The Journal of Productivity Analysis, 7, 161–176 (1996)

Statistical Inference and Nonparametric Efficiency: A Selective Survey

S. GROSSKOPF
Department of Economics, Southern Illinois University, Carbondale, IL 62901-4515

Abstract

The purpose of this paper is to provide a brief and selective survey of statistical inference in nonparametric, deterministic, linear programming-based frontier models. The survey starts with nonparametric regularity tests, sensitivity analysis, two-stage analysis with regression, and nonparametric statistical tests. It then turns to the more recent literature which shows that DEA-type estimators are maximum likelihood, and, more importantly the results concerning the asymptotic properties of these estimators. Also included is a discussion of recent attempts to employ resampling methods to derive empirical distributions for hypothesis testing.

Keywords. Statistical inference, nonparametric efficiency, DEA estimators

1. Introduction

The purpose of this paper is to provide a brief survey of statistical inference as applied to nonparametric efficiency measurement, which I mean to include what is generally referred to as data envelopment analysis (DEA) as well as Farrell efficiency and what has come to be known as FDH (free disposal hull). The general conception of the outside observer is that there is no statistical inference employed by those using these nonparametric methods. This impression arises from the fact that this approach typically employs linear programming models[1] which are classified as nonparametric, but deterministic. Researchers have, however, been fairly resourceful in finding ways to pursue hypothesis testing, for example. This paper gives a brief, and by necessity selective, overview of statistical inference in the nonparametric, deterministic, linear programming-based frontier models. I emphasize that this survey is not comprehensive—it is meant to sketch out some of the current approaches and give a few examples. These examples and methods will reflect the fact that I am a practitioner and an economist. That means that the traditional DEA, which is based in the operations research school of thought, will be underrepresented; my apologies to those whom I have neglected.

As a point of departure, we begin with Afriat (1972) who recognized Farrell's (1957) model as a nonparametric test of neoclassical production functions and foresaw the need for statistical inference in this context. The next section takes up where the Farrell literature left off; namely with attempts to apply some kind of sensitivity analysis to the basic Farrell model. Next we turn to what has become the practitioner's approach to statistical inference: the two- (or more) stage approach. The general idea is to first compute an efficiency index, and then use that as data in a second stage, generally employing some kind of regression analysis, most recently including nonparametric density estimation at the second stage.

Section 4 briefly mentions nonparametric tests as an option for statistical inference before turning to recent work which shows that DEA (and FDH) are maximum likelihood, and also provides asymptotic properties of those estimators. Section 6 includes recent attempts to employ resampling methods like bootstrapping and jackknifing to derive empirical distributions as a basis for statistical inference, particularly in small samples. In the last section we come back to some of the ideas suggested by Afriat in 1972 and picked up later by Varian (1990) to derive tests of goodness-of-fit that have some economic content.

2. Production Technology and Nonparametric Efficiency

The main goal of the nonparametric efficiency literature, including DEA, is to provide a measure of performance. In the context of production theory, one can think of many ways to measure performance. In the absence of prices, which was the main arena in which DEA was originally applied, performance by the firm (or decision-making unit) was based on its ability to produce outputs from inputs relative to best practice in some relevant group. In a multiple-output environment, a natural definition of technology is the distance function. Following Shephard (1970) or see Färe (1988), the output distance function is defined as

$$D_o(x, y) = \inf \{\theta : (x, y/\theta) \in S\}, \tag{1}$$

where[2]

$$S = \{x \in \mathfrak{R}_+^N : x \text{ can produce } y \in \mathfrak{R}_+^M\}. \tag{2}$$

As is well-known, the distance function is reciprocal to what Färe, Grosskopf, and Lovell (1985, 1994) have dubbed the output-oriented measure of Farrell technical efficiency.[3] For the scalar output case it is easy to show that the output distance function is equivalent to the ratio of observed to maximum potential output. That is

$$\begin{aligned} D_o(x, y) &= \inf \{\theta : y/\theta \leq f(x)\} \\ &= \inf \{\theta : 1/\theta \leq f(x)/y\} \\ &= y/f(x), \end{aligned} \tag{3}$$

where $f(x) = \max\{y : (x, y) \in S\}$ is a scalar-valued production function. This provides the intuition behind the efficiency measure; it is the ratio of (the size of) observed output relative to (the size of) maximum potential (or best practice) output.

