Applied Economics Letters, 2008, 1–4, iFirst

Routledge
Taylor & Francis Group

A new method for detecting outliers in Data Envelopment Analysis

Nam Anh Tran[a], Gerald Shively[b,*] and Paul Preckel[b]

[a]*Department of Economics, North Carolina State University, Box 8110, Raleigh, NC 27695, USA*
[b]*Department of Agricultural Economics, Purdue University, 403 West State Street, West Lafayette, IN 47907, USA*

Downloaded By: [Virginia Tech University Libraries] At: 17:29 19 May 2009

We introduce a simple method for detecting outliers in Data Envelopment Analysis. The method is based on two scalar measures. The first is the relative frequency with which an observation appears in the construction of the frontier when testing the efficiency of other observations, and the second is the cumulative weight of an observation in the construction of the frontier. We provide a link to computer programming code for implementing the procedure.

I. Introduction

Data Envelopment Analysis (DEA) is a computationally convenient way to measure efficiency that does not require an explicit functional relationship between inputs and outputs. However, because the frontier is constructed using extreme observations, DEA can be sensitive to extreme points, especially when data may be contaminated by measurement error. In such settings, the technical efficiency scores calculated from datasets that include outliers can be misleading. Timmer (1971) was among the first to recognize the potential sensitivity of computed technical efficiency measures to outliers when linear programming techniques are used to measure efficiency. In other settings, outliers may simply represent outcomes observed accurately, but with low frequency, and hence, worthy of further investigation.

Several approaches to detecting outliers in DEA have been described. Andrews and Pregibon (1978) proposed a geometric method based on calculating the proportion of geometric volume spanned by subsets of the data. The proportion of volume spanned by a sub-set of the data obtained by removing some observations is compared to the volume spanned by the entire dataset. Ratios of these spanning proportions are then used to detect outliers. While conceptually attractive, one drawback with this method is that it can only be applied to firms with a single output. This is especially limiting in light of the fact that one of the most appealing advantages of DEA for efficiency analysis is that the approach easily accommodates multiple outputs.

To overcome this difficulty, Wilson (1993) adapted the geometric approach for use with multiple outputs. Unfortunately, the method is computationally expensive and does not account for the frontier aspect of the problem (Simar, 2003). More recently Cazals *et al.* (2002) proposed a nonparametric efficiency estimator that is robust to extreme observations. This method is based on the concept of an expected minimum input function (or an expected maximum output function). An expected frontier is formed and then pushed away from the data as far as possible. Eventually, some points – the candidate outliers – will not be enveloped by the expected frontier. This approach is useful, but somewhat cumbersome in practice, especially with large datasets.

Here, we propose a simple alternative method to detect outliers based on indices that are constructed

*Corresponding author. E mail: shivelyg@purdue.edu

Applied Economics Letters ISSN 1350 4851 print/ISSN 1466 4291 online © 2008 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/13504850701765119

based on the weights applied to the observations as each one is sequentially tested for efficiency. We define outliers as those observations with large influence on the construction of the efficiency frontier. These impacts can be construed either in terms of the relative frequency with which an observation appears on the frontier, or in terms of the cumulative weight an observation carries when the frontier is being built. These metrics are not identical filters, but in practice they generate highly correlated outcomes. Moreover, the approach is intuitive, computationally simple, and can be easily incorporated into standard DEA computer code.[1]

II. DEA Framework

The DEA approach was introduced by Farrell (1957). Fried *et al.* (1993) provide a comprehensive overview of its use. We adopt the following definitions:

$j = 1, \ldots, n$	an index of firms
$i = 1, \ldots, m$	an index of inputs
$k = 1, \ldots, r$	an index of outputs
$x_j = (x_{1j}, \ldots, x_{mj})$	column vector of inputs of firm j
$y_j = (y_{1j}, \ldots, y_{rj})$	column vector of outputs of firm j
$\lambda = (\lambda_1, \ldots, \lambda_n)$	row vector of nonnegative weights
Θ	a scalar 'shrinking factor.'

