Productivity and Efficiency Analysis

2) Data envelopment analysis

d) Statistical approach to DEA

Timo Kuosmanen

Aalto University School of Business

https://people.aalto.fi/timo.kuosmanen

Econometric critique of DEA

Schmidt (1985) phrased the criticism as follows:

> "I am very skeptical of non-statistical measurement exercises, certainly as they are now carried out and perhaps in any way in which they could be carried out.... I see no virtue whatever in a non-statistical approach to data." p. 296

- DEA "models" phrased as linear programming problems
- What is the **model** that DEA is trying to "**estimate**"?

DEA model vs estimator

- Model

  $T = \{(\mathbf{x},\mathbf{y}) \mid \text{inputs } \mathbf{x} \text{ can produce outputs } \mathbf{y}\}$

- DEA estimator

  $T^{DEA\text{-}CRS} = \{(\mathbf{x},\mathbf{y}) \mid \mathbf{x} \geq \mathbf{X\lambda};\ \mathbf{y} \leq \mathbf{Y\lambda};\ \boldsymbol{\lambda} \geq \mathbf{0}\}$

- **Statistical consistency**: DEA estimator is statistically consistent if $T^{DEA\text{-}CRS} \rightarrow T$ as $n \rightarrow \infty$.

- But how to model the data generating process? Which assumptions are needed ?

Farrell (1957, *J. Royal Stat. Soc.*)

However, if one looks at it from a statistical point of view, one may wish to reformulate the problem as follows. There exists some efficient function, from which all the observed points deviate randomly but in the same direction. What, then, is the best estimate of the function?

It might seem at first sight as though the method chosen involved a great waste of information—only a handful of "efficient" points contribute directly to the estimate, the rest being "ignored". However, in the problem of estimating the extremes of the rectangular distribution, the relevant information is all contained in the extreme observations, and this seems to be an analogous case. On the other hand, the estimate chosen will almost certainly have a pessimistic bias (that is, a bias away from the origin) but in the absence of *a priori* knowledge of the distribution of the deviations, it is impossible to remove this bias.

Similarly, errors of observation will introduce an optimistic bias, which can only be eliminated if the distributions of both errors and efficiencies are known. This is an interesting problem for any theoretical statistician but for practical purposes the important fact is that if the errors are small compared with the variation in efficiencies, this bias will be negligible. In fact, the

Statistical basis of DEA

Banker (1993), *Man. Sci.*

DEA production function (compare with Afriat, 1972):

$$y_0^* = g^*(\mathbf{x}_0) = \text{Max}\left\{ y \,|\, y = \sum_{j=1}^{n} \lambda_j y_j, \quad \sum_{j=1}^{n} \lambda_j \mathbf{x}_j \leq \mathbf{x}_0, \right.$$
$$\left. \sum_{j=1}^{n} \lambda_j = 1, \quad \lambda_j \geq 0 \right\}. \quad (2)$$

Assumptions regarding the data generating process:

POSTULATE 3A. Envelopment. *Efficiency deviations* ϵ_j *are independently and identically distributed with probability density function* $f(\epsilon)$ *such that* $f(\epsilon) = 0$ *for all* $\epsilon < 0$.

POSTULATE 4A. Monotonicity of Density Function. *If* $0 \leq \epsilon' \leq \epsilon''$ *then* $f(\epsilon') \geq f(\epsilon'')$.

Statistical basis of DEA

Banker (1993), *Man. Sci.*

Nonparametric maximum likelihood estimator:

$$\underset{f(\cdot),g(\cdot)}{\text{Maximize}} \quad \prod_{j=1}^{n} f(\epsilon_j = g(\mathbf{x}_j) - y_j) \tag{1}$$

subject to

$g(\cdot)$ is monotone increasing and concave, and (1.1)

$f(\epsilon) = 0$ for $\epsilon < 0$, that is

$$\epsilon_j = g(\mathbf{x}_j) - y_j \geq 0. \tag{1.2}$$

DEA is MLE:

PROPOSITION 2. *If the probability density function $f(\epsilon)$ satisfies Postulates 3A and 4A, and $\mathbf{x}_j$ and ϵ_j are independently distributed, then the optimal solutions $y_j^* = g^*(\mathbf{x}_j)$ solving (2) for $j = 1, \cdots n$, and $\epsilon_j^* = g^*(\mathbf{x}_j) - y_j$ solve the MLE problem in (1). The piecewise-linear function estimated by solving (2) is a maximum likelihood estimate of the production function $g(\cdot)$ in (1) for all $\mathbf{x} \in X^*$.*

Statistical basis of DEA

Banker (1993), *Man. Sci.*

Statistical consistency:

PROPOSITION 5. *If the production frontier* $g(\mathbf{x})$ *is monotone increasing and concave for* $\mathbf{x} \in X$, *where* X *is a convex and compact subset of* R^m, *and if* $\mathbf{x}$ *and* ϵ *are independently distributed with probability density functions* $h(\cdot)$ *and* $f(\cdot)$ *such that* $h(\mathbf{x}) > 0$ *for all* $\mathbf{x} \in X$, $f(\epsilon) = 0$ *for* $\epsilon < 0$, *and* $F(\epsilon) = \int_{-\infty}^{\epsilon} f(\epsilon)\, d\epsilon > 0$ *for all* $\epsilon > 0$, *then the DEA estimators* $g^*(\mathbf{x})$ *are weakly consistent for all* $\mathbf{x}$ *in the interior of* X.

Statistical basis of DEA

Kneip, Park & Simar (1998), *Ectr. Th.*

- Consistency in the multi-output case
- Rate of convergence:

$$\widehat{\theta}_{DEA}(x_0, y_0) - \theta(x_0, y_0) = O_p\left(n^{-\frac{2}{p+q+1}}\right)$$

Here: $p + q$ inputs and outputs

- The curse of dimensionality
- Analogous to Stone's (1980) optimal rate of convergence for nonparametric estimators

Statistical inference - Bootstrapping
Simar & Wilson (1998 MS, 2000 *JPA)*

Smooth consistent bootstrap

1) Apply DEA to estimate efficiencies θ_i.
2) Apply nonparametric kernel density estimator to θ_i.
3) Draw random pseudo-inefficiencies η_i from the kernel density for each observation $(\mathbf{x}_i, \mathbf{y}_i)$.
4) Form a pseudo-sample $\{\mathbf{x}_i, \eta_i(\mathbf{y}_i / \theta_i)\}_b$
5) Apply DEA to pseudo-sample to estimate efficiencies $\{\varphi_i\}_b$
6) Compute the bootstrap estimate $E_{ib} = \varphi_{ib} / \eta_{ib}$
7) Repeat steps 2)-6) B times to obtain the bootstrap distribution $\{\mathbf{E}_1, \ldots, \mathbf{E}_B\}$, which can be used statistical inferences (hypothesis testing, confidence intervals)

Statistical inference - Bootstrapping
Simar & Wilson (1998 MS, 2000 *JPA)*

Variants of the so-called ***naive bootstrap*** either (i) use the empirical distribution of $\{(x_i, y_i),\ i = 1, \ldots, n\}$ to estimate $f(x, y)$; or (ii) alternatively, the empirical distribution of $\{(\widehat{\theta}_i, \eta_i, y_i),\ i = 1, \ldots, n\}$) is used to estimate $f(\theta, \eta, y)$. In the first variation,

The crucial aspect of the procedure revolves around how the density estimate $\widehat{f}(\theta)$ is defined. The naive bootstrap proposed by Ferrier and Hirschberg (1997) estimates $f(\theta)$ via the empirical density function, placing a mass $1/n$ at each observed $\widehat{\theta}_i$; (*i.e.*, Ferrier and Hirschberg resample θ_i^* by drawing independently, with replacement, from $\{\widehat{\theta}_i \mid i = 1, \ldots, n\}$). As discussed in Simar and Wilson (1999a, 1999b), this naive bootstrap is inconsistent.

5. Yet Another Bad Bootstrap Idea

While we have addressed the problems inherent in the naive bootstrap elsewhere, as noted in the previous section, another unfortunately bad idea has recently appeared. Löthgren and Tambour (1996, 1997) and Löthgren (1997, 1998) employ a peculiar variant of the homogeneous naive bootstrap described above in section (4.3.1). In their method, bootstrap values θ_i^* are drawn from the empirical distribution of the original efficiency estimates $\widehat{\theta}_i$ as in the ordinary homogeneous naive bootstrap. For the input-orientation, pseudo data are

Statistical inference in DEA: Parametric tests vs bootstrap

- Banker (1993) proposes asymptotic tests based on exponential or half-normally distributed inefficiency

- Bootstrap avoids the distributional assumptions, but it is still an *asymptotic* procedure. S&W (2000) note:

bootstrap approximation depends on both the number of replications B and and the sample size n. The approximation becomes exact as $B \rightarrow \infty$ and $n \rightarrow \infty$.

Bootstrap does correct for noise

- Bootstrap inferences in DEA only consider the sampling error
- Noise is assumed away. S&W (2000) conclude:

> While bootstrap methods offer a tractable approach to statistical inference in DEA or FDH models, a larger, remaining challenge is to find a way for allowing stochastic noise in the data

- DEA confidence intervals that ignore noise are often very tight
- If data are noisy, applying bootstrap bias correction can make things worse

Next lesson

2e) Pre-history of DEA in economics