European Journal of Operational Research 202 (2010) 810–818

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Decision Support

Integrating categorical variables in Data Envelopment Analysis models: A simple solution technique

Gerrit Löber [a], Matthias Staat [b,*]

[a] University of Mannheim, Department of Economics, 68131 Mannheim, Germany
[b] DSC GmbH, Carl-Benz-Straße 16a, 69191 Schriesheim, Germany

ARTICLE INFO

Article history:
Received 16 December 2007
Accepted 21 May 2009
Available online 2 June 2009

Keywords:
Non-discretionary variables
Categorical variables
Returns to scale

ABSTRACT

This paper introduces a novel method to incorporate categorical non-discretionary variables in Data Envelopment Analysis (DEA) models. While solutions to this problem have been introduced before, they have rarely been employed in applied work. We surmise that existing solution concepts pose problems for applied researchers and develop a simple and straightforward alternative based on indicator variables. We thereby provide a flexible tool for models with categorical variables that–unlike the approaches mentioned above–can be solved with standard DEA software irrespective of scale assumptions even if no option for non-discretionary variables is available. Furthermore, there is no need to split the data and run multiple DEA, one for each data set generated. The model is extensible to categorical discretionary variables and in addition to non-hierarchical data.

1. Introduction

One of the major innovations related to the Data Envelopment Analysis (DEA) model introduced by Charnes et al. (1978)[1] nearly 30 years ago were different approaches to handle so-called non-discretionary[2] variables. The importance of this topic is reflected in the number of citations the seminal papers by Banker and Morey (1986a,b) have generated: 239 (94) publications refer to non-discretionary or environmental variables (a: 146 (62); b: 93 (32)).[3]

Ordinary DEA models presuppose only variables representing a proper input or output which are an integral part of the technology to be estimated. For input-oriented approaches, the efficiency of an observation is assessed by calculating the minimal radial reduction of inputs that would be necessary to arrive at the technology frontier. Obviously, the reduction is zero for observations already efficient but for inefficient observations substantial reductions may be required. This implies that the observation–for instance a company being evaluated–has control over the variables specified as inputs; otherwise, the point of calculating input reductions would be moot.

However, already Charnes et al. (1981) mention that not all variables in their specification may be changed at the discretion of the observations (schools) analyzed; these factors are therefore not an integral part of the technology but nevertheless affect the efficiency of observations. Banker and Morey (1986a,b) discuss the incorporation of such variables into DEA models. These non-discretionary variables are mostly accounted for by employing their one-stage approaches or variants thereof; however, semi-parametric and fully non-parametric two-stage approaches (Simar and Wilson, 2007; Daraio and Simar, 2005; Daraio and Simar, 2007) have gained popularity.

Recent reviews[4] on the topic were presented by Syrjänen (2004) and Muñiz et al. (2006). While the latter summarizes the state of the current knowledge, Syrjänen (2004) points at several problems related to different returns to scale assumptions in connection with continuous non-discretionary variables and discusses a generalized model comprising a number of different models as special cases.

The literature distinguishes between two types of non-discretionary variables, namely continuous and categorical. A categorical non-discretionary variable is simply an ordinal variable. To facilitate comparison with other work, we label the cardinal case as

* Corresponding author.
E-mail addresses: gloeber@rumms.uni-mannheim.de (G. Löber), Matthias.Staat@dsc-gmbh.com (M. Staat).

[1] The model by Charnes et al. (1978) is associated with a constant returns to scale (CRS) technology.

[2] The term environmental variable is also used for non-discretionary variables in the context of DEA, see, Daraio and Simar (2007).

[3] To enable the reader to assess the relative importance of non-discretionary variables, we provide citation counts for a number of other important extensions to the basic DEA model, which is the constant returns to scale (CRS) case: the introduction of variable returns to scale (VRS) by Banker et al. (1984) has 838 (241) citations. The data base for the citation search is www.isiknowledge.com (EBSCO database) as of September 2007.,The proportions should reflect the relative importance of the respective topic.

[4] More dated surveys are provided by Ray (1998) as well as Golany and Roll (1993).

0377-2217/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2009.05.032

"continuous" and the ordinal case as "categorical". Banker and Morey (1986b) treat the population size in an area served by a pharmacy as a categorical variable. While this is not necessarily true for the population count chosen as an example by Banker and Morey (1986b) these categorical variables usually have a qualitative character; i.e., it is not clear how much of a difference in effect a higher/lower value signals.

The solution for non-discretionary variables is straightforward in case of a VRS technology; the details of the procedures are given below. The CRS case for continuous non-discretionary variables was treated in Banker and Morey (1986a) as well but – despite the fact that numerous applications of the CRS case with discretionary variables exist – seems to have received very limited attention by other researchers (see again Syrjänen, 2004). For instance, most of the models for energy regulation listed in Pollitt and Jamasb (2001) rest on CRS but none of them comprise any non-discretionary variables, while von Hirschhausen et al. (2006) in their model for the German electricity market use a continuous variable for customer density but specify their CRS model incorrectly.

We propose a solution for the case of categorical non-discretionary variables, a case which up to now has virtually gone unnoticed in the applied literature. There is no reason to believe that categorical variables are *a prioi* less relevant than cardinal ones and one can only speculate why the case of categorical non-discretionary variables has so far been ignored in applied research despite the fact that solutions exist. The two solution concepts available in principle are: (1) splitting the data set according to the distinct values of the categorical variables (in the spirit of Charnes et al., 1981) and (2) the models proposed by Banker and Morey (1986b).[5]

Splitting the data set may become very tedious especially when using more than one categorical variable because for each distinct combination of categorical variables a separate DEA run needs to be carried out and the results of these potentially many DEA partial runs need to be collected. The only advantage of the model by Banker and Morey (1986b) is that splitting the data into subsamples can be avoided. However, this comes at a cost as the approach cannot be solved using standard DEA software – Syrjänen (2004, p. 24) points out that a some of the popular DEA software packages offer an option for non-discretionary variables for a CRS technology (Banker and Morey, 1986b) but do not execute it properly which may have kept researchers from applying this method.

With the method we propose it is possible to treat the case of categorical variables regardless of the returns to scale assumption within a simple framework by constructing special indicator variables. This model does not require splitting the data. In addition, it can be solved by any standard linear program (LP) solver or DEA software, even if the software does not explicitly handle the non-discretionary case.[6] It is therefore easier to use than both the approach of Charnes et al. (1981) and the method proposed by Banker and Morey (1986b). We expect this method to facilitate the application of models for non-discretionary categorical variables, which up to now has been neglected in applied research. The proposed idea can in addition be applied to discretionary categorical variables and non-hierarchical categoricals as proposed by Førsund (2002) even in the absence of numerical data. We will follow Banker and Morey (1986b) with our description since these two additions are straightforward.

The paper is organized as follows: the next section contains a brief exposition of DEA. At the same time, the standard solution concepts for the problem of categorical non-discretionary variables are introduced. This is followed by some observations on LPs that are particularly relevant for DEA problems. The consequences of these observations for the method introduced in the sequel are discussed in a separate section. Our approach is then demonstrated by a simple numerical example. We also discuss its application using the pharmacy data from Banker and Morey (1986b). A demonstration using to data for German electricity distribution utilities demonstrates the full potential of our methodology before we offer some concluding remarks.

2. DEA: standard models

2.1. Basics

This section presents the relevant background on DEA.[7] A technology Ψ is defined as: $\Psi = \{(x,y) \in \mathbb{R}_+^{P+Q} : x \text{ can produce } y\}$. The technology comprises P inputs x and Q outputs y. Provided Ψ satisfies a particular set of axioms, a DEA can be applied to assess the efficiency of some observation. The axioms are convexity, monotonicity, inclusion of observations and minimum extrapolation (see Banker and Morey, 1986b, Appendix B).[8] The following LP (1), representing an input-oriented specification[9] resting on the assumption of VRS may be used to calculate an estimate $\hat{\theta}(x_0, y_0)$ of the true efficiency score θ for an observation (x_0,y_0):

$$
\begin{aligned}
&\min \theta^{\mathrm{LP}(1)} \\
&\text{s.t.} \quad \sum_{n=1}^{N} \lambda_n y_{qn} \geqslant y_{q0}, \quad q = 1,\ldots,Q \\
&\qquad \sum_{n=1}^{N} \lambda_n x_{pn} \leqslant \theta x_{p0}, \quad p = 1,\ldots,P \\
&\qquad \sum_{n=1}^{N} \lambda_n = 1 \\
&\qquad \lambda_n, x_{pn}, y_{qn} \geqslant 0, \quad n = 1,\ldots,N
\end{aligned}
\tag{1}
$$

where the data contain N observations indexed by n. The VRS-assumption is embodied in the (convexity) constraint on the intensity variables $\sum_{n=1}^{N} \lambda_i = 1$. Referent points[10] are formed as the linear combination of the input and output values of efficient peers for the observation that is benchmarked and which have $\lambda_n > 0$. LP (1) therefore represents a model with $P + Q + 1$ constraints (not counting the non-negativity constraints) and must be solved N times, once for each observation. LP (1) serves as a starting point of our discussion.

2.2. Extensions standard DEA models

In the present section, we discuss several extensions related to non-discretionary variables that have been proposed in the literature; we focus on one-step approaches. As mentioned in the introduction, some of the material to be discussed below was originally presented in a somewhat inaccessible manner. We therefore enter in an extended discussion of these models, making an effort to present them in a clear and concise way. Following up on Banker and Morey (1986a,b) some authors (Lovell, 1994; Ruggiero, 1996) have proposed alternative approaches to the problem by modifying their model. However, our approach aims at incorporating categorical non-discretionary variables without modifications to the standard DEA program, which none of the alternative approaches achieve. We already mentioned Syrjänen (2004) generalized

[5] An alternative model was proposed by Lovell (1994) and later by Ruggiero (1996). These could be solved with the algorithm proposed below.

[6] The popular DEAP package by Coelli (1996) is a case in point.

[7] The presentation essentially follows Banker and Morey (1986b).

[8] For a transparent presentation on the different sets of axioms underlying the variants of the models to be discussed, see Syrjänen (2004).

[9] We chose an input-oriented approach in this paper. The extension to output-orientation is straightforward.

[10] A referent point is a set of comparable values on the frontier which could be achieved by an observation and thus a part of the reference technology. An efficient observation, which is part of a referent point is called (efficient) peer.