} orthogonal array. The sequential approximation converged after six trials (Table 5).

Table 4
Initial values of coefficients for initial estimation equation

i	Variable x_i	Initial Value of Coefficient a_i at Level: 1	2	3
1	Number of files	0	15	30
4	Number of table creation programs	0	15	30
12	Number of subsystems	0	50	100
13	Use of special devices	0	50	100

Next we compared coefficients of common items in the equation for estimation of actual development working hours and one for initial estimation. In Table 6, we compare the values at level 2.

While the coefficients for number of subsystems and use of special devices are not so different, those for number of files and number of table creation programs are very distinct. The reason is that many items that are included from this analysis largely affect number of files and number of table creation programs. In other words, there is a correlation between the excluded items and each of number of files and number of table creation programs.

Table 5
Converged values of initial estimation equation after six trials

Coefficient	Level 1	2	3
a_1	15.0	15.46875	15.9375
a_4	11.25	11.71875	12.1875
a_{12}	12.5	14.0625	15.625
a_{13}	43.75	50.0	56.25

Table 6
Coefficients of common items in two equations

Coefficient	Development Working Hours: Actual	Estimated
Number of files	3.1875	15.46875
Number of table creation programs	3.1875	11.71875
Number of subsystems	15.9375	14.0625
Use of special devices	58.125	50.0

As the coefficients of the initial estimation equation, we adopted the values at level 2. The sum of squared residuals S_e results in 977,692.517 when the values at level 2 are used. The percent contribution was 92.4% and the standard deviation of errors was 162.55. Then, this is approximately 1.2 times as much as 135.73, the counterpart for the estimation equation of actual development working hours based on 13 items.

In a conventional process, most companies have determined a contract fee according to only the numbers of different screen displays and documents. Thus, we calculated a regression equation based on them. Even in the case of using a general linear regression equation computed by the least squares method that minimizes errors [y = 216.68 + 16.9 (number of documents) + 9.82 (number of screen displays)], the resulting standard deviation of errors amounts to 256.33, which is 1.6 times that for the initial estimation equation and about twice as much as that for the actual estimation equation. In addition, since the regression equation grounded on actual labor-hour results was not used, the resulting standard deviation of errors would be much larger. Therefore, our initial estimation equation is worthwhile to adopt.

This case study is contributed by Yoshiko Yokoyama.