Available online at www.sciencedirect.com
SCIENCE DIRECT®

International Journal of Project Management 23 (2005) 275–282

INTERNATIONAL JOURNAL OF PROJECT MANAGEMENT

www.elsevier.com/locate/ijproman

Simulation-based estimation for correlated cost elements

I.-T. Yang

Department of Civil Engineering, Tamkang University, No. 151 Ying-chuan Road, Tamsui, Taipei 251, Taiwan

Received 12 August 2004; received in revised form 10 November 2004; accepted 1 December 2004

Abstract

In this paper, we present a general method to incorporate correlations between cost elements in the process of cost estimation. The proposed method first checks the feasibility of the correlation (Pearson or Spearman) matrix, adjusts it if necessary, then uses the correlations to generate correlated multivariate random vectors, which are employed to model possible outcomes of the cost elements. The method is applied to a full data set of 216 British office buildings to illustrate its practical use. The application result indicates that the impact of correlations is significant and may cause serious problems if neglected. The result is also used to validate that the proposed method can capture the correlations with relatively small deviations.

Keywords: Cost; Estimation; Risk; Simulation; Project management

1. Introduction

Cost estimation is a process of collecting, analyzing, and summarizing data in order to prepare an educated projection of the anticipated cost of a project [26]. This process begins in the early stages of the project and repeats frequently during the entire life cycle. The reliability of cost estimation is important to ensure the success of the project since it serves as the foundation for making critical decisions on financial investments and borrowings.

The prices of all the resources (material, equipment, and labor) are exposed to certain levels of uncertainty, particularly when the project life cycle is lengthy [25]. To manage the inherent uncertainty, Monte-Carlo simulation methods have been widely applied for various types of projects, such as [3,11,15,31]. In Monte-Carlo simulation, a mathematical model is constructed based on pre-specified probability distributions, which describes the possible outcomes of major cost elements (e.g., substructure, exterior walls, and electrical finishing) involved in a project, and run to see what the overall project cost will be for each simulation replication. After a certain number of replications, the collected samples are used to derive the output distribution of the overall project cost.

An enhancement of ordinary simulation methods has been directed to consider statistical correlations (dependencies) between cost elements. The correlation represents the co-movement of two cost elements; when one is more expensive, the other tends to cost more as well (or cost less for a negative correlation). Arguments and evidences for the existence of correlations and their profound impact on simulation results have been presented in the literature [10,23,33]. To treat the correlations, various approaches have been proposed, such as [29,30,34].

The goal of this paper is to present a simulation-based method to incorporate correlations between cost elements with more modeling capabilities. Specifically, the present method is unique in handling the following requirements:

1. To allow the distributions (i.e., marginal distributions) of individual cost elements to be of different types. Namely, some of them may only be expressed

E-mail address: ityang@mail.tku.edu.tw

0263-7863/$30.00 © 2005 Elsevier Ltd and IPMA. All rights reserved.
doi:10.1016/j.ijproman.2004.12.002

with discrete and finite options whereas others can be expressed as continuous functions. In addition, those continuous distributions may come from different families (e.g., some are lognormal while some are beta).
2. To provide an automatic procedure to check the feasibility (a mathematical definition will be given later) of a correlation matrix and adjust it if infeasible.

This paper is organized as follows. In the following two sections, we introduce two different approaches (historical and subjective) to obtain necessary input for the analysis: the marginal distributions of individual cost elements and their correlations. The characteristics of the two approaches lead to the need for a more general method. In the section followed, we present the proposed method in two stages and provide computational guidelines. The proposed method is then applied to a modified British data set composed of 216 office buildings to illustrate its practical use. The results are used to validate the proposed method. Closing remarks are presented in the last section.

2. Required input: distributions and correlations

There are two sets of input data required to perform a simulation-based cost analysis considering correlations. The first set describes marginal distributions of individual cost elements and the second is a correlation matrix consisting of the correlation coefficients between pairs of cost elements. Both sets of data can be estimated in two ways: (1) by summary statistics on historical data, or (2) by subjective judgments. In what follows we give brief introduction to these two approaches and elucidate how their characteristics give rise to a more general simulation method.

2.1. Choice of distribution

When historical data is used to describe the marginal distributions, it involves an attempt to fit theoretical distributions to the data and verify goodness-of-fit statistically. While a complete review of goodness-of-fit tests and their theoretical backgrounds is given in [18], the fitting process can be done very efficiently by commercial software packages (such as Palisade BestFit, Arena, ProModel, and Crystal Ball). A typical result is a list of several "good" distributions and their associated parameters, based on which the estimator can select the most proper one. Previous studies suggested that the lognormal distribution fits historical cost data better than other well-known distributions, such as normal or beta [30,33].

Despite its theoretical maturity, using historical data to forecast possible outcomes has some pitfalls. First, actual values may lie outside the range of historical records due to new technology, equipment, and material. Second, historical data may not adequately represent the true underlying population because of sampling error [12]. Last, the prices of resources may not always be repeatable, thus, the historical approach may be fallacious [23].

In the absence of reliable historical data, the second best alternative is for a cost estimator to rely on his/her experiences to subjectively specify the marginal distributions. In the context of probabilistic estimation, it is usual to assume the underlying distribution is a beta distribution whose parameters are specified by three point estimates: the minimum, maximum, and most likely values [4]. Some recent controversy has been whether the beta distribution should be replaced with the triangular distribution since the former does not have clear-cutting bounds [16] and requires four parameters, which do not have a one-to-one correspondence with the three estimates [32].

Both approaches mentioned above (historical and subjective) may involve the following practical concerns. First, they may actually be mixed in practice. A cost estimator may not have historical data of *all* the cost elements if some works are usually outsourced or subcontracted. Thus, he/she can obtain summary statistics only on those data on hand and has to rely upon subjective estimation for the remaining cost elements. Another possibility occurs when the estimator has reasons to believe some of the cost elements in a new project have their own bounds (minimum and maximum) and thereby cannot be represented by historical data. For these particular elements, the subjective approach is more appropriate. Second, it has been frequently encountered that the price options of some cost elements are collections of discrete outcomes in lieu of continuous functions [2].

The practical concerns above give rise to the need for a more general simulation model, which should be able to treat all different types of distributions in one framework (some are discrete while some are continuous; some are lognormal while some are beta). This is the first goal of this paper.

2.2. Specification of correlations: product-moment and rank

Two different measures have been used to reflect the degree of relation between cost elements. The first one is an ordinary product-moment (Pearson) correlation coefficient and the second is a rank (Spearman) correlation coefficient. The product-moment correlation can be calculated as follows: