Study of influence of test case parameters on test suite optimality in order to automate the information systems design process

A S Artyukhova¹, S I Rodzin¹

¹ Federal State-Owned Autonomy Educational Establishment of Higher Education “Southern Federal University”, 44, Nekrasovsky lane, Taganrog, 347928, Russia

E-mail: antonina1992@gmail.com

Abstract. The paper presents data on the influence of test parameters like average execution time, number of runs, test’s severity and complexity, number of related bugs on test’s optimality to include a test set for regression testing of information system software components. The paper includes research using such methods as analysis of variance, correlation and regression analysis of test parameters. The studies have shown a degree of the influence of test’s parameters to test suite efficiency and optimality in order to perform regression testing. Solving the problem of creation of optimal test suites aimed at regression testing gives a chance to automate the information systems design process.

1. Introduction
Nowadays the agile software development has a leading position in the software development area. The feature of this kind of processes is a very short software life cycle. Often a software changes; as a result there is lack of resources for verification of the correct functionality of information systems. The verification of the correct functionality of the information system needs to be carried out through the entire software life cycle even after the production release. The regression testing becomes a great important kind of testing at this stage [1, 2].

The creation of the test set that allows checking the work of information system software components and integration with the minimum resource costs is an actual problem. So, the test suite should include test cases that allows verifying the maximum number of program modules, finding bugs in their work with the minimum costs of time and resources [3].

The paper contains the study of test attributes such as $x^{(1)}$ – average execution time, $x^{(2)}$ – number of runs, $x^{(3)}$ – test’s severity, $x^{(4)}$ – complexity and $x^{(5)}$ – number of related bugs. Research of the dependencies between test case efficiency during finding the bugs process in the information system and listed above attributes gives additional knowledge about test suites creation. In the future this knowledge can be used for optimization verification of the correct functionality of information systems.

2. Methods
Analysis of variance, correlation and regression analysis were used for research of dependency of test’s optimality for including it to test suite using attributes: average execution time, number of runs, its severity and complexity, number of related bugs. In this case the term ‘test’s optimality’ is treated as the
best possible variants in terms of meeting several criteria separately or together. The actual question is about selection of test parameters that characterize its optimality [3, 4].

The linear relationship of target variable $Y$—expected test’s optimality from five regressors ($x^{(1)}$–average execution time, $x^{(2)}$–number of runs, $x^{(3)}$–severity, $x^{(4)}$–complexity, $x^{(5)}$–number of related bugs) is researched in the paper.

3. The study of the influence
Step 1. The model of multiple linear regression takes the form:

$$Y_i = a_0 + a_1 x^{(1)}_i + a_2 x^{(2)}_i + a_3 x^{(3)}_i + a_4 x^{(4)}_i + a_5 x^{(5)}_i + \varepsilon_i ; i = 1, 2, ..., 52,$$

unobserved random variables $\varepsilon_i$ are independent (random effects on the target variable of uncontrolled factors) and have the same normal distribution $\varepsilon_i \sim N(0; \sigma_{ELR})$, or otherwise $Y_i$ is independent and has normal distribution [5];

$$Y_i = N(MY_i = a_0 + a_1 x^{(1)}_i + a_2 x^{(2)}_i + a_3 x^{(3)}_i + a_4 x^{(4)}_i + a_5 x^{(5)}_i; \sigma_{Yi} = \sigma_{ELR}).$$

Function $Y_i = MY_i x^{(1)}_i x^{(2)}_i x^{(3)}_i x^{(4)}_i x^{(5)}_i = a_0 + a_1 x^{(1)} + a_2 x^{(2)} + a_3 x^{(3)} + a_4 x^{(4)} + a_5 x^{(5)} ; \sigma_{Yi} = \sigma_{ELR}$ is called a multiple linear regression function.

Step 2. We need to find the correlation coefficient between variables that possibly affect test’s optimality. The result numbers show values of corresponding correlation coefficients ($r_{xy}$).

| Table 1. Correlation coefficients values |
|------------------------------------------|
| Regressors          | Correlation coefficients |
|---------------------|--------------------------|
| $x^{(1)}$           | -0.808381893             |
| $x^{(2)}$           | 0.09767698               |
| $x^{(3)}$           | -0.06136937              |
| $x^{(4)}$           | 0.754271463              |
| $x^{(5)}$           | -0.775608401             |

Step 3. In this step we calculate the matrix pairwise correlation. According to the table below we see that linear dependency between regressors $X^{(1)}$, $X^{(4)}$ and $X^{(3)}$ is the strongest (bold font).

| Table 2. The matrix pairwise correlation |
|------------------------------------------|
| $Y$  | $x^{(1)}$ | $x^{(2)}$ | $x^{(3)}$ | $x^{(4)}$ | $x^{(5)}$ |
|------|-----------|-----------|-----------|-----------|-----------|
| $Y$  | 1         |           |           |           |           |
| $x^{(1)}$ | -0.808381893 | 1         |           |           |           |
| $x^{(2)}$ | 0.09767698 | -0.105500018 | 1         |           |           |
| $x^{(3)}$ | -0.06136937 | 0.08390344 | -0.141444047 | 1         |           |
| $x^{(4)}$ | 0.754271463 | -0.833018247 | 0.137661092 | -0.147048395 | 1       |
| $x^{(5)}$ | -0.775608401 | 0.940458919 | -0.067763543 | 0.132281249 | -0.808100935 | 1       |

We calculate estimates $\hat{a}_0$, $\hat{a}_1$, $\hat{a}_2$, $\hat{a}_3$, $\hat{a}_4$, $\hat{a}_5$ and $s_{ELR}$ parameters of model linear regression. Estimates of linear function regression are

$$\hat{y}_i = \hat{a}_0 + \hat{a}_1 x^{(1)} + \hat{a}_2 x^{(2)} + \hat{a}_3 x^{(3)} + \hat{a}_4 x^{(4)} + \hat{a}_5 x^{(5)} = 61.97 - 3.37 x^{(1)} + 0.000384 x^{(2)} + 0.23 x^{(3)} + 0.085 x^{(4)} + 0.207 x^{(5)}.$$

Step 4. The next step is analysis of variance. In Table 3, the column «df» contains the degrees of freedom $m = 5$, $n - m = 46$, $n - 1 = 51$ of corresponding random values.
**Table 3. Analysis of variance.**

|                | df | SS         | MS      | F         | Significance F |
|----------------|----|------------|---------|-----------|---------------|
| Regression     | 5  | 2302.307616| 460.4615233 | 19.24388726 | 2.73596E-10   |
| Residual       | 46 | 1100.673153| 23.92767724  |            |               |
| Total          | 51 | 3402.980769|         |           |               |

\[ SS_{\text{Regression}} = \sum_{i=1}^{n} (\bar{Y}_i - \bar{Y})^2 \quad SS_{\text{Residual}} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \quad SS_{\text{Total}} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2. \]

Their values correspond to 2302.31, 1100.67 and 3402.98 and are provided in the «SS» column; «MS» column provides values of:

\[ MS_{\text{Regression}} = SS_{\text{Regression}} / m, \quad MS_{\text{Residual}} = SS_{\text{Residual}} / (n - m - 1), \quad \text{that correspond to } 460.46 \text{ and } 23.93. \]

\[ SS_{\text{Regression}} = n\sigma_y^2 R^2, \quad SS_{\text{Residual}} = n\sigma_y^2 (1 - R^2). \]

Hypothesis testing \( H_0: a_1 = a_2 = \ldots = a_m = 0 \) is based on analysis statistic:

\[ F_{m,n-m-1} = \frac{MS_{\text{Regression}}}{MS_{\text{Residual}}} = \frac{SS_{\text{Regression}}/m}{SS_{\text{Residual}}/(n - m - 1)} = \frac{R^2/m}{(1 - R^2)/(n - m - 1)} \]

that has (statistical assumption about truth \( H_0 \)) Fisher–Snedecor distribution with \( m \) and \( (n - m - 1) \) degrees of freedom. In this case the observed value of statistic \( F_{5, 46} = 19.24 \) is more than \( f_{0.05; 5, 46} = 2.4 \), so hypothesis \( H_0 \) is rejected with a 5% significance level.

Hypothesis \( H_0 \) can be tested with a different approach: if significance \( F \) (calculated significance level of hypothesis \( H_0 \)) is more than the selected significance level \( \alpha \) (at this case \( \alpha = 0.05 \)), then hypothesis \( H_0 \) is accepted (equation is statistically non-significant), but if significance \( F \) is less than \( \alpha \), hypothesis \( H_0 \) is rejected (equation is significant). For the current model significance \( F \) is \( 2.74 \times 10^{-10} \) — equation is significant.

The observed value of statistic \( F_{m,n-m-1} \) and calculated significance level of hypothesis \( H_0 \) were provided in the table «Analysis of variance» (columns «F» and «Significance F»).

