



Technical White Paper

Use of Multivariate Regression Analysis and Simulation for Defect Prediction

Mahapatra Smarjit Sahoo
Project Manager
Sonata Software Limited

STATEMENT OF CONFIDENTIALITY

Information included in this document, in its entirety, is considered both confidential and proprietary to Sonata Software and may not be copied or disclosed to any other party without its prior written consent.

Abstract

This white paper discusses the use of Multivariate Regression Model, instead of the Empirical Model used earlier, for prediction of defects in a project. A large variation between the number of actual defects and predicted defects was the key reason behind moving from Empirical Model to the Multivariate Regression Model.

Once the Multivariate Regression Model was deduced, Monte Carlo Simulation technique was used to check its reliability. Using the simulation technique, the Deterministic Model was essentially converted into Stochastic Model.

The benefits of using Monte Carlo Simulation technique are also discussed in this white paper.

About the Author

M S Sahoo is working as a Project Manager with Sonata Software Ltd. He has around 7 years of experience in the Groupware domain. For the last two years, he has been working as a Project Manager, managing projects in the technology space of Lotus Notes, Oracle PL/SQL and Java. He has also been actively involved in the CMMI initiative at Sonata.

Table of Contents

| | |
|--|---|
| 1. Challenge..... | 1 |
| 2. Solution Delivered..... | 1 |
| 3. Result..... | 4 |
| 4. Problems with the Multivariate Regression Model | 5 |
| 5. Monte Carlo Simulation | 5 |
| 6. Benefits of Monte Carlo Simulation technique | 6 |
| 7. Recommendations | 6 |

1. Challenge

Sonata was developing office automation applications for the customer, a Fortune 500 manufacturing company, offering it a gamut of services, such as application development, enhancement and production support.

Typically, application development and large enhancements are considered as individual projects. On an average, Sonata executes around 10 to 12 projects, varying from 01 to 18 man-months, annually. The project processes are well-defined and mature.

In 2007, as part of the CMMI initiative, Sonata brought about some key process changes, one of them being the introduction of the concept of Defect Prediction in Process Performance Management (PPM). The objective was two-fold:

- To predict the total number of defects in a project.
- To project the distribution of defects over phases.

The idea was to use historical data to predict/estimate the number of defects in a project. The defects were then distributed over the 3 phases of the project i.e. Review and Unit Testing, System Testing, and User Acceptance Testing. Sonata also planned to monitor the latent defects in each phase.

2. Solution Delivered

Empirical Model

For Defect Prediction, Sonata came up with an Empirical Model, the formula for which is:

Predicted Defects = Effort Estimates * Defect Density.

Based on historical data, the Average/Mean Defect detected per Man-Day of Effort i.e. Defect Density (D) of the completed projects was calculated. For a new project, the number of defects was predicted/estimated by multiplying the Defect Density (D) with the Estimated Effort (E) i.e. $D * E$.

Hence, the formula for Empirical Model was somewhat similar to the formula given below:

Predicted Defects = Linear Function of Defect Density + constant.

This formula implied that the Empirical Model could be approximated as a Linear Regression Model (single variable), where the method of Least Squares had not been used to calculate the Constant and Coefficients of the Independent Variables.

- Intercept – It was the value of Predicted Defects when the Effort Estimate was 0.
- Coefficient of Independent Variable – It was the change in the value of Predicted Defects per unit change in the Effort Estimates.

More precisely, in this case, the Empirical Model approximated to a Linear Model, where the Constant (Intercept) was 0 and the Coefficient of Independent Variable was the Defect Density (Mean Defects/Man-Day of Effort).

Now, the problem before Sonata was that the Predicted Defects were way off the Actual Defects found. The explained portion of the Total Variation, i.e. sum of the squares of unexplained deviations between the Actual and the Predicted Defects, was low at 67.7%. This meant that the Coefficient of Determination was low. The problem was two-pronged; either the Empirical Model was not good enough to predict the defects or Effort Estimate, as the lone Explanatory Variable was not able to explain all the variation in defects. Hence, Sonata decided to opt for a Deterministic Model to explain the variation in the number of defects predicted and defects actually detected.