We also define weighted combinations of the input and output vectors, namely $x_1\lambda_1 +, \ldots, + x_n\lambda_n$ and $y_1\lambda_1 +, \ldots, + y_n\lambda_n$, where all weights are assumed to be nonnegative. Following the standard approach, if we can find a weighting vector λ that solves the input-oriented linear programming problem:

$$\begin{aligned} &\text{Minimize } \theta \\ &\text{subject to :} \\ &\sum_j \lambda_j y_j \geq y_o \\ &\sum_j \lambda_j x_j \leq \theta x_o \\ &\lambda_j \geq 0 \end{aligned} \tag{1}$$

and $\theta < 1$, then we can say firm (x^0, y^0) is inefficient. Inefficiency here reflects the fact that we can find a weighted combination of firms in the sample that produces equal or greater output with fewer inputs. If, in contrast, the optimal $\theta = 1$, then we conclude firm (x^0,y^0) is efficient in the context of the sample.

Note that no restrictions are placed on λ (the weight vector) in [1], beyond the nonnegativity condition. This means the efficiency score of one firm is calculated based on all the other firms in the dataset, regardless of their size, and the frontier is a linear combination of those firms' inputs and outputs. For a firm on the efficient frontier, if inputs increase by n times then output also increase n times. This constant returns to scale (CRS) approach is only meaningful when firms produce at optimal levels. In reality, many factors preclude firms from operating at optimal levels, and an increase in inputs may result in a nonproportional increase in output. Banker *et al.* (1984) modified Charnes *et al.* (1978) to introduce variable returns to scale (VRS) in DEA. The CRS model can be rewritten for VRS by adding to problem [1] the constraint $\sum_j \lambda_j = 1$. The VRS approach divides the sample into different classes of firms based on size. The most efficient firms within each class form the frontier. In this case, the VRS efficiency frontier is a convex combination of inputs and outputs. The VRS output-oriented DEA model is:

$$\begin{aligned} &\text{Maximize } \theta \\ &\text{subject to :} \\ &\sum_j \lambda_j y_j \geq \theta y_o \\ &\sum_j \lambda_j x_j \leq x_o \\ &\sum_j \lambda_j = 1 \\ &\lambda_j \geq 0. \end{aligned} \tag{2}$$

The efficiency score in the VRS output-oriented DEA model is defined as $1/\theta$.

III. Outlier Detection

Without loss of generality, consider the VRS output-oriented DEA model in which λ is a row vector of weights. Recall that (x^0,y^0) represents a specific producer under consideration; therefore, the efficiency score $1/\theta$ is the technical efficiency score for this producer only. In order to derive technical efficiency scores for all j firms in the sample,

[1] GAMS code and an application of our method are available at www.agecon.purdue.edu/staff/shively/DEA

Downloaded By: [Virginia Tech University Libraries] At: 17:29 19 May 2009

problem [2] must be solved j times. We can arrange the resulting λ values in matrix form:

$$M_\lambda = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \dots & \lambda_{1j} \\ \lambda_{21} & \lambda_{22} & \dots & \lambda_{2j} \\ \dots & \dots & \dots & \dots \\ \lambda_{j1} & \lambda_{j2} & \dots & \lambda_{jj} \end{bmatrix} \tag{3}$$

where the ith row is the weight vector associated with the efficiency test [2] for the ith firm. M_λ contains j rows and j columns, and consists of the weights each observation receives in the process of finding all technical efficiency scores in the sample. We can interpret the weights in M_λ in one of two ways. First, each nonzero weight represents an occasion when one observation appears during the construction of the DEA hull. The total number of occurrences is indicated by the number of times a nonzero value for λ appears in the corresponding column of λ values. Second, we can compute the cumulative weight of one observation across all constructed efficient sets. This is the column sum of all weights for a single observation.

Each of these metrics constitutes a possible method of identifying an outlier. Accordingly, we define two new indexes to represent these λ-weights. We define λ-count (C_j) as the number of times an observation appears during the construction of the DEA hull. It is computed as:

$$C_j = \sum_{j \text{ if } \lambda_{ij} > 0} 1 \tag{4}$$

We define λ-sum (S_j) as the cumulative weight of an observation in all constructed efficient sets. It is computed as:

$$S_j = \sum_j \lambda_{ji} \tag{5}$$

For efficient firms, the DEA model yields nonzero values for λ-count and λ-sum. All inefficient firms have zero values of both λ-count and λ-sum.

