Operations Research Perspectives 1 (2014) 6–17

Contents lists available at ScienceDirect

Operations Research Perspectives

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

Using the latent class approach to cluster firms in benchmarking: An application to the US electricity transmission industry

Manuel Llorca [a,*], Luis Orea [a], Michael G. Pollitt [b]

[a] Oviedo Efficiency Group, Department of Economics, University of Oviedo, Spain

[b] Energy Policy Research Group and Judge Business School, University of Cambridge, United Kingdom

ARTICLE INFO

Article history:
Received 29 January 2014
Received in revised form
26 February 2014
Accepted 18 March 2014
Available online 13 April 2014

JEL classification:
D22
L51
L94

Keywords:
Electricity transmission
Utilities regulation
Latent class model approach
Nonparametric analysis

ABSTRACT

In this paper we advocate using the latent class model (LCM) approach to control for technological differences in traditional efficiency analysis of regulated electricity networks. Our proposal relies on the fact that latent class models are designed to cluster firms by uncovering differences in technology parameters. Moreover, it can be viewed as a supervised method for clustering data that takes into account the same (production or cost) relationship that is analysed later, often using nonparametric frontier techniques. The simulation exercises show that the proposed approach outperforms other sample selection procedures. The proposed methodology is illustrated with an application to a sample of US electricity transmission firms for the period 2001–2009.

1. Introduction

Electricity networks are often regulated by implementing incentive-based regulation schemes that use some types of benchmarking, i.e. a comparison of utilities' performance with best-practice references. As shown by Zhou et al. [1], the nonparametric DEA (Data Envelopment Analysis) has become a very popular tool in energy and environmental studies, especially for benchmarking electric utilities. Unlike the econometric SFA (Stochastic Frontier Analysis) that requires the specification of a particular functional form for the cost or production functions to be estimated, DEA imposes fewer assumptions on the shape of firms' technology and it allows regulators to address traditional convergence problems and the well-known 'wrong skewness problem' in the SFA literature.

A key issue that is sometimes not taken into account by regulators (and researchers) is the heterogeneity or unobserved differences among firms, although utilities are usually quick to mention this issue to the regulators. This concern underlies the negotiations between regulators and utilities, where utilities wield uniqueness as a reason to avoid being compared with their peers. However, it is often assumed in this setting that the whole set of benchmarked firms share the same technology, and hence differences in behaviour are attributed to inefficient use of factors that are under the control of the companies. Possible differences among utilities associated with different technologies are either overlooked or are addressed using simple sample selection procedures, mostly based on factors that may affect performance such as geographic location or utilities' size. Therefore, the efficiency scores obtained from these analyses might be biased and some firms might be penalized (or rewarded) in excess if their underlying technology is less (more) productive than the technology used by other firms operating with more (less) advantageous conditions. This is particularly important in the case of incentive regulation and benchmarking of electricity networks where the results of efficiency analysis have important financial implications for the firms.

In this paper we examine whether we should (a) split the sample arbitrarily on the basis of a single size variable, or (b) use a

* Correspondence to: Faculty of Economics and Business, University of Oviedo, Av. del Cristo, s/n, 33006, Oviedo, Spain. Tel.: +34 985104885.

E-mail addresses: llorcamanuel@gmail.com, llorcamanuel@uniovi.es (M. Llorca).

http://dx.doi.org/10.1016/j.orp.2014.03.002
2214-7160/

comprehensive statistical procedure to control for technological differences, before carrying out a traditional efficiency analysis of regulated electricity networks. We advocate using the so-called latent class model (LCM) approach that allows us to split the electricity networks into a number of different classes, where each class is associated with a different technology. We advocate this approach for several reasons. First, LCM clusters firms by searching for differences in production or cost parameters, which is exactly what regulators are looking for. Second, our approach can be viewed as a "supervised" method for clustering data as it takes into account in the first stage the same (production or cost) relationship that is analysed later, often using nonparametric frontier techniques. Indeed, the literature on data dimension reduction uses this expression for those methods that not only use the information contained in the explanatory variables to be aggregated, but also the information of the dependent variable that will be predicted later on. And third, our approach is not more "technical" than other clustering methods as it can be implemented using standard software and using the same variables that will be used to get efficiency scores in a later stage. Having practicality in mind, we have proposed some simplifications such as the use of simple specifications for both the deterministic (e.g. Cobb–Douglas) and stochastic (e.g. normal distribution) parts of the model to facilitate its application. The use of the same variables in both the latent class stage and the second, DEA, stage also contributes to simplify the use of the proposed procedure.

