UNIVERSITY OF APPLIED SCIENCES

FACHHOCHSCHULE
GIESSEN
FRIEDBERG

Friedberger Hochschulschriften

Manfred Börgens
Thomas Hemmerich
Ludwig B. Rüssel

Use of Discriminant Analysis in Forecasting the Success of a Software Development Project

Friedberger Hochschulschriften Nr. 4

Friedberger Hochschulschriften
Herausgeber:
Die Dekane der Fachbereiche des Bereichs Friedberg der FH Gießen-Friedberg
Wilhelm-Leuschner-Straße 13, D-61169 Friedberg
http://www.fh-friedberg.de

Friedberg 2000
ISSN 1439-1112

USE OF DISCRIMINANT ANALYSIS IN FORECASTING THE SUCCESS OF A SOFTWARE DEVELOPMENT PROJECT

M. Börgens, T. Hemmerich and L. B. Rüssel

Cost overruns are widespread in software development projects. Though various tools for risk analysis are in use, a manager still needs a simple and powerful method predicting success or failure of software projects. This paper proposes discriminant analysis as an appropriate tool. It has successfully been used for predicting bancruptcy and for credit contract classification. Discriminant analysis needs predictor variables for assigning a group classification (successful / failing) to each project. The data base for predictors and groups was established in a major German company by analyzing 107 software projects. The discriminant functions calculated can be used as a caveat against cost overruns as the reliability of failure predictions could be proven to be sufficiently high.

Keywords: Discriminant analysis, software projects, risk management, cost estimation

COST OVERRUNS IN DEVELOPMENT PROJECTS

Many high-technology development projects have been plagued by cost overruns and therefore constituted a loss for the contractor. Within that category, software development projects impose an even higher risk as cost estimation has proven to be more unreliable than in traditional hardware development. For senior management to decide whether to bid for a contract and commit to a fixed price or cost ceiling seems to be a task which comes close to gambling. At best the senior manager can rely on his experience and judge according to "proven" success or failure factors, such as who will be the project manager or whether the project develops in state of the art. What he actually needs is a statistically reliable and early indicator for deciding whether to bid that gives him an unbiased prediction of the

project outcome. This article describes the result of a case study in a German company applying discriminant analysis as a risk analysis tool for supporting senior management bidding decisions.

TOOLS FOR COST ESTIMATION AND RISK ANALYSIS

Various tools are applied for software cost estimation. Analogy models estimate software cost by comparing the cost of the program to be developed to a similar known program and adjust the cost of the known program by a factor accounting for the different size (source lines of code) of the programs. Regression-based models use equations derived from a regression analysis of a historical data base that relate cost to a set of known independent variables (input factors or predictors). A set of regression derived cost estimating relationships are the basis of parametric models which have been prevalent in estimating software cost. Many models can be calibrated to the specific environment of the user. Examples of parametric models are those well-known commercial models like COntructive COst MOdel (COCOMO) and PRICE-S [1].

These and other state of the art cost models represent valuable tools that supply a decision-maker with additional confidence that the company can commit to the price offered to the customer. However, the decision-maker is left alone if he is to judge the risk of the price being exceeded by actual cost. The point estimate does not give him any information about the reliability of the estimate. Several techniques for quantifying risk and revealing the underlying probability distribution of the cost estimate can be used in the pre-contract phase such as Probabilistic Event Analysis (PEA), Monte-Carlo-simulation for cost estimating relationships, stochastic decision trees or stochastic networks [2]. These techniques may lead to a better insight into the uncertainty of the estimate, but they all require plenty of input data (i.e., subjective probalitities), still leaving alone the senior manager with his judgement of the risk of the project. He rather needs a simple, yet powerful project distress model making predictions on the likelihood of a project failure, a failure to keep within the estimated cost margin, and classifying a project accordingly. Pugh has suggested the use of a multivariate statistical procedure known as discriminant analysis that has been previously applied in predicting corporate bancruptcy [3]. Altman has pioneered the use of discriminant analysis for financial distress [4], and his models are in widespread use today. In analogy to the success or bancruptcy of a company, projects can be

classified by applying a discriminant model. This paper derives such a model based on the statistical analysis of accomplished software development projects.

CLASSIFICATION OF SOFTWARE PROJECTS BY DISCRIMINANT ANALYSIS

Theoretical background. Discriminant analysis, in its first step, tries to find rules for optimally dividing an existing sample into different well-defined homogeneous groups. In the second step, these rules are used for the group classification of new members of the sample. The rules consist of an appropriate choice of "predictors" and their aggregation in linear combinations. So the predictors are independent variables (metric or dichotomous) in linear functions, the "discriminant functions". Their coefficients are weights linking the predictors to the dependent variable which indicates the group the new sample member belongs to, see [5], [6], [9], [10]. If there are only two groups, the number of functions is reduced to one (see table 1).