To identify outliers, we focus attention only on efficient firms.[2] When we construct the frontier, it is positioned to envelop all observations, including outliers. Based on the values of C_j and S_j, we can identify observations in the dataset that exert an especially strong influence on the construction of the efficient frontier. These observations are potentially outliers.

After identifying an observation with a surprisingly high frequency or level for its weight in the efficiency tests for other firms based on λ-count or λ-sum, one can investigate further and consider dropping the observation from the sample. Doing so results in a new dataset with a sample size of $j - 1$. One then repeats the DEA to obtain new values for C_j and S_j, exclusive of the dropped observation. In an iterative fashion, one can continue to drop those observations with high values of λ-count or λ-sum after each DEA run. The process stops once a desired degree of convergence in the observed weights has been reached. One easy way to identify convergence is through visual interpretation of the data, using a graph that plots iterations on the x-axis and λ-weights for corresponding observations on the y-axis.

IV. Discussion

Our approach for detecting outliers is based on the weights the observations receive during the construction of the DEA hull. It is important to keep in mind that the construction of the efficient frontier differs under CRS and VRS assumptions. With CRS, the frontier consists of a *linear combination* of inputs and outputs of the most efficient firms. The VRS approach consists of a *convex combination* of inputs and outputs of the most efficient firms. For this reason, the λ-weights of observations will typically differ between CRS and VRS models. For CRS, in the one-input one-output case, because the frontier is a linear combination of observations, all weight is placed on the most overly efficient outlier in the dataset; therefore, our method of detecting outliers will find the observation in the sample with the greatest weight every time one calculates weights. For VRS, this need not happen. As the most efficient observations are identified based on the scale of operations, in some cases the highest level of output per unit of input will not be the one with the greatest weight. For CRS, in the one-input one-output case, the most efficient observation is always the one with the largest λ-weight.Our approach to detecting outliers based on dropping observations with the greatest weight is intuitive and analogous to statistical measures of leverage. The procedure is easily incorporated into existing DEA programmes and the necessary λ-weights can be recovered directly as by-products of the DEA computation. In practice, we find that it takes somewhat more iterations to identify outliers in a VRS model than in a CRS model, but we

[2] In terms of identifying the efficiency of other firms in the sample, erroneous outliers that appear as inefficient firms have no deleterious effect on the construction of the frontier. That is, they simply seem to be more or less inefficient than they actually are. Thus, when we refer to outliers, we explicitly refer to observations, which are on the efficient frontier.

Downloaded By: [Virginia Tech University Libraries] At: 17:29 19 May 2009

have found the computational burden of the procedure to be extremely low and the results to be helpful in applied settings, especially when one is confronted with data that may be subject to measurement error.

References

Andrews, D. F. and Pregibon, D. (1978) Finding the outliers that matter, *Journal of the Royal Statistical Society*, **40**, 85 93, Series B (Methodological).

Banker, R. D., Charnes, A. and Cooper, W. W. (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis, *Management Science*, **30**, 1078 92.

Cazals, C., Florens, J. P. and Simar, L. (2002) Nonparametric frontier estimation: a robust approach, *Journal of Econometrics*, **106**, 1 25.

Charnes, A., Cooper, W. W. and Rhodes, E. (1978) Measuring the efficiency of decision making units, *European Journal of Operational Research*, **2**, 429 44.

Farrell, M. (1957) The measurement of productive effi ciency, *Journal of the Royal Statistical Society*, **120**, 253 90, Series A (General).

Fried, H. O., Lovell, C. A. K. and Schmidt, S. S. (Eds.) (1993) *The Measurement of Productive Efficiency*, Oxford University Press, Oxford.

Simar, L. (2003) Detecting outliers in frontier models: a simple approach, *Journal of Productivity Analysis*, **20**, 391 424.

Timmer, C. P. (1971) Using a probabilistic frontier function to measure technical efficiency, *The Journal of Political Economy*, **79**, 776 94.

Wilson, P. (1993) Detecting outliers in deterministic nonparametric frontier models with multiple outputs, *Journal of Business and Economic Statistics*, **11**, 319 23.

Downloaded By: [Virginia Tech University Libraries] At: 17:29 19 May 2009