International Journal of Mechanical, Industrial and Aerospace Sciences
ISSN: 2517-9950
Vol:7, No:12, 2013

A Study on a Research and Development Cost-Estimation Model in Korea

Babakina Alexandra, Yong Soo Kim

***Abstract*—**In this study, we analyzed the factors that affect research funds using linear regression analysis to increase the effectiveness of investments in national research projects. We collected 7,916 items of data on research projects that were in the process of being finished or were completed between 2010 and 2011. Data pre-processing and visualization were performed to derive statistically significant results. We identified factors that affected funding using analysis of fit distributions and estimated increasing or decreasing tendencies based on these factors.

***Keywords*—**R&D funding, Cost estimation, Linear regression, Preliminary feasibility study.

I. Introduction

GOVERNMENT interest in estimating research and development (R&D) expenses increases every year in an attempt to use the limited budget allocations more efficiently. In this R&D preliminary feasibility study, we analyzed the R&D expenses per project between 2010 and 2011 based on the research field and steps required for project completion. By performing a preliminary feasibility analysis, we were able to provide guidelines and a practical approach for research-fund distribution that could be used by project managers.

In this study, from a variety of fund distribution factors considered, we were able to identify the most important parameters. Our results showed that R&D budgets depended on the scale of the project, the stage of development, specific departments, and industry classification. Thus, the R&D budget distributions were classified by management department, industrial classifications, and technology classifications and, in some cases, were separated further by project attributes for the referring field to identify the commutated fit distribution, mean, and range of research funds. Based on the information obtained, guidelines were developed for R&D budget estimation.

II. Related Works

The field of cost estimation has received much attention over the years from manufacturing engineers and R&D analysts [1]. In conventional studies, several cost estimation techniques have been used, such as variant-based cost estimation, generative cost estimation, and hybrid cost estimation. Variant-based cost estimation is based on the actual cost of similar products manufactured previously. This approach focuses on part coding for converting the parts' features into numerical properties to be used for clustering [2]. Generative cost estimation is performed using a detailed analysis of the different production processes to assign a cost to the various design features of a product that, taken together, have engineering "meaning" [3]–[5]. Finally, hybrid cost estimation can be used when some of the parts have detailed information available, whereas other parts are still in the earliest stages of development and have insufficient data. Generative methods are used for parts for which the required data are available, and the variant-based approach can be used for parts still in the early stages [1]. In this study, the variant-based approach was performed based on actual R&D investment data in Korea.

TABLE I
Data Attributes Based On Research Characteristics

Attributes	Contents
Research Types	Basic research
Stage	Development
Support Ministries in Korea	Ministry of Education and Science Technology(MEST), Ministry of Land, Transport and Maritime Affairs(MOLIT), Korea Meteorological Administration(KMA), Ministry for Food, Agriculture, Forestry and Fisheries, Rural Development Administration, Ministry of Culture, Sports and Tourism, Ministry of knowledge economy (MKE)
Research Classification	BT (biotechnology), CT (Culture Technology), ET (Environmental Technology) IT (Information Technology), NT (Nano Technology), ST (Space Technology)
Science and Technology Classification	Construction / Transportation, Economy / Management, Scientific Technology and Society of humanity, Education, Machinery, Food, Agriculture, Forestry and Fisheries, Brain science, Chemistry

III. Research Methods

In this paper, we analyzed the factors that affect research funding using linear regression analysis to increase the effectiveness of investments in national research projects. We collected 7,916 pieces of data on research projects that were in the process of being finished or were completed between 2010 and 2011. The starting point of the analysis was to examine the total project cost with respect to the government R&D department classification, the stage of R&D, the department's research classification, and the science and technology standard classification.

In the diagram shown below (Fig. 2), the arrow in the graphic represents the logical path from the preparation and installation of the "targeted selection" to its ultimate impact on the fitting distribution.

Babakina Alexandra is with the Department of Industrial and Management Engineering, Kyonggi University Graduate School, 154-42, Gwanggyosan-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do, South Korea (e-mail: babakina25@hotmail.com).

Yong Soo Kim is an Assistant Professor of the Department of Industrial and Management Engineering, Kyonggi University, 154-42, Gwanggyosan-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do, South Korea (Corresponding author to provide phone: +82 31-249-9771; fax: +82 31-244-3534; e-mail: kimys@kgu.ac.kr).

A. Step 1: Target Selection

The data were divided into the categories, including the government departments of the analysis target, the R&D stage of development, the research classification, and the science and technology standard classification. The data, once categorized, were computed as a monthly project expense. Candidate distributions were selected to optimize the fit, and the distributions were then fit to our data.

B. Step2: Selection of the Best-fit Distribution

In this stage, a number of mathematical distributions were used to model the distribution of fit. Seven distributions were considered to determine the best fit: a normal distribution, lognormal distribution, gamma distribution, exponential distribution, Weibull distribution, logistic distribution, and a log-logistic distribution.