These distance functions/efficiency measures can be computed as solutions to simple linear programming problems, for example as the solution to the following problem for each observation $k' = 1, \ldots, K$

$$(D_o(x, y))^{-1} = \max \theta_{k'} \tag{4}$$

subject to

$$\sum_k z_k y_{km} \geq \theta_{k'} y_{k'm}, \quad m = 1, \ldots, M$$

$$\sum_k z_k x_{kn} \leq x_{k'n}, \quad n = 1, \ldots, N$$

$$z_k \geq 0, \quad k = 1, \ldots, K.$$

The ability to compute distance functions (alias efficiency) so easily has contributed to their usefulness and popularity. The fact that this computational procedure does not require specifying a parametric functional form also proved useful in the production literature on nonparametric regularity tests. This was recognized by Afriat (1972), who suggested using nested linear programming models to test for consistency of a set of data with the properties of neoclassical production functions, including tests for disposability, nondecreasing concave, and classical constant returns to scale. He saw this as an alternative to imposing an ad hoc functional form and estimating Cobb-Douglas or CES production functions which he found to be too restrictive.[4]

He also recognized that Farrell's measure of productive efficiency could prove useful as a measure of goodness of fit:

> Another kind of objection can be brought to bear on a determination of error, apart from error in the data, which is not based on an economic concept. Economic exactitude is efficiency so an economic error can be expressed as an inefficiency. A Euclidean sum of squares is in itself devoid of economic meaning. Farrell's efficiency method is safe from such objection but it bears on some econometric methods. Any distance however large is negligible economically if it corresponds to a negligible difference in the economic account. In economic analysis error can have expression in terms of failure to reach an optimum value, and this gives basis for an economic principle of estimation or approximation which is generally applicable as the commonly used least squares. (p. 569)

This idea was later taken up by Varian (1990), which we discuss briefly in the last section.

Afriat goes on to discuss how to go about providing a test of the efficiency hypothesis which does not require that every observation have an efficiency score of one. He suggests using a simple model for a distribution of efficiency scores such as the beta distribution, since "with neither any theory which brings a particular (probability) mechanism in to view, nor any empirical basis for a particular form of distribution, a simple model is appropriate." (p. 579) Then one can determine the parameters of the distribution and efficiency scores under maximum likelihood. Variations on this theme were taken up by Banker and Maindiratta (1992), Banker (1993), and Korostelev et al. (1992, 1995).

3. Sensitivity Analysis

Those using the programming approach to efficiency measurement have always been aware that their technique does not explicitly account for the possibility of "noisy" data, i.e.,

measurement error and natural randomness. The fact that the "frontier" is determined by extreme points in the data (which might be extreme because of data errors) has branded this approach as sensitive to outliers.[5] We note that the stochastic frontier models typically assume that their independent variables are deterministic and not contaminated by noise as well, however, that approach explicitly allows for noise in the dependent variable, although not generally explicitly for outliers (but see the papers which do address this issue discussed at the end of this section).

Perhaps the first attempt to modify the original Farrell model to adjust for the possibility of the frontier being determined by contaminated data was due to Timmer (1971). He solved the following linear programming problem:

$$\min \sum_{j=1}^{n} E_{jt} = \sum_{j=1}^{n} \sum_{i=1}^{m} \alpha_{ij} X_{ijt} - \sum_{j=1}^{n} Y_{jt} \tag{5}$$

subject to

$$\sum_{i=0}^{m} \alpha_i X_{ijt} \geq Y_j t, \quad j = 1, \ldots, n. \tag{6}$$

The formulation of this problem follows the industry frontier production function proposed by Aigner and Chu (1968). Timmer proposed restating the constraints as probability statements, with the constraint holding with some externally specified probability. As a practical matter he suggested discarding the first prespecified

> percent of the efficient observations until a prespecified level of P (the probability level) is reached. Alternatively efficient observations might be discarded one at a time until the resulting estimated coefficients stabilize. (p. 782)

This general idea was taken up and formalized as chance-constrained programming and applied to the nonparametric DEA problem by a number of authors. Like Timmer, these authors added probabilities to the input and output constraints. This allows for stochastic inputs and outputs, but the statistical properties of the frontier and associated efficiency measures are not known.[6] See Charnes and Cooper (1963), Land et al. (1988, 1993), and Olesen and Petersen (1995) for this approach, and Desai et al. (1994) for a critique of this approach, as well as a Monte Carlo simulation which they propose as an alternative, which is essentially the bootstrapping approach discussed in Section 6.

Others have also employed various sorts of sensitivity analysis in their applications. See Sengupta (1987) and Valdmanis (1992) for some examples of the variety of approaches taken in this vein. Although these sensitivity tests give us some idea of the robustness of the results to sample size, specification of inputs and outputs, measurement of inputs and outputs, etc., they do not provide us with a basis for statistical inference.

