Dealing with small samples and dimensionality issues in data envelopment analysis

Panagiotis Zervopoulos

University of Western Greece

5. February 2012

Dealing with small samples and dimensionality issues in Data Envelopment Analysis

Panagiotis D. Zervopoulos

Department of Business Administration of Food and Agricultural Enterprises

University of Western Greece, 2 Georgiou Seferi St, Agrinio, Greece

pzervopoulos@uoi.gr

Abstract

Data Envelopment Analysis (DEA) is a widely applied nonparametric method for comparative evaluation of firms' efficiency. A deficiency of DEA is that the efficiency scores assigned to each firm are sensitive to sampling variations, particularly when small samples are used. In addition, an upward bias is present due to dimensionality issues when the sample size is limited compared to the number of inputs and output. As a result, in case of small samples, DEA efficiency scores cannot be considered as reliable measures. The DEA Bootstrap addresses this limitation of the DEA method as it provides the efficiency scores with stochastic properties. However, the DEA Bootstrap is still inappropriate in the presence of small samples. In this context, we introduce a new method that draws on random data generation procedures, unlike Bootstrap which is based on resampling, and Monte Carlo simulations.

Keywords: Data envelopment analysis; Data generation process; Random data; Bootstrap; Bias correction; Efficiency

1. Introduction

Data Envelopment Analysis (DEA) is a widely applied nonparametric method for assessing operational efficiency of homogeneous units. The units or, decision making units (DMUs) involved in the efficiency evaluation process are predominantly a sample of a broader population. Population data are either difficult to collect or unknown. Considering the nonparametric property of DEA, or even its limited statistical underpinning, the yielded efficiency scores are sensitive to sampling variations (Simar and Wilson 1998). Hence, the efficiency scores assigned to the sample DMUs should not be considered as "global" relative assessment measures, but rather solely as "local".

Another issue raised in the DEA literature is associated with the dimensionality "curse" that plagues DEA efficiency scores. A plethora of scholars highlight the upward bias of the DEA efficiency scores when the sample size is inadequate for the number of input and output

variables (Perelman and Santin 2009; Cooper et al. 2007; Simar 2007; Sherman and Zhu 2006; Coelli et al. 2005; Staat 2001; Smith 1997; Banker 1993). Cooper et al. (2007), Zhang and Bartels (1998), and Smith (1997) have defined an appropriate sample size for bias-free estimations of up to 160 units, or a sample adjusted accordingly to the number of utilized input and output variables.

Bootstrap, and particularly the DEA Bootstrap put forth by Simar and Wilson (1998) tackles the problem of reliability of the DEA efficiency scores when sample data are utilized in the evaluation process. The DEA Bootstrap, or smoothed Bootstrap, is a combination of the original Bootstrap (Efron 1979) modified with a smoothing parameter (Silverman 1986) and DEA (Charnes et al. 1978). To be more precise, Simar and Wilson manage to estimate bias in the DEA efficiency scores that is due to sampling variations. They apply a smoothed Bootstrap for generating randomly sampled efficiency scores that are then used for estimating bootstrapped inputs (input-oriented approach) or outputs (output-oriented approach). Subsequently, the bootstrapped inputs or outputs are introduced to the DEA linear programming models for bias-corrected efficiency scores. The DEA Bootstrap inherits the virtues of the original Bootstrap without avoiding though its limitations. A major limitation of the Bootstrap method when it is applied to nonparametric settings is the minimum required sample data for estimating the variability of the population data (Chernick 2008). This weakness is also implied by Efron and Tibshirani (1998). In this context, Chernick (2008) proposed a minimum sample size of 50 observations for estimating reliable scores consistent with the population distribution.

The proposed method overcomes the limitation of Bootstrap, particularly of the DEA Bootstrap, as it yields efficiency scores to DMUs that resemble, more so than those obtained by the DEA Bootstrap, the true efficiency scores when small samples of observations are available. The new method also cures the dimensionality problem of DEA as the adaptability of the estimated sample efficiency scores to the true population scores increases against the DEA Bootstrap results when more input and output variables are incorporated in the production process.

2. Breakdown of the new bias-correction method

The introduced method is not a resampling as Bootstrap, rather it draws on truncated random data generation processes to estimate the unknown population distribution F from the empirical distribution $\hat{F}$.

The scope of the new method is to estimate the population efficiency scores $\Theta = \{\theta_p,\ p = 1, 2, ..., m\}$ by producing an estimator $\hat{F}$ of the population distribution

F from the efficiency scores $\hat{\Theta}=\left\{\hat{\theta}_i,\ i=1,2,...,n\right\}$ defined by DEA. Bias-corrected efficiency scores $\tilde{\Theta}^*=\left\{\tilde{\theta}_i^*,\ i=1,2,...,n\right\}$ are generated by $\hat{F}$ in the pursuit of $\tilde{\Theta}^*\approx\Theta$ and $\tilde{\theta}_i^*\approx\theta_p$.