Discriminant function:

$$Y = a_0+a_1*X_1+a_2*X_2+...+a_n*X_n$$

Y = dependent variable indicating the group membership
a_i = coefficient for the i^{th} independent variable (i^{th} predictor), i = 1 ... n
X_i = i^{th} independent variable (i^{th} predictor), i = 1 ... n
a_0 = constant value

Table 1. General structure of the discriminant function (Two groups case)

There are several mathematical methods supporting the elimination of predictors which are obsolete because they contribute very little to the dependent variable. An important by-product of discriminant analysis is a table of probabilities indicating the reliability of the prediction (see the chapter RESULTS on pages 6 – 7 and tables 2 - 4).

Application to software projects. The *groups* have to be defined according to the company's notion of "success". We propose to define only two groups of projects: those which are sufficiently profitable to the company and those which are not. The *predictors* for the success of software projects may be, e.g.,:

- Experience of the manager in charge of the project
- Total amount of the order
- Number of software engineers involved
- Number of subcontractors

Before discriminant analysis can be used for prediction a (random or complete) sample of accomplished projects is needed in order to establish the discriminant function. It is derived by a simple (though mostly computer-assisted) algebraic procedure ([5], [6], [9], [10]) using as input the sample data of predictors and groups. For a first approach to an optimal result, it is advisable to take into account every reasonable (and data-based) predictor. Its specific coefficient will be shown by the discriminant function. The product of a coefficient and the standard deviation of the corresponding predictor in the sample is proportional to the weight the predictor contributes to the dependent variable "success".

To diminish computational complexity, one should try to keep only those predictors the elimination of which would substantially decrease the reliability of the analysis.

Once the discriminant function has been established, it is an easy task to use it for predictions. A software developing company bidding for a contract has to estimate the total project cost and to plan human and material resources before starting negotiations. In this phase, data for the predictors should be collected. They represent the input of the discriminant function, the output of which is a single number. Positive numbers forecast a success of the project, negative numbers a failure.

Case study. A subsidiary of a diversified German high-technology company has a major department for software engineering. Its software is required by external customers as well as by other internal departments. The company tries to reduce the risk of decision in the bidding and negotiating phase by using discriminant analysis [7]. Two samples were used for establishing the groups and predictors. The first one comprised all 56 software projects finished in a 10-year interval with substantial deviations between estimated and real cost and with complete data about eight previously selected predictors. The second sample comprised all 107 software

projects finished in the same time interval with complete data about six predictors chosen out of the former eight and was composed of the first sample and 51 additional projects with only small cost deviations.

The main task: Choice of predictors. In general, the main problem before starting a discriminant analysis is to get a sufficient data base and an adequate number of predictors in order to create a discriminant function which ensures a prediction probability of more than 50%. The first discriminant analysis started with the following eight predictors which - by management experience and by rough inspection of the sample data - promised to be a good base for classifying future software projects.

- Total value of the order
- Difficult versus non-difficult contract negotiations
- Government sector customer versus private customer
- External versus internal order
- Experience of project manager
- Run-time of contract
- Project managers' current occupation with other projects
- Number of subcontractors

For these predictors, complete data were available in the first sample of 56 software projects. The last two predictors were omitted from the second sample of 107 projects.

DATA EVALUATION

For the computation of the discriminant functions and its statistical by-products, the statistical software package SPSS [8] was used. It applies "linear" discriminant analysis, a parametric decision procedure with rather strong prerequisites concerning quality of data, namely, normally distributed predictor variables with identical covariance matrices. These assumptions are hard to prove and probably often violated by data collected in a company environment like the present one. Fortunately, linear discriminant analysis has been shown to be robust against violations of these conditions [5], [6]. Hence the results will not suffer substantially from suboptimal data conditions.

RESULTS

Three discriminant functions, like the one presented in table 1, were computed. The first and second function were based on the first sample of 56 projects. The first one used eight predictors, as outlined above. For the second one they were reduced to the following two:

- Difficult versus non-difficult contract negotiations
- Experience of project manager

These were selected because they proved to be by far most substantial. (This selection was done step by step and is supported by SPSS.)

