Chapter 17
A Framework of Statistical and Visualization Techniques for Missing Data Analysis in Software Cost Estimation

Lefteris Angelis
Aristotle University of Thessaloniki, Greece

Nikolaos Mittas
Aristotle University of Thessaloniki, Greece

Panagiota Chatzipetrou
Aristotle University of Thessaloniki, Greece

ABSTRACT

Software Cost Estimation (SCE) is a critical phase in software development projects. However, due to the growing complexity of the software itself, a common problem in building software cost models is that the available datasets contain lots of missing categorical data. The purpose of this chapter is to show how a framework of statistical, computational, and visualization techniques can be used to evaluate and compare the effect of missing data techniques on the accuracy of cost estimation models. Hence, the authors use five missing data techniques: Multinomial Logistic Regression, Listwise Deletion, Mean Imputation, Expectation Maximization, and Regression Imputation. The evaluation and the comparisons are conducted using Regression Error Characteristic curves, which provide visual comparison of different prediction models, and Regression Error Operating Curves, which examine predictive power of models with respect to under- or over-estimation.

INTRODUCTION

Software has become the key element of any computer-based system and product. The complicated structure of software and the continuously increasing demand for quality products justify the high importance of software engineering in today's world as it offers a systematic framework for development and

DOI: 10.4018/978-1-5225-3923-0.ch017

maintenance of software. One of the most important activities in the initial project phases is Software Cost Estimation (SCE). During this stage a software project manager attempts to estimate the effort and time required for the development of a software product. The importance of software engineering and the role of cost estimation in software project planning has been discussed widely in literature. (Jorgensen & Shepperd 2007). Cost estimations may be performed before, during or even after the development of software.

The complicated nature of a software project and therefore the difficult problems involved in the SCE procedures emerged a whole area of research within the wider field of software engineering. A substantial part of the research on SCE concerns the construction of software cost estimation models. These models are built by applying statistical methodologies to historical datasets which contain attributes of finished software projects. The scope of cost estimation models is twofold: first, they can provide a theoretical framework for describing and interpreting the dependencies of cost with the characteristics of the project and second they can be utilized to produce efficient cost predictions. Although the second utility is the most important for practical purposes, the first utility is equally significant, since it provides a basis for thorough studies of how the various project attributes interact and affect the cost. Therefore, the cost models are valuable not only to practitioners but also to researchers whose work is to analyse and interpret.

In the process of constructing cost models, a major problem arises from the fact that missing values are often encountered in some historical datasets. Very often missing data are responsible for the misleading results regarding the accuracy of the cost models and may reduce their explanatory and prediction ability. The aforementioned problem is very important in the area of software project management because most of the software databases suffer from missing values and this can happen for several reasons.

A common reason is the cost and the difficulties that some companies face in the collection of the data. In some cases, the cost of money and time needed to collect certain information is forbidding for a company or an organization. In other cases, the collection of data is very difficult because it demands consistence, experience, time and methodology for a company. An additional source of incomplete values is the fact that data are often collected with a different purpose in mind, or that the measurement categories are generic and thus not applicable to all projects. This seems especially likely when data are collected from a number of companies. So, for researchers whose purpose is to study projects from different companies and build cost models on them, the handling of missing data is an essential preliminary step (Chen, Boehm, Menzies & Port 2005).

Many techniques deal with missing data. The most common and straightforward one is *Listwise Deletion* (LD), which simply ignores the projects with missing values. The major advantage of the method is its simplicity and the ability to do statistical calculations on a common sample base of cases. The disadvantages of the method are the dramatic loss of information in cases with high percentages of missing values and possible bias in the data. These problems can occur when there is some type of pattern in the missing data, i.e. when the distribution of missing values in some variables is depended on certain valid observations of other variables in the data.

Other techniques estimate or "impute" the missing values. The resulting complete data can then be analyzed and modelled by standard methods (for example regression analysis). These methods are called *imputation methods*. The problem is that most of the imputation methods produce continuous estimates, which are not realistic replacements of the missing values when the variables are categorical. Since the majority of the variables in the software datasets are categorical with many missing values, it is reasonable to use an imputation method producing categorical values in order to fill the incomplete dataset and then to use it for constructing a prediction model.