Let a DMU $u_i\left\{x_k,y_\lambda\right\}_{i=1}^n$ where x_k denotes the *k*-number inputs ($x_k\in\mathbb{R}^+$) and y_λ stands for the λ-number outputs ($y_\lambda\in\mathbb{R}^+$). By applying DEA, for instance, the input-oriented Variable Returns to Scale (VRS) model (Banker et al., 1984)

$$\begin{aligned}&\min\theta\\&s.t.\ X\lambda\le\theta x_o\\&\quad Y\lambda\ge y_o\\&\quad\Lambda=1\\&\quad\lambda\ge0\end{aligned}\qquad(1)$$

we obtain $\hat{\theta}_i$ efficiency scores $\left\{\hat{\theta}_i\middle|0<\hat{\theta}_i\le1\right\}_{i=1}^n$ for every DMU_i. Accordingly, in the case of the output-oriented VRS DEA model, we define $\hat{\varphi}_i$ efficiency scores $\left\{\hat{\varphi}_i\middle|\hat{\varphi}_i\ge1\right\}_{i=1}^n$ for every DMU_i. In the following analysis we presume input orientation is applied.

Based on the efficiency scores (i.e., $\hat{\theta}_i\ \forall\ i=1,2,...,n$) assigned to the sample DMUs, a truncated random data generation process *T* is utilized to produce a sequence of pseudo-numbers $\left\{x^{*\psi}\right\}_{\psi=1}^{\omega}$ for every DMU. Every sequence of pseudo-numbers originates from every single efficiency score or from a combination of a targeted efficiency score and the average scores of the sample.

$$T\left\{\hat{\theta}\middle|\hat{\theta}\text{ to produce }x_o^{*\psi}\ \forall\ \psi=1,2,...,\omega\right\}$$

or
$$T\left\{\hat{\theta}_i\middle|z\hat{\theta}_i+(1-z)n^{-1}\sum_{i=1}^{n}\hat{\theta}_i\text{ to produce }x_{io}^{*\psi}\ \forall\ i=1,2,...,n;\ \psi=1,2,...,\omega\right\}\qquad(2)$$

where
$$x_{io}^{*\psi}=\min\left\{x_i^{*\psi},\hat{\theta}_i\right\}$$

In addition, $T(x^*) \sim N(\hat{\theta}, \widehat{se}^2)$

and $T(x_i^{*\psi}) \sim N(\hat{\theta}_i(\bullet),\ \hat{\theta}_i(\bullet)\widehat{cv})$ (3)

$$\hat{\theta}_i(\bullet) = z\hat{\theta}_i + (1-z)n^{-1}\sum_{i=1}^{n}\hat{\theta}_i$$

where z is a user-defined credibility score that denotes the magnitude of a single efficiency score, and complementary of the sample mean efficiency scores, on the generation of a truncated random sequence of data (scores). In fact, there is inherent dependency between the efficiency scores of the sample DMUs that is due to the comparative assessment procedure applied through DEA.

Moreover, x^* represents the randomly generated data, the $x_{io}^{*\psi}$ expresses selected randomly generated replicas of the efficiency score for the ψ-number elements of the sequence, and cv stands for the coefficient of variation.

The bias-corrected efficiency score for every DMU is defined as follows

$$\tilde{\theta}_i^* = s(x_{io}^{*\psi}) \ \ \forall\ i = 1,2,...,n;\ \ \psi = 1,2,...,\omega \tag{4}$$

where *s* is a statistic (i.e., mean)

It is straightforward that the bias is expressed as

$$\tilde{\theta}_i^* = \hat{\theta}_i - \widehat{bias}_i^{TRDG} \qquad \text{where} \ \ \tilde{\theta}_i^* \in [0,1) \tag{5}$$

The standard error of the proposed truncated random data generation (TRDG) process is

$$\widehat{se}_i^{TRDG} = \left\{ \omega^{-1}\sum_{\psi=1}^{\omega}[x_{io}^{*\psi} - s_i(\bullet)]^2 \right\}^{1/2} \tag{6}$$

where $$s_i(\bullet) = \omega^{-1}\sum_{\psi=1}^{\omega} x_{io}^{*\psi}$$

Taking into account equations (4) and (6), the confidence interval of the bias-corrected efficiency scores are formed as follows

$$\left[\tilde{\theta}^* - t_{(\omega-1)}^{(1-a/2)}\widehat{se}^{TRDG},\ \tilde{\theta}^* + t_{(\omega-1)}^{(1+a/2)}\widehat{se}^{TRDG} \right] \tag{7}$$

where α denotes the level of significance, we prove that

$$\Pr ob\left\{\hat{\theta}_i < \tilde{\theta}_i^{*ub},\ \forall\ i=1,2,...,n\right\}=0 \tag{8}$$

and

$$L^{-1}\sum_{l=1}^{L}\Pr ob_l\left\{\hat{\theta}_i < \tilde{\theta}_i^{*ub},\ \forall\ i=1,2,...,n\right\}=e \tag{9}$$

where *ub* stands for the upper bound of the confidence interval of the bias-corrected efficiency scores. Acknowledging the inherit randomness in the proposed method, all the provided proofs or statements result from iterative procedures. In formulation (9), the probability, that is the average of L=1000 iterations, is equal to an infinitesimal value. The cases in which this infinitesimal probability is present are identified and presented in order to be avoided by the user of the proposed method.