To overcome this problem, Sonata decided to use the Multivariate (Multiple Variable) Regression Model for Defect Prediction for a project. The steps followed to implement the same are as underlined below:

- (a) A brainstorming session was held to earmark the Explanatory/Independent Variables that would explain most of the variation in the Dependent Variable (Defects Detected). The Independent Variables were:
- Size of the project in terms of Lotus Notes Points.
 - Number of Trainees (T) involved in the project. This was a measure of the productivity.
 - Elapsed days for the project.
- (b) The historical data on Defects, Size, Number of Trainees (T) and Elapsed Days for the completed projects was used to generate the Regression Equation:

$$\text{Defects} = A + B * \text{Size} + C * T + D * \text{Elapsed Days}.$$

For this calculation, the Minitab software was used. The historical data and the result of the Regression Analysis are specified below:

| SIZE (LN Points) | No of Trainees | Elapsed Days | Defects Detected |
|------------------|----------------|--------------|------------------|
| 568 | 1 | 35 | 81 |
| 1396 | 1 | 91 | 177 |
| 371 | 0 | 25 | 56 |
| 813 | 1 | 33 | 97 |
| 312 | 0 | 30 | 43 |
| 771 | 2 | 16 | 87 |
| 346 | 1 | 11 | 40 |
| 1304 | 1 | 30 | 133 |
| 253 | 1 | 26 | 35 |
| 121 | 0 | 15 | 9 |
| 234 | 0 | 22 | 18 |
| 388 | 0 | 25 | 32 |

Regression Analysis: Defects versus SIZE, No of Trainees (T), Elapsed Days

The regression equation is

Defects = - 7.30 + 0.0884 SIZE + 9.10 T + 0.599 Elapsed Days

| Predictor | Coef | SE Coef | T | P |
|--------------|---------|---------|-------|-------|
| Constant | -7.302 | 4.893 | -1.49 | 0.174 |
| SIZE | 0.08836 | 0.01080 | 8.18 | 0.000 |
| T | 9.102 | 5.081 | 1.79 | 0.950 |
| Elapsed Days | 0.5994 | 0.1853 | 3.24 | 0.012 |

S = 8.38730 R-Sq = 97.9% R-Sq(adj) = 97.2%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|----------------|----|---------|--------|--------|-------|
| Regression | 3 | 26827.9 | 8942.6 | 127.12 | 0.000 |
| Residual Error | 8 | 562.8 | 70.3 | | |
| Total | 11 | 27390.7 | | | |

Interpretation of the result:

The formula of the Regression Model, as obtained from Minitab, is:

Defects = 7.30+0.0884 Size+ 9.10T+0.599 Elapsed Days.

R-Sq = Coefficient of Determination = 97.9%: This implied that 97.9% variation in the number of defects predicted and defects detected could be explained by the 3 Independent Variables considered above. Hence, the model looked a good criterion for the best-fit.

R-Sq (adj) = Adjusted Coefficient of Determination = 97.2%: As the R-sq (adj) was close to R-Sq, it was safe to say that there were no unnecessary terms in the Regression Model.

S = Standard Error of Estimate = 8.38%: This was a measure of dispersion of the sample points around the fitted model. In this case, the dispersion was high.

- (c) The hypothesis, that the 3 Explanatory/Independent Variables mentioned above were significant for Defects Detected (Dependent Variable), was put to test.
- The Null Hypothesis: The Explanatory Variables are not significant -- $B=0$, $C=0$ and $D=0$.
 - The Alternate Hypothesis: The Explanatory Variables are not significant -- $B \neq 0$, $C \neq 0$ and $D \neq 0$.