The problem of dealing with messy data and outliers has also been taken up using alternative approaches. Seaver and Triantis (1989) address this issue in the context of parametric frontiers. More recent contributions include Seaver and Triantis (1992) and Wilson (1993, 1995).

4. Practical Approaches: The Two-Step Procedure

Perhaps since DEA and its relatives represent a practical applied tool, researchers using this approach appealed to practical fixes to go about undertaking hypothesis testing. It soon became clear that computing individual efficiency scores really was not enough for either consulting purposes or for policy analysis. One of the first questions that arises is the following: Even if these efficiency scores were measured perfectly with perfect data and represented significant deviations from best practice performance, how do we explain their variation? I can't speak for those in operations research, but the economist naturally reaches for a familiar tool: regression analysis.[7]

The basic idea of the two-step procedure is to treat the efficiency scores as data or indexes and use linear regression to explain the variation in the efficiency scores. This implied specifying a host of independent variables intended to explain inefficiency. Perhaps the first refinement of this model was to account for the fact that the efficiency scores were censored: Depending on the type of efficiency score computed, we do not observe values of the efficiency score above (below) values of unity. As a consequence, OLS was not appropriate; a model which explicitly accounted for the fact that the dependent variable was limited was preferred. One of the first studies to my knowledge to worry about this issue was Lovell, Walters, and Wood (1995).

Other conceptual issues arise. Perhaps the most intuitive is the fact that the variables in the second stage were obviously expected to affect performance, so why weren't they included in the original model?[8] More subtle, and perhaps more troubling, is the issue of the data generating process, and the related issue of distribution of the errors. If the variables used in specifying the original efficiency model are correlated with the explanatory variables used in the second stage, then the second-stage estimates will be inconsistent and biased. See Deprins and Simar (1989) and Simar, Lovell, and Vanden Eeckaut (1994).

Another variation of the two-step procedure suggested by Thiry and Tulkens (1992) is provided in Simar (1992). In this paper, Simar compares several approaches to computing efficiency in the presence of panel data, including what he calls a semi-parametric approach. Following Thiry and Tulkens, in this approach efficiency is first computed using FDH (although in principle DEA could also be employed). The undominated firms are then employed to estimate a parametric frontier, i.e., the FDH performs the role of a filter.[9] As noted by Simar, however, the analytic sampling distribution of the efficiency measures so computed is unknown; therefore he proposes using bootstrapping to provide an empirical distribution and hence means of assessing the statistical significance of the estimators.[10]

A further variation on the theme was to combine DEA, bootstrapping and nonparametric density estimation. This hybrid approach is discussed in the section on bootstrapping methods.

5. Nonparametric Statistical Tests

Since efficiency scores computed using DEA and its relatives are based on a nonparametric method, it is natural to appeal to nonparametric statistics to provide a basis for statistical inference. This approach is appealing in the sense that many of them are distribution free,

i.e., there is no need to presume an underlying normal distribution, for example, to proceed with these tests.

Because of its ready availability, several researchers have used the nonparametric tests which are included in the SAS procedure NPAR1WAY and RANK. The tests in NPAR1WAY include a battery of simple linear rank statistics. These can be used to determine if the distribution of efficiency scores has the same location parameter (for example the median) across various groups. Also included are several statistics based on the empirical distribution of the sample, which allow the researcher to test if the distribution of a variable is the same across different groups.

The rank statistics in the NPAR1WAY procedure are based on the following:

$$S = \sum_{j=1}^{m} c_j a(R_j) \tag{7}$$

where R_j is the rank of observation j; c_j is the class variable which identifies the group to which an observation belongs, and $a(R_j)$ is the rank score. The NPAR1WAY procedure computes the following rank order statistics: Wilcoxon scores (based on the sum of the ranks of the efficiency scores in each group), median scores (which assign a score of 1 to observations above the median and 0 to observations below the median and sums these by group), Van der Waerden scores (which use approximations of the expected value of the efficiency scores under a normal distribution as the rank score), and Savage scores (which use expected values of order statistics based on the exponential distribution). Clearly the usefulness of these tests for DEA applications is limited by the fact that with DEA, there are typically a mass of scores with identical rank, namely one. These tests vary in power with respect to the form of the underlying distribution. As noted in the SAS manual, the Wilcoxon scores "are locally most powerful for location shifts of a logistic distribution," "median scores are locally most powerful for double exponential distributions," Van der Waerden scores "are powerful for normal distributions" and Savage scores are "powerful for comparing scale differences in exponential distributions or location shifts in extreme value distributions." (p. 717)

The statistics computed based on empirical distribution functions of the sample include the Kolmogorov-Smirnov statistic, the Cramer-von Mises statistic, and the Kuiper statistic. The Kolmogorov-Smirnov statistic was employed in Sengupta (1987), and recommended by Banker (1993).

6. DEA (and FHD) as Maximum Likelihood Estimators

The best hope for finding a statistical foundation for nonparametric efficiency measures was hinted at by Afriat (1972)—namely maximum likelihood. Schmidt (1976) showed that the Aigner and Chu parametric frontier models (which are deterministic and computed as solutions to linear or quadratic programming problems) are equivalent to maximum likelihood estimation (with an exponential probability density function of the efficiency scores for the linear programming specification, and half-normal for the quadratic programming specification).

Banker (1993) modified the maximum likelihood model used by Schmidt to show that DEA is also maximum likelihood

> with the principal difference being the specification of the production frontier in DEA as a nonparametric monotone increasing and concave function, instead of a parametric form linear in the parameters. (p. 1266)

Perhaps more importantly, Banker also shows that the DEA estimators are consistent,[11] establishing asymptotic statistical properties of DEA, as well as suggesting statistics to be used in hypothesis testing.

Independently, Korostelev, Simar, and Tsybakov (1992, 1995) established that FDH and DEA were maximum likelihood estimators of the boundary of a set, where that boundary is either a monotone (FDH) or convex and monotone (DEA) function of its arguments. They too analyzed the asymptotic statistical properties of these estimators. They derive the rate of convergence of the FDH and DEA estimators and show that no other estimator converges at a faster rate. Specifically the rate of convergence is $n^{-1/(s+1)}$ for FDH, and $n^{-2/(s+2)}$ for DEA, where s is the number of inputs and outputs and n is the number of observations. This is obviously a very important result.

These results finally established a statistical foundation for nonparametric efficiency measures, at least in terms of asymptotic properties. Unfortunately, the bulk of applied work in this area is based on relatively small samples. Thus, the nature of the small sample properties of these estimators was clearly the next order of business. Banker (1993) cautions that these

> ...results should be interpreted very cautiously, at least until systematic evidence is obtained from Monte Carlo experimentation with finite samples of varying sizes. (p. 1272)

Banker and Chang (1994, 1995) provide simulation results for the means tests proposed in Banker (1993). Banker and Chang (1993) propose DEA tests of returns to scale and provide simulation evidence, and Banker, Chang, and Sinha (1994) propose DEA input substitution tests and simulation evidence. For a nice summary of this work see Banker (1995) in this volume.

The Monte Carlo experiment challenge was also taken up by Kittelsen (1995) who provides evidence as to the small sample properties of DEA as efficiency estimators based on a Monte Carlo study. He was particularly interested in addressing the problems noted by Banker (1993), especially the problems of bias and nonindependence. As stated by the author, his paper "aims at providing some simulation evidence on the bias of the DEA efficiency estimators, and the approximating power of hypothesis tests suggested in the literature. (p. 2)

The issue of bias arises in DEA since it provides estimates of efficiency relative to the sample employed. If one is willing to assume that there is no underlying model or reference technology, then these relative measures might be assumed to be observed without bias. (Kittelsen claims that this is one interpretation of the FDH approach.) On the other hand, if one believes in an underlying model, the issue of bias arises. As stated by the author, "the problem of bias in DEA follows from the fact that the probability of observing a truly

efficient unit in a sample is less than one" (p. 9). Kittelsen notes that this bias increases with the dimensionality of the problem: for example, since the VRS model includes one more constraint than the CRS model (and the models are nested), the bias of the VRS model can be no less than the bias of the CRS model.

The underlying model employed by Kittelsen is a simple single-output production model. The DEA model he employs is the constant returns to scale (CRS) input-saving model, equivalent to the original Farrell measure of technical efficiency. He includes seven experiments, including the base trial in which technology is a CRS linear function with one input and one output, $F(y, x) = y - x$. The output is generated randomly from a normal distribution, and input is computed as $x = (1 - \gamma)y$, where γ is the inefficiency term generated randomly from a half-normal distribution (in the base trial). Each trial includes 1000 samples; in the base trial each sample has 100 observations. The author computes the two F-tests proposed by Banker (1993) (one based on a half-normal distribution, the other exponential) as well as two t-tests for equality of group means. For most trials the CRS and VRS efficiency measures were computed.

