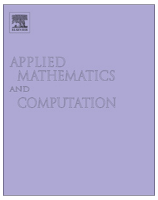
Applied Mathematics and Computation 253 (2015) 126–134

Contents lists available at ScienceDirect

Applied Mathematics and Computation

journal homepage: www.elsevier.com/locate/amc



APPLIED
MATHEMATICS
AND
COMPUTATION

Stepwise regression data envelopment analysis for variable reduction

Mithun J. Sharma [a], Yu Song Jin [b]

[a] Centre for Management Studies, Dibrugarh University, India
[b] Department of Shipping Management, Korea Maritime Ocean University, Republic of Korea

ARTICLE INFO

Keywords:
Data envelopment analysis
Stepwise regression
Variable reduction

ABSTRACT

In this paper, we develop stepwise regression data envelopment model to select important variables. We formulate null hypothesis to understand the importance of each variable and use Kruskal–Wallis test for this purpose. If the Kruskal–Wallis test does not reject the null hypothesis then we can conclude that all the variables are of equal importance as their presence and on the other hand absence of other variable does not create huge fluctuations in efficiency scores in fact give a complete ranking relative to base model. If the Kruskal–Wallis test does reject the null hypothesis this will imply there is significant fluctuation in the efficiency score relative to base model. And therefore we have to further check the pair of variables that causes the fluctuation in order to determine its importance using Conover–Inman test. The results of the proposed models are compared with the results of previously published models of the same dataset. The proposed models helps understand the extent of misclassification decision making units as efficient/inefficient when variables are retained or discarded alongside provides useful managerial prescription to make improvement strategies.

1. Introduction

In analyzing the performance of a supply chain using data envelopment analysis (DEA) a lot of process variables has to be determined as inputs and outputs. Operations managers are often interested in finding a vital few that would be most influential in the quality of analysis. Although stepwise regression methods are often used for finding important variables, multicollinearity among variables often impacts on its performance [1]. The DEA classification depends on the choice of input and output variables. Often there are many alternatives. Since, there is a rule of thumb – the number of inputs and outputs should be less than one-third of the number of units, sometimes we need to detect redundant variables [2]. Sengupta [3] recommends using the canonical correlation analysis (CCA) for identifying the maximal correlation between linear combinations of two sets of variables. A detailed literature review on variable selection models and DEA is given in the next section.

In practice the input and output variables are usually highly correlated with one another, often reflecting no more than the relative size of each decision making unit (DMU). To counteract the limited distinction provided by a DEA with many variables, analysts for many years have taken the approach of retaining only some of the variables originally planned the analysis omitting, on an ad hoc basis, variables that are highly correlated with those retained [4]. In this paper, we describe a systematic mathematical and statistical model for deciding which of the original correlated variables can be omitted with

E-mail addresses: mithunjsharma@dibru.ac.in (M.J. Sharma), coppers@kmou.ac.kr (S.J. Yu)

http://dx.doi.org/10.1016/j.amc.2014.12.050
0096-3003/© 2014 Elsevier Inc. All rights reserved.

least loss of information, and which should be retained. In Section 2 we give a detailed literature review on variable selection models using DEA and in the Section 3 we propose step-wise regression DEA model and test the model with real world data in Section 4. Finally, in Section 5 we conclude our paper.

2. Variable selection models and DEA

The data envelopment analysis (DEA) [5] classifies n organizational units into efficient and inefficient units, given their multiple inputs and their multiple outputs. Let x_{ij} be a given level of the ith input of unit $j(i = 1, \ldots, m,\ j = 1, \ldots, n)$ and y_{rj} a given level of the rth output of unit and the outputs (u_r^k) which maximizes the ratio between the weighted output and the weighted input. These ratios for all the n units are bounded from above by one, and the weights are all positive. Each unit k is assigned the highest possible efficiency score (ratio) by choosing the most favorable weights. Therefore, if a unit does not reach the maximum possible value (1) it is inefficient, otherwise it is efficient [2].

The major advantage of DEA over other methods that determine efficiency, such as cost-benefit analysis or regression, is that the relative weights of the variables do not need to be known, a priori. Jenkins and Anderson [4] explicitly points out that “more the DMUs in the DEA analysis, the more the constrained weights likely to be, and the more discerning the DEA result. Conversely, the more variables (inputs and outputs) in the DEA, the less discerning is the analysis”. Therefore there is great motivation to have limited variable in the DEA analysis.

In this section we will summarize the literature review in particular from the two papers [4,6] that has dealt with variable reduction algorithm and its statistical properties. In the context of DEA results to be more discerning, some of the researchers have worked to combine DEA with multivariate statistics. Friedman and Sinuany-Stern [2] used cannonical correlation and DEA in a way similar to [3] where [3] computes two sets of weights, one of the inputs and another for the outputs, in such a way that the correlation between a weighted vector of inputs, and a weighted input vector of outputs, is highest for the given dataset. [2] then used hybrid DEA/Discriminant model to classify DMUs as either efficient or inefficient from the function that best discriminated between the efficient and inefficient set. Zhu [7] applied principal component analysis to his 16 observation to develop a unique value for the ratio of every input and output for each DMU. The ratios thus obtained were used to rank his data and compared with those of DEA results. Ueda and Hoshiai [8] obtained the constraints by aggregating the data with one or two of the principal components of the original data and the objective for each LP was derived from the standard objective ratio form. These studies do not sufficiently distinguish the effect of presence or absence of a variable in a dataset instead according to [4] they are plausible “add-ons”.

Another commonly used method for variable reduction for inclusion in DEA model is to apply regression and correlation analysis [9] where highly correlated variables are retained and redundant variables are discarded. Norman and Stoker [10] proposed a forward method where variables are added one at a time and the efficiency scores obtained from forward method is tested for correlation. Their indicator to identify the importance level was the degree of correlation. Higher the degree of correlation higher its importance. Jenkins and Anderson [4] used regression and correlation on DEA based model to omit variables with least loss of information. The variables to be omitted are chosen based on partial correlation, in which the variance of an input or output around its mean value indicates the importance of a specific variable. They use partial correlation as a measure of information, instead of a simple correlation matrix. However, partial correlation matrix is based on the assumptions that the data is drawn from an approximately normal distribution and the conditional variance is homoscedastic. On the other hand DEA is a non parametric approach and it is unclear, particularly with a relatively small dataset, whether such conditions exist. Banker [11,12] used statistical tests to evaluate the marginal impact on efficiencies of omitting or retaining a variable. Pastor et al. [13] explored nested DEA models with reduced and extended forms of models obtained by dropping and adding one variable to each model. Efficiency was calculated for each model under reduced and extended forms of the model. Statistical test was carried out to determine the significance of each variable under study. Wagner and Shimshak [6] modelled a stepwise selection of DEA variables using backward approach. Their algorithm evaluated all the possible inputs and outputs in considering the efficiency score. Then step-wise one variable was dropped and the average change in efficiency score was computed. This procedure continued until there remained a single input and a single output. The objective of the method was its ease of use or data storage, and ease of explanation. Although, these models give statistical validation and some insights into effect of variable omission and inclusion, they fail to exclusively distinguish the importance or ranks of each such omission or inclusion. In this paper we propose an algorithm, step-wise regression in DEA. This paper advances the work of variable reduction methods in DEA by formalizing the proposed algorithm.

3. Stepwise regression-DEA model

Stepwise regression-DEA algorithm proposed in this paper is a hybrid extension of stepwise regression algorithm [14]. Stepwise regression algorithm [14] is an automatic procedure for statistical model selection in cases where there is a large number of potential explanatory variables, and no underlying theory on which to base the model selection. The procedure is used primarily in regression analysis, though the basic approach is applicable in many forms of model selection.