The significance level Sonata was looking at was 10% i.e. 0.1. The results of the Regression Analysis obtained from Minitab showed that for all the 3 Explanatory/Independent Variables, $p < 0.1$. Hence, the Null Hypothesis was rejected. This implied that the three Explanatory Variables were significant. The same result could have been achieved using the T-test.

- (d) It was examined whether the value R-Sq really indicated that the Explanatory Variables explained the variation in the number of Defects Predicted and Defects Detected or was it a coincidence.
- The Null Hypothesis: Defects detected do not depend on any of the 3 Explanatory Variables, $B=C=D=0$.
 - The Alternate Hypothesis: Defects detected depend on at least one of the 3 Explanatory Variables, either $B \neq 0$ or $C \neq 0$ or $D \neq 0$.

The significance level being looked at was 10% i.e. 0.1. The result of the Regression Analysis obtained from Minitab showed that for the Analysis of Variance (ANOVA), $p < 0.1$. Hence, the Null Hypothesis was rejected. This implied that Regression Analysis, as a whole, was quite significant. Using the F-test, the same result was achieved.

- (e) Multi-collinearity in Multivariate Regression was tested. For this purpose, simple Regression of Defects with each of the Explanatory/Independent Variable was determined. Also, the Regression of Defects, taking 2 of the 3 Explanatory Variables at a time, was determined. The following conclusion was reached on R-Sq:

| Regression | R-Sq |
|--------------------------------|--------|
| Defects vs. Size | 95.20% |
| Defects vs. T | 30.80% |
| Defects vs. Elapsed Days | 59.50% |
| Defects vs. Size, T | 95.30% |
| Defects vs. Size, Elapsed Days | 97.10% |
| Defects vs. T, Elapsed Days | 80.80% |

The table above clearly suggests that there is a strong correlation between the Size and Number of Trainees (T), and Size and Elapsed Days. But still the Number of Trainees and Elapsed Days were retained as Explanatory Variables.

- (f) Finally, the residuals were analyzed. The residuals were random and did not exhibit any non-random patterns. Hence, the Multivariate Regression Model emerged as a pretty good one for estimation/prediction of defects.

3. Result

Sonata arrived at a Multivariate Regression Model and performed many tests to authenticate its validity. The final Regression Equation was:

$$\text{Defects} = - 7.30 + 0.0884 \text{ Size} + 9.10 \text{ T} + 0.599 \text{ Elapsed Days.}$$

This Multivariate Regression Model could be used to predict the defects in a project as long as its Size, Number of Trainees and Elapsed Days were known.

4. Problems with the Multivariate Regression Model

There were, however, some inherent concerns Sonata had about this model. The historical data that was used to derive the model had only 12 data points. Hence, reliability was a concern. It was also not certain whether the model could reliably predict the number of defects. To address these issues, it was decided to run a Monte Carlo Simulation on the model.

5. Monte Carlo Simulation

Using Monte Carlo Simulation, the derived Multivariate Regression Model was iteratively evaluated, using Random Numbers as inputs for the Explanatory/Independent Variables (Uncertain Parameters). Microsoft Excel was used to execute the Monte Carlo Simulation. The steps followed are as detailed below:

- (a) The type of distribution to be considered for Uncertain Parameters was determined. For Size, Number of Trainees and Elapsed Days, a Uniform Distribution was chosen, which was then used to generate a set of random numbers for Uncertain Parameters. The formula used for generating Random Numbers is specified below:

$$\text{Min} + \text{Rand} () * (\text{Max} - \text{Min})$$

The table below shows a few Random Numbers generated for Uncertain Parameters.

| SIZE | No of Trainees | Elapsed Days |
|------|----------------|--------------|
| 2823 | 0 | 116 |
| 2080 | 2 | 103 |
| 753 | 1 | 83 |
| 2246 | 0 | 37 |
| 1155 | 0 | 45 |
| 2226 | 1 | 60 |