The trials performed include the following:

A. Base trial: True model CRS, half-normal inefficiency, $n = 100$.
B. Varying sample size: Same as above, but sample size is varied from $n = 20$, to $n = 1000$.
C. Magnitude of inefficiency: $n = 100$, inefficiency half-normal but with different levels.
D. Distribution of inefficiency: $n = 100$, but distribution of inefficiency compared for half-normal, gamma, and exponential.
E. Distribution of output: $n = 100$, y distributed as normal, uniform and lognormal, same mean, and standard deviation.
F. Cobb-Douglas technology: $n = 100$, number of inputs varies from one to three.
G. Misspecification: $n = 100$, irrelevant input included.

The author draws two major conclusions from his experiments: (1) That applications of the Banker F-tests or simple t-tests should be based on a split sample rather than a pooled sample (avoiding dependence between the two samples being tested); (2) "bias is important, increasing with dimensionality, and decreasing with sample size, average efficiency, and a high density of observations near the frontier." (p. 27) The author suggests that for small samples, bootstrapping may be a better approach. We take this up in the following section.

7. Resampling Methods

DEA and its nonparametric relatives were developed as empirical methods; only recently are there serious attempts to try to provide it with some statistical underpinnings. An appealing alternative is to use an empirical approach to estimate the distributions of the statistics we are interested in instead of trying to figure out the underlying model and analytical distribution. This is the general idea of resampling methods like bootstrapping and jackknifing.

Some early attempts to use resampling in the general context of nonparametric efficiency measurement include Färe, Grosskopf, and Weber (1989), Yaisawarng (1988), Grosskopf and Yaisawarng (1990), and Ferrier et al. (1993). The last two studies used resampling

to measure economies of scope. Resampling was used to derive a distribution of measures of scope in order to compute a point estimate for each observation based on the mean of this empirical distribution.[12] The first two studies used jackknifing for the same general purpose. These studies did not exploit the possibility of constructing confidence intervals from the empirical distributions.

As mentioned in the discussion of two-step procedures, Simar (1991) was perhaps the first published paper to propose using the bootstrap for computing confidence intervals for efficiency scores derived from nonparametric frontier methods (in his application these were computed using the FDH frontier). His approach implemented bootstrapping based on the residuals from the second-stage estimation. Later, Hall, Härdle, and Simar (1995) proposed using an interated bootstrap procedure.

One use of the bootstrap, then, is to provide an empirical distribution of efficiency scores for each observation in the sample. An example of this approach is Ferrier and Hirschberg (1994) who apply this technique to a sample of Italian banks. They proceed as follows: first, efficiency scores are computed according to the standard Farrell input based technical efficiency under variable returns to scale. The resulting scores are considered to be the original statistics. The next step is to use these scores to correct the input data such that all observations are on the frontier. Next, efficiency scores are randomly drawn from the original distribution and used to construct pseudo input data by multiplying the corrected inputs by the randomly drawn efficiency scores. Efficiency is recomputed using the pseudo data for the reference technology but maintaining the original data for the observation under evaluation. (As we shall see, the fact that they include the original data for the observation under evaluation in the reference technology causes problems, particularly for observations originally found to be efficient.) The last two steps of this procedure are repeated 1000 times. These new efficiency scores are used to derive an empirical distribution of efficiency for each observation.

In another application Ferrier and Hirschberg (1995) apply the same bootstrap technique described above, namely Davison, Hinkley and Schechtman's (1986) balanced bootstrap, to data they used in a previous study to determine the efficiency of buildings in terms of their climate control technology. Here, as before, they set the number of bootstraps at 1000. Again, the goal of these two studies is to provide a means of determining the precision of efficiency estimates computed using DEA. Their technique does provide a means of constructing empirical distributions and therefore confidence intervals based on the percentage method for those observations which are initially found to be inefficient. It does not necessarily provide a basis for judging the precision of scores of those observations found to be initially on the best practice frontier. As the authors put it "the strength and the drawback of the bootstrap is that it is so firmly based on the original data. (p. 4)

Even though the bootstrapping techniques employed by Ferrier and Hirschberg are based on the original distribution of efficiency scores, the bootstrap can still be employed to provide some evidence of the bias of the original scores. Following Efron (1979), the authors use the difference between the mean of the bootstrap replications and the original efficiency score to illustrate that the original efficiency score is biased upwards (see Ferrier and Hirschberg (1994)).

Atkinson and Wilson (1995) provide a bootstrap methodology for constructing approximate confidence intervals for mean efficiency scores in small samples. They propose using a

small sample correction of the original efficiency scores, then resampling from the corrected data with replacement to compute the mean. The resampling procedure is repeated to arrive at an empirical distribution of means. Next they use the Efron percentile method to construct confidence intervals for the means. As an alternative they suggest the bias-corrected and accelerated method described by Efron and Tibshirani (1993).