The process of fitting the distributions involved the use of certain statistical methods that allowed the distribution parameters to be estimated based on the sample data. This is where distribution-fitting software can be very useful. In our research, the best-fit distribution for each case among seven probability distributions was selected using Minitab statistical software. To finish this step, the fits were evaluated using the Anderson–Darling (A–D) value to determine the best-fit (lowest value) distribution. Additional information on the (A–D) procedure is provided below.

- The A–D procedure is a general test to compare the fit of an observed cumulative distribution function to the expected cumulative distribution function.
- When an A–D goodness-of-fit test can be applied to a continuous distribution, then it is considered a goodness-of-fit test method.

C. Step3: Goodness- of- fit test

In the next step, we performed a goodness-of-fit test for the distribution with the lowest A–D value, which was selected in the previous step. In this step, the fit of the distribution chosen depended on the *p*-value and the significance level specified. In this study, the significance level was set to 0.05. We observed that typically, most of the distributed data outliers were mainly in the vicinity of the upper and lower limits; thus, ~5–20% of the outliers were eliminated from these areas of the distribution. By excluding the upper and lower threshold values, we were able to obtain a better-fit distribution.

D. Step4: Estimation of distribution parameters

In this step, we estimated the parameters and confidence intervals for the fit distribution. When the distribution passed the goodness-of-fit test, the parameters were then estimated for the distribution. Minitab was used to compute probabilities, the parameter's estimated value, the distribution characteristics, and the percentile table for the relevant distribution. In our research, among various distribution-estimation methods, we chose the maximum-likelihood method to estimate the probability interval for a specific range.

E. Step5: Completion

The most appropriate distribution was derived based on the previous four processes (Steps 1–4). The current step represents completion of the distribution analysis.

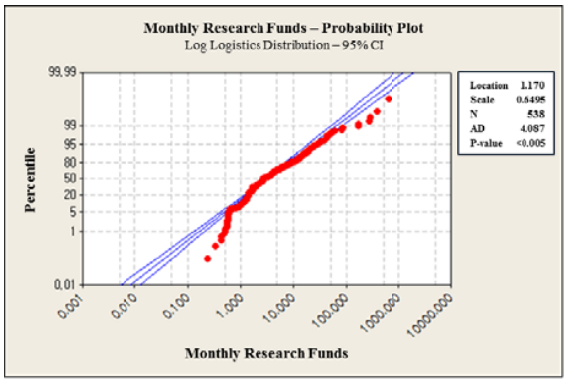

Fig. 1 Probability plot for monthly funds

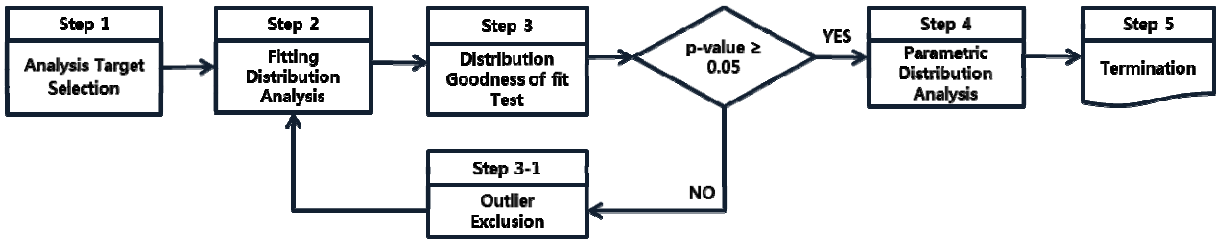

Fig. 2 Fit-distribution analysis graphic

IV. Experimental Results

In this study, we collected 7,916 data points on research projects that were in the process of being finished or were completed between 2010 and 2011. An analysis of fit distribution, mean, and confidence intervals of monthly funds (unit: millions) was determined based on the following classifications: government R&D classification, stage of R&D, the department's nomination, research classification (6T), and the science and technology standard classification.

The analysis was carried out considering R&D departments, different R&D stages, research classification (6T), and project type (number of projects targeted: >30).

Studies classified but not fitted were subdivided using the science and technology standard classification and then analyzed accordingly. The same procedure was used to sort the

outlier values. Each Ministry category and its research classification depended on a specific p-value to determine the appropriate distribution for their department (i.e., normal distribution, lognormal distribution, gamma distribution, exponential distribution, Weibull distribution, logistic distribution, or log-logistic distribution). The density plot characteristics of each distribution were used to define the appropriate amount of research funds for each category.

V. Conclusion

By performing a preliminary feasibility study, it became possible to explore a plan that provided practical help for project managers with regard to research-funds distribution. Specifically, in this study, we identified the best distribution parameters among a variety of distributions to explain the existing research-funds distribution. This study described a five-step analysis process based on Minitab software to estimate the distribution parameters. The Anderson–Darling goodness-of-fit test was used to determine the appropriate distribution for each department.