The inherit randomness in the proposed method is regarded as a drawback because it is a source of instability for the obtained results when the method is applied repeatedly. To overcome this drawback, a stabilization parameter γ is introduced in the procedure that eliminates up to 99% the variation of the bias-corrected scores. The parameter γ expresses the number of iterations for the formulations (2)-(7). The reported results are average scores.

The proposed method for dealing with sampling variations and dimensionality issues in DEA is expressed by the following function

$$f_{\hat{\theta}}(\alpha, cv, z, \omega, \gamma, n^{ex}, \text{var}^{ex}) \equiv \tilde{\theta}^{*TRDG} \tag{10}$$

In formulation (10), two exogenous parameters n^{ex} and var^{ex} are included which denote the number of DMUs in the original sample and the number of input and output variables, respectively, that are utilized for defining the efficiency scores through DEA. These two parameters implicitly influence the bias-correction procedure.

Based on a numerical example and on the results that are tested through Monte Carlo so that to eliminate randomness, the proposed method yields better estimators ($\tilde{\theta}^{*TRDG}$) for the population efficiency scores (θ) than the DEA Bootstrap ($\tilde{\theta}^{*boot}$) when the original sample consists of less than 50 DMUs. In addition, the adaptive power of $\tilde{\theta}^{*TRDG}$s to θ increases against $\tilde{\theta}^{*boot}$s when the number of input and output variables increases.

3. Conclusion

In this paper, a new method for correcting bias in DEA efficiency scores is presented. Commonly, DEA yields overestimated efficiency scores when sample data rather than population data are used, and the number of DMUs is limited compared to the number of variables. In some studies, adequate sample sizes have been determined for obtaining unbiased efficiency scores. However, in many cases the required sample size cannot be collected (e.g., automobile industry, power companies, water companies).

In this paper is presented a new method for correcting bias in DEA efficiency scores when small samples are available (i.e., $n<50$ DMUs). The new method enhances the applicability of DEA when the DEA Bootstrap fails due to the limited number of DMUs under evaluation, or the inadequate sample size compared to the number of input and output variables. The new approach does not draw on resampling but on an iterative truncated random number generation procedure. Despite the inherit randomness of the new method, the results are robust and the proposed procedure does not suffer from instability. In addition, it is proved that the results obtained by the proposed method are more adaptive to reality than those estimated by the DEA Bootstrap when small samples are available.

References

Banker RD (1993) Maximum-Likelihood, Consistency and Data Envelopment Analysis - a Statistical Foundation. Manage Sci 39 (10):1265-1273

Charnes A, Cooper WW, Rhodes E (1978) Measuring the efficiency of decision making units. Eur J Oper Res 2 (6):429-444

Chernick MR (2008) Bootstrap Methods: A Guide for Practitioners and Researchers. John Wiley & Sons, New Jersey

Coelli T, Rao P, O'Donnell CJ, Battese G (2005) An Introduction to Efficiency and Productivity Analysis. Springer, New York

Cooper WW, Seiford LM, Tone K (2007) Data envelopment analysis: a comprehensive text with models, applications, references and DEA-Solver software. 2nd edn. Springer Science + Business Media, New York

Efron B (1979) Bootstrap methods; another look at the jacknife. Annals of Statistics 7:1-26

Efron B, Tibshirani RJ (1998) An Introduction to the Bootstrap. Chapman & Hall/CRC, Boca Raton

Perelman S, Santin D (2009) How to generate regularly behaved production data? A Monte Carlo experimentation of DEA scale efficiency measurement. Eur J Oper Res 19:303-310

Sherman HD, Zhu J (2006) Benchmarking with quality-adjusted DEA (Q-DEA) to seek lower-cost high-quality service: Evidence from a US bank application. Ann Oper Res 145:301-319. doi:DOI 10.1007/s10479-006-0037-4

Silverman BW (1986) Density Estimation for Statistics and Data Analysis. Chapman and Hall, London

Simar L (2007) How to improve the performances of DEA/FDH estimators in the presence of noise? J Prod Anal 28:183-201

Simar L, Wilson PW (1998) Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models. Manage Sci 44 (1):49-61

Smith P (1997) Model misspecification in Data Envelopment Analysis. Ann Oper Res 73:233-252

Staat M (2001) The effect of sample size on the mean efficiency in DEA: Comment. J Prod Anal 15:129-137

Zhang Y, Bartels R (1998) The effect of sample size on the mean efficiency in DEA with an applicatino to electricity distribution in Australia, Sweden and New Zealand. J Prod Anal 9:187-204