Färe and Whittaker (1995) use bootstrapping in yet another context. Their rationale for using bootstrapping is to compensate for the survey data they used, which was not representative of the total population from which the survey was drawn.

> The use of nonparametric efficiency measurement and the subsequent application of bootstrapping and kernel density estimation to the results allow inferences to be drawn concerning the whole population. (abstract).

The purpose of their paper was to compare the empirical performance of two models of dairy production: one which accounts for intermediate production of grain/feed, and a second which ignores the intermediate production. Both are formulated as linear programming problems; the latter looks like an output-oriented Farrell technical efficiency measure.[13] The procedure they used was to employ a modified bootstrap for complex survey data proposed by Sitter (1992) to draw an independently, identically distributed (iid) sample of DEA estimates[14] for the computing models. There were 100 bootstrap samples drawn for each model. The next step was to compute a kernel density estimate of the probability density function (at 50 points in an interval between 1 and 2.5). This was repeated for each bootstrap sample. Since the distribution of efficiency measures is skewed, the authors used the median as a measure of central tendency. The median and the 95% confidence interval were estimated for each of the 50 points of the density estimates.

The nonparametric density estimation used by Färe and Whittaker provides a means of graphically inspecting the (estimated) distribution of the (medians of the) efficiency measures and the error bounds. They conclude that for their results from the standard DEA model (without intermediate products), the estimated medians are not meaningful, which implies that "the estimated confidence intervals are meaningless." (p. 14) This is due to the fact that the dimensionality of the problem is high (i.e., almost all farms are efficient), therefore the median is on the frontier. This notion of dimensionality was formalized by Thrall (1988). It is also known as a boundary problem.

Simar and Wilson (1995) and Wilson and Simar (1995) suggest an alternative bootstrap method which takes into account boundary problems associated with using the original empirical distribution of efficiency measures as the basis for resampling. Their goal is to use the bootstrap to analyze the sensitivity of efficiency scores to the sampling variations of the estimated frontier. In Simar and Wilson (1995), they argue that the bootstrap should be designed to mimic the data generating process underlying the efficiency measurement. The data generating process they suggest is an improvement on that used by Ferrier and Hirschberg in that all of the data in the reference technologies are treated the same way in the bootstrap, as discussed below. In addition, they address the boundary problem of the empirical distribution (i.e., the truncation at value one) by employing the Silverman (1986) reflection method combined with the nonparametric estimation method used in Färe and Whittaker. Their method yields a bootstrap that is consistent and unbiased.

More specifically they employ the following procedure:

1. For each observation, compute the efficiency score in the usual way.
2. Define the empirical density function putting mass $1/n$ on the individual estimated efficiency scores.
3. Generate a random sample of size n of the efficiency scores using the reflection method and kernel density estimation.
4. Construct pseudo data by first correcting the data for inefficiency based on the original efficiency scores, then multiplying the corrected data by the randomly drawn scores from step 3.
5. Compute the bootstrap estimate of efficiency using the data from step 4 to form the reference technology. The observation being evaluated takes on its original value (but not in the reference technology, in contrast to Ferrier and Hirschberg). As usual, efficiency scores are computed for each observation.
6. Repeat steps 3–5 a large number of times, B, providing B bootstrap estimates of efficiency for each observation in the data set.

These smoothed bootstrap estimates can then be used to construct confidence intervals for the original efficiency scores of the individual observations. Simar and Wilson (1995) provide one empirical illustration; Wilson and Simar (1995) include three empirical illustrations. Gstach (1994) has also proposed a combination of bootstrapping and density estimation that may prove fruitful.

We note that bootstrapping is also a viable alternative to provide statistical precision to the efficiency scores derived from stochastic frontier models. This has been employed by Grosskopf and Hayes (1993), and Grosskopf, Hayes, and Hirschberg (1995) to name two examples.