- (b) The defects for each set of Random Inputs were calculated using the Regression Model.
- (c) Steps 1 and 2 were performed for around 5,000 iterations.
- (d) The results were evaluated by following the steps given below:
- Creation and analysis of the histogram for $n = 5000$ and 40 bins. The following inferences were drawn:
 - Defects were always greater than 0, as expected.
 - The uncertainty in defects was large, varying from 13 to 520. The standard deviation needed to be examined to reach the final decision.
 - The distribution did not look like a perfect Normal Distribution.
 - Analysis of the summary statistics:

- The mean was very close to the median. Hence, the distribution of defects was symmetric.
- The standard deviation, which is a measure of the spread of defects, was 127.61. Though large, the spread was still reasonable as defects were to be detected for a fairly wide range of project size (100 to 5000 Lotus Notes Points). Hence, it was concluded that the variance was reasonable enough and consequently, it was safe to assume that the Multivariate Regression Model was also amply reliable.
- Calculation of the final statistic i.e. Confidence Interval of the true population mean. The UCL and LCL were 263.57 and 269.51 respectively. So it was highly expected that the population mean would fall between the UCL and LCL specified above.

Based on the historical data, a Deterministic Model (Multivariate Linear Regression Model) was deduced to predict the number of defects in a project. As the number of data points was less, the Monte Carlo Simulation technique was used to test the reliability of the model. It was found that the model was reasonably reliable and could be used to predict defects.

Recently, Sonata employed the Multivariate Regression Model, described above, internally to predict defects in a new project. Once the project was completed, the reliability of the model was gauged on the basis of the number of actual defects vis-a-vis the number of predicted defects (i.e. how big or small was the residue).

6. Benefits of Monte Carlo Simulation technique

The Monte Carlo Simulation technique has certain unique benefits for prediction of defects in a project, such as:

- It is very simple and easy to use. Once the model is finalized, prediction of defects for a new project would require information on Size, Productivity (Number of Trainees) and Elapsed Days. All this information is computed during the execution of the project. Hence the effort required for prediction of defects is less.
- It can be employed for projects with fairly matured and stable processes. Also, there is no need for any kind of Causal Analysis. E.g. If a few defects were found in a project during testing, would that imply that the testing was poor or that the development was outstanding and the software had few defects to find? The Regression Model cannot answer these questions.
- It uses the previous defect data to arrive at the Defect Prediction Model. So due importance is given to the data of previous similar projects to predict defects in the upcoming projects. Hence, historical data is important for using this Defect Prediction Technique.

7. Recommendations

Certain factors should be borne in mind while predicting defects in a project. These are:

Keep refining the Defect Prediction Model (run the regression) as more data points are found. This would help align the Prediction Model with the latest behavior and patterns of defects.

In case it is difficult to arrive at the key Explanatory Variables that explain the maximum variation in defects, techniques like Principal Component Analysis can be used to deduce them.

Use of multiple Explanatory Variables is recommended to predict defects in a new project. This is because sometimes defects cannot be predicted accurately using a single Explanatory Variable. Sonata employed a Multivariate Linear Regression Model to arrive at the best-fit Defect Prediction Model. However, sometimes a Non-Linear Regression Model (quadratic, exponential etc.) provides a better-fit model. This possibility should be explored as well, more so if the Linear Regression Model is not able to explain a significant portion of variance in the Dependent Variable. Analysis of residues would give a good insight on this.

The reliability of the model was tested using Monte Carlo Simulation. Typically, it is used to test the reliability, performance or sensitivity of a model due to random variations, lack of knowledge or error in uncertain parameters. It can be used often when the model in question is complex, non-linear or involves more than just a couple of uncertain parameters.

A drawback of the Multivariate Regression Model is the difficulty in its interpretation. So, a more comprehensible technique, i.e. Regression via Classification can be used for Defect Prediction.

For more information, contact info@sonata-software.com

Click here to know more about Sonata's [Product Quality Assurance Services](#)