The same idea is currently being developed by Agrell et al. [2] in a very recent study where they use the LCM approach to control for technological differences in an application to Norwegian power distribution firms. Our paper reinforces the approach from both a theoretical and an empirical point of view. In particular, we carry out a simulation analysis to examine whether the latent class approach outperforms other more arbitrary and less robust procedures for splitting a sample of observations—such as the k-means clustering algorithm or simply using the median of some relevant variables. The simulation exercises confirm our expectations and show that the proposed approach outperforms alternative sample selection procedures. We illustrate this procedure with an application to the US electricity transmission firms examined in [3]. We find two statistically different groups of firms that should be compared or treated separately. In order to confirm the results from the simulation exercise, we compare the partition of the sample obtained through this method with those from alternative clustering procedures.

This paper is organized as follows. Section 2 introduces the two-stage procedure that is proposed to control for unobservable differences in firms' technology (environment) in energy regulation. Section 3 introduces the simulation analysis performed and its main outcomes. Section 4 uses data from the US electricity transmission industry to compare the relative performance of our approach and alternative procedures. Section 5 concludes.

2. A two-stage procedure to address unobserved heterogeneity in utility regulation

As Haney and Pollitt [4] pointed out in a recent survey, regulators have been using several statistical methods to determine the performance of energy utilities. Obtaining reliable measures of firms' performance requires dealing with controllable factors and monitoring for the different environmental conditions under which firms operate. However, both regulators' reports and academic studies do not usually deal with these technological differences. Statistical methods have recently been developed to address this issue. In most of these methods, heterogeneity is understood as an unobserved determinant of the production/cost frontier, while inefficiency is interpreted as the 'distance' to the frontier once heterogeneity has been taken into account.

Following Greene [5,6] we can distinguish two types of models that allow us to achieve our aim, namely the so-called True Fixed Effects (TFE) and True Random Effects (TRE) models introduced by this author, and the LCM, also known as finite mixture models, which have been broadly used in several fields of research (see [7]; or [8], for simple applications; and [9]; or [10], for more comprehensive applications that aim to examine technological gaps using a metafrontier approach). Both approaches have their own strengths and weaknesses. In the TFE/TRE models, unobserved heterogeneity is captured through a set of firm-specific intercepts that are simultaneously estimated with other parameters. Hence, this approach assumes that there are as many technologies as firms. However, as it imposes common slopes for all firms, all of them share the same marginal costs, economies of scale and other technological characteristics.

In contrast to the TFE/TRE models, the LCM approach allows the estimation of different parameters for firms belonging to different groups. This can be easily seen if the general specification of a cost function in this framework is expressed as follows:

$$\ln X_{it} = \alpha_j + \beta_j \ln Y_{it} + v_{it|j} \tag{1}$$

where i stands for firms, t for time and $j = 1, \ldots, J$ for class. X_{it} is a measure of firms' cost, Y_{it} is a vector of explanatory variables, and the random term v_{it} follows a normal distribution with zero mean and variance σ_v^2. As both α_j and β_j, are j-specific parameters, the technological characteristics vary across classes.

Letting θ_j denote all parameters associated with class j, the conditional likelihood function of a firm i belonging to class j is $LF_{ij}(\theta_j)$. The unconditional likelihood for firm i is then obtained as the weighted sum of their j-class likelihood functions, where the weights are the probabilities of class membership, P_{ij}. That is:

$$LF_i(\theta, \delta) = \sum_{j=1}^{J} LF_{ij}(\theta_j) P_{ij}(\delta_j), \quad 0 \le P_{ij}(\delta_j) \le 1,$$

$$\sum_{j=1}^{J} P_{ij}(\delta_j) = 1 \tag{2}$$

where $\theta = (\theta_1, \ldots, \theta_J), \delta = (\delta_1, \ldots, \delta_J)$ and the class probabilities are parameterized as a multinomial logit model:

$$P_{ij}(\delta_j) = \frac{\exp(\delta_j' q_i)}{\sum_{j=1}^{J} \exp(\delta_j' q_i)}, \quad j = 1, \ldots, J, \qquad \delta_J = 0 \tag{3}$$

where q_i is either an intercept or a vector of individual-specific variables. Therefore, the overall likelihood function resulting from (2) and (3) is a continuous function of the vectors of parameters θ and δ, and can be written as:

$$\ln LF\,(\theta, \delta) = \sum_{i=1}^{N} \ln LF_i\,(\theta, \delta) = \sum_{i=1}^{N} \ln \left\{ \sum_{j=1}^{J} LF_{ij}\left(\theta_j\right) P_{ij}\left(\delta_j\right) \right\}. \tag{4}$$

Maximizing the above maximum likelihood gives asymptotically efficient estimates of all parameters. A necessary condition to identify the whole set of parameters is that the sample must be generated from at least two different technologies or two noise terms.

Several comments are in order. First, in this framework each firm belongs to one and only one class.[1] Therefore, the probabilities

[1] This does not mean that a specific firm is going to be always in the same class. The clusters are created without taking into account the panel structure of the data,