8. Goodness-of-Fit and Measurement Error

"It is perhaps also instructive to look at the frequency distribution of efficiencies.... It is to such frequency distributions that one must look for a measure of the success of the analysis, corresponding to the multiple correlation coefficient in regression analysis." (Farrell, 1957)

This section takes us back to Farrell (1957) and Afriat (1972). The quote above suggests that Farrell had thought about the idea of the efficiency distribution providing a measure of goodness-of-fit that has some economic content. This was also recognized by Afriat (1972), see the quote in the introductory section. Varian (1990) picks up on the same ideas:

> What matters for most purposes in economics is not whether a...violation of the optimizing model is *statistically* significant, but whether it is *economically* significant.... Hence the conventional methods are lacking in two senses: first, they have an excess reliance on parametric forms, and second, they test for statistically significant violations of optimization rather than economically significant violations. (p. 116)

Varian proposes using a measure of the size of the deviations from optimizing behavior as a measure of goodness-of-fit. For example, he proposes using the percent by which a firm departs from cost minimization as a measure of goodness-of-fit, and recognizes, as did Farrell, that ". . .the distribution of these measures. . . may be of considerable interest themselves." (p. 130)

In an earlier paper, Varian (1985) focuses explicitly on nonparametric analysis of optimizing behavior. In our context, we can follow up on Afriat's idea that one may think of Farrell type efficiency measures as nonparametric tests of consistency of a set of data with the regularity conditions implied by neoclassical production functions, for example. This is discussed in Färe and Grosskopf (forthcoming).

Varian also noted that due to measurement error, for example, some data may violate consistency which in the Farrell context, means their efficiency scores are not exactly one. He suggests allowing for approximate satisfaction, or specification of a weak consistency condition.[15] Banker and Maindiratta (1988) suggest searching for the largest set of observations that do satisfy the regularity conditions. Following Varian, they also point out that DEA can be used to provide an inner approximation to technology.

9. Summing Up

This paper was intended to provide a brief and selective overview of statistical inference and nonparametric efficiency measuring from the viewpoint of an economist/practitioner. From this perspective, it seems to me that there has been considerable progress—more than I realized before writing this—toward providing reasonable and practical tools for pursuing statistical inference and hypothesis testing. I was also surprised to discover that there really might be something called "statistical underpinnings of DEA." Last fall, when I first saw that this was my suggested topic, I considered that idea to be some kind of oxymoron. Now, thanks to efforts espeically by Banker and Korostelev, Simar and Tsybakov, we can really talk about DEA estimators.

I look forward to further work based on Monte Carlo studies of the small sample properties of these estimators. I also still find the idea of using empirical methods like bootstrapping to construct confidence intervals, for example, extremely appealing, and think that Simar and Wilson (1995) have suggested a very reasonable way of thinking about the data generating process, as well as providing innovative solutions to the boundary and bias problems involved with applying bootstrapping to nonparametric technical efficiency. Extending their ideas to allocative efficiency, overall efficiency and productivity is perhaps the next order of business. I also like the idea, suggested by Simar during the meetings, of first applying nonparametric density estimation to the data and then proceeding with nonparametric efficiency measurement.

Acknowledgments

I am grateful for comments from Rolf Färe, Rajiv Banker, Léopold Simar, Paul Wilson, and Gary Ferrier.

Notes

1. There are exceptions. FDH does not require computation of a linear programming problem. Also, some formulations of efficiency problems may also require specification of nonlinear programming problems.
2. The input distance function is defined as $D_i(y, x) = \sup\{\lambda : (x/\lambda, y) \in S\}$.
3. The reciprocal of the input distance function is the better-known input-oriented measure of Farrell technical efficiency.
4. This nonparametric test approach was further developed by Diewert and Parkan (1983), Hanoch and Rothschild (1972), Varian (1984), and Banker and Maindiratta (1988).
5. Outliers which are due to measurement error will have their most extreme effect if measurement error falsely results in observations determining the frontier, since that will then affect the efficiency scores of all the observations for which that observation forms the basis, causing their inefficiency to be overstated. If measurement error only occurs in nonextremal observations, the problem is less severe—it will only distort the score of that particular observation. On the other hand, if some observations are extreme but not due to measurement error, the programming approach certainly provides a good screening device to identify such observations.
6. The stochastic DEA literature has also been subject to this criticism.
7. In the case of those using the traditional DEA models, this can become a three step procedure: compute the Farrell type efficiency scores, correct the data and go back and remove slacks to arrive at a revised score, then use regression.
8. This issue also arose in the stochastic frontier literature, see Reifschneider and Stevenson (1991).
9. This approach is no longer recommended by Simar. He suggests using nonparametric smoothing methods to first remove "noise," then use the cleaned data to proceed with nonparametric efficiency measurement.
10. In later work, Hall, Härdle, and Simar (1995) propose using an iterated bootstrap in the panel data regression context. They use an iterated bootstrap to correct a percentile confidence interval (in their case, there were ties in the maximum of the intercept in a fixed effects model).
11. Specifically, "when the true production function is monotone increasing and concave, DEA estimators are consistent if the probability of arbitrarily small deviations ϵ is strictly positive," pp. 1266–1267.
12. The computation of economies of scope required construction of pseudo data which satisfied the hypothesis of strict additivity.
13. The authors also calculated maximal revenue, which allowed them to derive the output analog of a Farrell decomposition.
14. This is required for the nonparametric density estimation.
15. This is presumably also the motivation behind Timmer's probabilistic frontier and the associated work on chance-constrained programming and stochastic DEA discussed earlier.