By excluding the upper and lower threshold values, we were able to better fit the distribution resulting from 7,916 items of data on research projects that were in the process of being finished or were finished between 2010 and 2011. We identified factors that affected funding using analysis of fit distributions and estimated the parameters using the maximum-likelihood method. The distribution identified using this method was in good agreement with the data obtained.

Acknowledgment

This work was supported by Kyonggi University's Graduate Research Assistantship 2013.

TABLE II
Analysis Results of Projects Conducted during 2010–2011

Department name	6T	Probability distribution	P-value	N	Mean	Standard deviation	Median	The first quartile	The third quartile
Small and Medium Business Administration	BT	Log logistic distribution	<0.005	493					
	CT	Log logistic distribution	>0.250	93	8.5399 (7.6864, 9.4882)	5.0385	7.5305 (6.8390, 8.2919)	5.5789	10.1648
	ET	Log logistic distribution	<0.005	1022					
	IT	Log logistic distribution	<0.005	1349					
	NT	Log logistic distribution	0.0220	209	9.6343 (8.9897, 10.3253)	5.5715	8.5263 (8.0073, 9.0788)	6.3433	11.4604
	ST	Log logistic distribution	>0.250	43	11.0984 (8.8355, 13.9409)	11.1305	8.7451 (7.2011, 10.6200)	5.8145	13.1525
MKE	BT	Log logistic distribution	<0.010	135					
	ET	Log logistic distribution	<0.005	641					
	IT	Log logistic distribution	<0.005	398					
	NT	Log logistic distribution	<0.005	114					
	ST	Log logistic distribution	0.2190	36	35.7852 (20.8814, 61.3263)	*	18.4089 (13.2464, 25.5835)	9.5992	35.3038
Ministry of Environment	ET	Log logistic distribution	0.0210	133					

TABLE III
OUTLIERS ANALYSIS RESULTS OF PROJECTS CONDUCTED DURING 2010–2011

Department name	6T	Main Category	Probability distribution	P-value	N	Mean	standard deviation	Median	The first quartile	The third quartile
MEST	BT	Health and Medical	Log-normal distribution	0.486	74	4.9850 (3.7505, 6.6258)	6.6881	2.9792 (2.3642, 3.7540)	1.5026	5.9064
		Life Sciences	Log logistic distribution	0.043	53					
	IT	Information and Communication	Log logistic distribution	0.009	39					
MOLIT	ET	Construction and Transportation	Log logistic distribution	0.052	35	12.4252 (8.2914, 18.6200)	57.4447	8.0554 (6.0851, 10.6638)	4.6982	13.8117
Agriculture, Forestry and Fisheries ministry	BT	Food, Agriculture, Forestry and Fisheries	Log logistic distribution	0.028	255					
Rural Development ministry	BT	Food, Agriculture, Forestry and Fisheries	Log logistic distribution	0.023	269					
		Life Sciences	Log logistic distribution	0.068	32	7.6346 (5.8235, 10.0090)	8.0635	5.9462 (4.7550, 7.4357)	3.9169	13.8117
Department of Health and Human Services	BT	Health and Medical	Log-normal distribution	<0.005	132					
Small and Medium Business Administration	BT	Food, Agriculture, Forestry and Fisheries	Log logistic distribution	<0.005	147					
		Health and Medical	Log logistic distribution	<0.005	140					
		Life Sciences	Log logistic distribution	<0.005	127					
	CT	Information and Communication	Log logistic distribution	0.154	33	8.4364 (7.1054, 10.0810)	4.8163	7.5113 (6.4300, 8.7745)	5.6072	10.0621
	ET	Construction and Transportation	Log logistic distribution	<0.005	37					
		machinery	Log logistic distribution	<0.005	370					
		Energy Resources	Log logistic distribution	0.007	146					
		Materials	Log logistic distribution	0.03	87					
		Electrical and Electronic	Log-normal distribution	0.087	99	8.8171 (8.0041, 9.7127)	4.3463	5.3740 (5.2152, 6.3926)	10.8320	5.0581
		Chemical Engineering	Log logistic distribution	<0.005	54					
		Chemistry	Log logistic distribution	0.039	45					
		Environment	Log logistic distribution	0.006	146					

REFERENCES

[1] B. Verlinden, J. R. Duflou, P. Collin and D. Cattrysse, "Cost estimation for sheet metal parts using multiple regression and artificial neural networks: A case study", *Int. J. Production Economics*, Vol. 111, pp.484-492, 2008.
[2] C.C. Gallagher and W.A. Knight, "Group technology, production methods in manufacturing", Ellis Horwood, Chichester, England, 1986.
[3] L. Wierda, "Cost information tools for designers, Delft University Press, Delft, Netherlands, 1990.
[4] T.S., Geiger and D.M. Dilts, "Automated design-to-cost: Integrating costing into the design decision", *Computer-Aided Design*, vol.28, no.6-7, pp.423-438, 1996.
[5] S. Staub-French, M. Fischer, J. Kunz and B. Paulson, "A generic feature-driven activity-based cost estimation process", *Advanced Engineering Informatics*, vol.17, pp.23-29, 2003.