Now let us start hypothesis testing of \( H_0^{(j)}: a_j = 0 \) with alternative hypotheses \( H_1^{(j)}: a_j \neq 0, j=1,2,3,4,5. \)

The value of «t-statistic» \( t_{n-m-1}^{(j)} = \frac{\bar{a}_j}{s_a} \) has Student's t-distribution with \( (n - m - 1) \) degrees of freedom in case hypothesis \( H_0^{(j)} \) is accepted. The hypothesis rejection area of \( H_0^{(j)} \) at the significance level \( \alpha \) is \( |t_{n-m-1}^{(j)}| > t_{\alpha,n-m-1}. \)

The value of test statistic \( t_{46}^{(1)} \) is \(-1.86\); the value of statistic \( t_{46}^{(2)} \) is \(0.077\); the value of statistic \( t_{46}^{(3)} \) is \(0.38\); the value of statistic \( t_{46}^{(4)} \) is \(1.71\), value of statistic \( t_{46}^{(5)} \) value of statistic \(-0.37\). If the critical region is \( f_{0.05; 46} = 2.0 \), then hypotheses \( H_0^{(1)}: a_1 = 0, H_0^{(2)}: a_2 = 0, H_0^{(3)}: a_3 = 0, H_0^{(4)}: a_4 = 0, H_0^{(5)}: a_5 = 0 \) are not rejected \((\bar{a}_1, \bar{a}_2, \bar{a}_3, \bar{a}_4, \bar{a}_5 \) are non-significant).

«P-value» is calculated significance levels of hypotheses \( H_0^{(j)} \), probabilities \( p_j = 2P \left( |t_{n-m-1}^{(j)}| \geq |t| \right) \). Hypothesis \( H_0^{(j)} \) is rejected with alternatives \( H_1^{(j)} \), if \( p_j < \alpha \).

Since \( p_1 = 1.403, p_2 = 0.009, p_3 = 0.586, p_4 = 0.061, p_5 = 1.718 \), then hypotheses \( H_0^{(1)}: a_1 = 0, H_0^{(2)}: a_2 = 0, H_0^{(3)}: a_3 = 0, H_0^{(4)}: a_4 = 0, H_0^{(5)}: a_5 = 0 \) are not rejected.

The results which were obtained in paragraphs 4 and 5 are systematized in table 4.
Table 4. The regression equation

| Step | Equation, t-statistic, P-value | $R^2$ | $R^2$ | $S_{ELR}$ | $\delta$ | $F_{0.05,m:n−m−1}$ |
|------|-------------------------------|-------|-------|-----------|---------|-------------------|
| 1    | $\hat{y}_1 = 61.97 - 3.5 x^{(1)} + 0.00038 x^{(2)} + \frac{0.23 x^{(1)}}{71.264} + \frac{0.085 x^{(4)}}{0.016} + \frac{0.207 x^{(5)}}{0.012} + 0.45$ | 0.74  | 0.71  | 4.38      | 4.6%    | 26.31             |
|      | ($47.21;76.73$)                |       |       |           |         | 2.417             |
|      | $\hat{y}_1 = 0.00127;0.000641$ |       |       |           |         |                   |
|      | $\hat{y}_1 = 0.588$            |       |       |           |         |                   |
|      | $\hat{y}_1 = 0.004$            |       |       |           |         |                   |
|      | $\hat{y}_1 = 62.93 - 3.52 x^{(1)} + 0.000383 x^{(2)} + \frac{0.082 x^{(4)}}{48.95;76.91} + \frac{0.206 x^{(5)}}{0.012}$ | 0.74  | 0.72  | 4.34      | 4.6%    | 33.41             |
|      | ($-7.110;0.073$)               |       |       |           |         | 2.570             |
|      | ($0.000128;0.000638$)          |       |       |           |         |                   |
|      | $\hat{y}_1 = 0.055$            |       |       |           |         |                   |
|      | $\hat{y}_1 = 0.004$            |       |       |           |         |                   |
| 3    | $\hat{y}_1 = 64.88 - 2.41 x^{(1)} + 0.000328 x^{(2)} + 0.079 x^{(4)}$ | 0.74  | 0.72  | 4.32      | 4.6%    | 44.90             |
|      | ($52.24;77.52$)                |       |       |           |         | 2.798             |
|      | $\hat{y}_1 = 0.00133;0.000524$ |       |       |           |         |                   |
|      | $\hat{y}_1 = 0.029;0.187$     |       |       |           |         |                   |
|      | ($-0.027;0.191$)               |       |       |           |         |                   |
|      | ($0.000128;0.000638$)          |       |       |           |         |                   |
|      | $\hat{y}_1 = 0.052$            |       |       |           |         |                   |
| 4    | $\hat{y}_1 = 73.68 - 3.22 x^{(1)} + 0.000349 x^{(2)}$ | 0.73  | 0.71  | 4.37      | 4.8%    | 64.70             |
|      | ($69.59;77.77$)                |       |       |           |         | 3.187             |
|      | $\hat{y}_1 = 0.00153;0.000544$ |       |       |           |         |                   |
|      | $\hat{y}_1 = 0.146$            |       |       |           |         |                   |
|      | ($0.000000015;0.000079$)       |       |       |           |         |                   |
|      | $\hat{y}_1 = -6.77$            |       |       |           |         |                   |
|      | $\hat{y}_1 = 3.58$             |       |       |           |         |                   |