The third function was based on the second sample of 107 projects. Of the six predictors supplied by project data, the following four turned out to be substantial:

- Difficult versus non-difficult contract negotiations
- Government sector customer versus private customer
- External versus internal order
- Experience of project manager

The reliability of the three discriminant functions is shown in tables 2 - 4.

Group	% of sample members correctly assigned	**% of correct group predictions**
Success	80.0	**40.0**
Failure	56.1	**88.5**
Total	62.5	**62.5**

Table 2. 56 projects, 8 predictors

Group	% of sample members correctly assigned	**% of correct group predictions**
Success	60.0	**36.0**
Failure	61.0	**80.7**
Total	60.7	**60.7**

Table 3. 56 projects, 2 predictors

Group	% of sample members correctly assigned	**% of correct group predictions**
Success	39.3	**35.5**
Failure	74.0	**77.0**
Total	64.8	**64.8**

Table 4. 107 projects, 4 predictors

The numbers in the tables are conditional probabilities (in %). *"% of sample members correctly assigned"* is p(A│B) meaning the post-hoc analysis where A is *"Group is assigned"* and B *"Sample member belongs to the group"*, whereas *"% of correct group predictions"* is p(B│A), the basis for future prospect. It should be mentioned that the probabilities for correct group prediction are not provided by SPSS. They can easily be computed by Bayes' formula. They are of greater importance than the ratios of correctly assigned cases, because they give the probabilities for correct future predictions (provided that the samples are representative).

The results shown for the percentages of correct group predictions are quite similar. This fact suggests to use all three discriminant functions side by side, because they will support each other in many cases. To increase the data base for discriminant analysis, it is advisable to join every new finished case to the sample; this will of course slightly alter the discriminant functions.

The results of the case study are promising. They show that discriminant analysis can be a helpful tool for software project decisions. Especially the failure prediction is very reliable. But obviously, the results are still suboptimal. Future analysis should

yield more appropriate predictors to increase the percentage of correct predictions, especially in the success case.

REFERENCES

[1] **Boehm, B. W.** *Software Engineering Economics* Englewood Cliffs (1981); *General Electric PRICE Systems: PRICE-Software Model Reference Manual* Moorestown (1989)

[2] **Whatley, N. M.** *Cost/Schedule/Technical Performance Risk Analysis*, in: Stewart, R. D. and Wyskida, R. M. (ed.) *Cost Estimators Reference Manual* New York, Chichester, Brisbane (1987), pp. 259-310

[3] **Pugh, P. G.** *Who Can Tell What Might Happen? Risks and Contingency Allowances*, Memorandum, Directorate of Project Time and Cost Analysis, Ministery of Defence, London (1985), pp. 11ff

[4] **Altman, E.** *Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bancruptcy* Journal of Finance, Vol. XXIII, No. 4 (1968) pp. 589-609

[5] **Fahrmeier, L., Hamerle, A.** *Multivariate statistische Verfahren* Berlin, New York (1984)

[6] **Lachenbruch, P. A.** *Discriminant Analysis* London (1975)

[7] **Hemmerich, T.** *Indikatorgestützte Risikoanalyse von Projektaufträgen mittels einer Diskriminanzanalyse* Diploma thesis, Fachhochschule Giessen-Friedberg, Friedberg (1992)

[8] **Bühl, A., Zöfel, P.** *SPSS* München 1999

[9] **Backhaus, K., Erichson, B., Plinke, W., Weiber, R.** *Multivariate Analyse-methoden* 6th ed. Berlin 1990

[10] **Hartung, J., Elpelt, B.** *Multivariate Statistik* München 1990

The authors

Manfred Börgens is professor for mathematics at the Fachhochschule Giessen-Friedberg in Germany, a University of Applied Sciences. After gaining his Diploma and Ph.D. in mathematics from the University of Düsseldorf in 1977 resp. 1979, he joined the Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen and worked for governmental software and hardware projects. He joined the Fachhochschule in 1984 and has since been a teacher and researcher in several application-oriented fields. His current work includes Applied Statistics and Quality Control.

Thomas Hemmerich started his professional career with ABB Netzleittechnik GmbH in 1993. When preparing this paper he was working for the company's department for Supply Management. He gained his Diploma in Commercial Engineering in 1992 from the Fachhochschule Giessen-Friedberg. The major results of this paper are based on his Diploma thesis.

Ludwig B. Rüssel started his professional career with Deutsche Aerospace AG (now DaimlerChrysler Aerospace AG) in 1991. When preparing this paper he was heading the department of budgeting and control of the space group. He gained his Diploma in Business Administration in 1987 from the University of Cologne and his Ph.D. from the Coblenz School of Corporate Management (WHU), Germany.