5. Equations analysis is provided in table 4. The best equation is an equation that has been obtained at step 4 because the equation and its coefficients have statistical significance. Variables $X^{(1)}$ and $X^{(2)}$ are included in the equation, the observed linear relation between variables is low: $\hat{r}(X^{(1)};X^{(2)}) = 0.576$. According to the equation:

a. more than 70% of test’s optimality variance ($\bar{Y}$) are connected with linear influence $x^{(1)}$ and $x^{(2)}$ ($R^2 = 0.73$);

b. Number $\hat{y}_1$ is calculated by the equation; $\hat{y}_1$ is the evaluation of general mean of test’s optimality on conditions when regressors ($x^{(1)}$ and $x^{(2)}$) are fixed at specific levels, so $x^{(1)} = x^{(1)}_1$, $x^{(2)} = x^{(2)}_2$. For example evaluation of general mean of test’s optimality in case of regressors values on the first object is $\hat{y}_1 = 73.68 - 3.22 \cdot 1.9 + 0.000349 \cdot 16 = 84.8 = 73.43$. The real optimality of test case $y_1 = 74$, residual is $y_1 - \hat{y}_1 = 0.57$. If residuals $y_j - \hat{y}_j$ are positive then verification costs are more than average level and if residuals are negative then verification costs are low then average level. As example on the first object is $y_1 - \hat{y}_1 = 0.57$, but on the second is $y_j - \hat{y}_j = -2.25$.

c. The increasing $x^{(1)}$ on the one unit with fixed value $x^{(2)}$ is accompanied the greatest change of average optimality of test (decreasing on 3.22 units); the increasing $x^{(1)}$ on the one unit is accompanied maximum possible 95% probability change target variable (decreasing optimality on 4.18), so 95% interval evaluation parameters $a_1$ and $a_2$ are ($-4.18; -2.27$) and (0.000153; 0.000544);

d. Analysis of elasticity coefficient

$$E_{Y|x^{(1)}} = a_1 \frac{\chi^{(1)}}{y} = -3.22 \cdot \frac{2.38}{67.48} = -0.135$$

$$E_{Y|x^{(2)}} = a_2 \frac{\chi^{(2)}}{y} = 0.000349 \cdot \frac{8421.9}{67.48} = 0.044$$

shows that increasing $x^{(1)}$ by 1% (with fixed value $x^{(2)}$) is accompanied by the greatest percent change
in the average optimality of test — decreasing by 0.135%; increasing \( x^{(i)} \) by 1% is accompanied by the maximum possible (with 95% probability) percentage change in the optimality, its decreasing by 
\[
\left| -4.18 \frac{2.83}{67.48} \right| = 0.175.
\]

4. Conclusion
The paper contains the study of the influence of test parameters on its expected optimality with the purpose to test a set for verification of the correct functionality of information systems. The dependencies of test optimality on its parameters have been obtained in the study. It allows in the future starting verification of not just a specific information system but the whole class program products of the information systems. It will lead to unification and formalization of the information system verification process and give a chance to automate the information systems design process.

5. Acknowledgments
The reported study was funded by RFBR according to the research project № 19-07-00570.

References
[1] Black R 2006 *Critical Testing Processes: Plan, Prepare, Perform, Perfect* (M: LORI) p 544
[2] Kaner C, Folk J, Nguyen H Q 2001 *Testing Computer Software* (K: DiaSoft) p 544
[3] Artyukhova A S 2018 The research of test’s efficiency of different parameters *Izvestiya SFedU. Engineering Sciences*. vol 198(4) pp 210-223
[4] Rodzin S 2015 Design of test generator for embedded self-testing *Proc. of 2015 IEEE East-West Design and Test Symp., EWDTs 2015* art no 7493098 pp 343-346
[5] Kalinina V N and Soloviev V I 2010 *Kompyuternyy Praktikum po Priklaudnoy Statistike i Osnovam Ekonometriki: Uchebnoye Posobiye dlya Studentov Vuzov Vega-Info* p 140