References

Afriat, S. (1972). "Efficiency Estimation of Production Functions." *International Economic Review* 13:3, Oct, 568–598.

Aigner, D.J. and S.F. Chu. (1968). "On Estimating the Industry Production Function." *American Economic Review* 58, 226–239.

Atkinson, S. and P. Wilson. (1995). "Comparing Mean Efficiency and Productivity Scores from Small Samples: A Bootstrap Methodology." *Journal of Productivity Analysis*.

Banker, R.D. (1993). "Maximum Likelihood, Consistency and Data Envelopment Analysis: A Statistical Foundation." *Management Science* 39:10, 1265–1273.

Banker, R.D. (1995). "Hypothesis Testing Using Data Envelopment Analysis." *Journal of Productivity Analysis*, this volume.

Banker, R.D. and H. Chang. (1993). "Tests of Returns to Scale for Monotone Concave Production Functions." Working paper.

Banker, R.D. and H. Chang. (1995). "A Simulation Study of Hypothesis Tests for Differences in Efficiencies." *International Journal of Production Economics*..

Banker, R.D. and H. Chang. (1995). "A Simulation Study of Efficiency Differences for Multiple Outputs with Measurement Error." Working paper.

Banker, R.D., H. Chang, and K.K. Sinha. (1994). "Tests to Evaluate the Separability or Substitutability of Inputs to a Production System." Working paper.

Banker, R.D and A. Maindiratta. (1988). "Nonparametric Analysis of Technical and Allocative Efficiencies in Production." *Econometrica* 56:6, 1315–1332.

Banker, R.D. and A. Maindiratta. (1992). "Maximum Likelihood Estimation of Monotone and Concave Production Functions." *Journal of Productivity Analysis* 3:4, 401–415.

Charnes, A. and W.W. Cooper. (1963). "Deterministic Equivalence for Optimizing and Satisficing under Chance Constraints." *Operations Research* 11:1, 18–39.

Davison, A.C., D.V. Hinkley, and E. Schechtman. (1986). "Efficient Bootstrap Simulation." *Biometrika* 73, 555–566.

Deprins, D. and L. Simar. (1989a). "Estimation de Frontière Dèterministes avec Facteurs Exogéne d'Inefficacité. *Annales d'Economie et de Statistique* 14, 117–150.

Deprins, D. and L. Simar. (1989b). "Estimating Technical Inefficiencies with Corrections for Environmental Conditions with an Application to Railway Companies." *Annals of Public and Cooperative Economics* 60:1, Jan-Mar, 81–102.

Desai, Anad, Samuel J. Ratick, and Arie Schinnar. (1994). "DEA with Stochastic Variations in Data." Working Paper Series 94-51, November, Max M. Fisher College of Business, The Ohio State University.

Diewert, W.E. and C. Parkan. (1983). "Linear Programming Tests of Regularity Conditions for Production Frontiers." In W. Eichhorn, et al. (eds.), *Quantitative Studies of Production and Prices*, Würzburg and Vienna: Physica-Verlag.

Efron, B. (1979). "Bootstrapping Methods: Another Look at the Jackknife." *Annals of Statistics* 7, 1–26.

Efron, B. and R.J. Tibshirani. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Färe, R. (1988). *Fundamentals of Production Theory*. Berlin: Springer-Verlag.

Färe, R. and S. Grosskopf. "Nonparametric Tests of Regularity, Farrell Efficiency and Goodness-of-Fit." *Journal of Econometrics* (forthcoming).

Färe, R., S. Grosskopf, and C.A.K. Lovell. (1985). *The Measurement of Efficiency of Production*. Boston: Kluwer-Nijhoff.

Färe, R., S. Grosskopf, and C.A.K. Lovell. (1994). *Production Frontiers*. Cambridge, U.K.: Cambridge University Press.

Färe, R., S. Grosskopf, and W. Weber. (1989). "Measuring School District Performance." *Public Finance Quarterly* 17:4, 409–429.

Färe, R. and G. Whittaker. (1995). "An Intermediate Input Model of Dairy Production Using Complex Survey Data." *Journal of Agricultural Economics* 46:2, 201–213.

Farrell, M.J. (1957). "The Measurement of Productive Efficiency." *Journal of the Royal Statistical Society*, Series A, General, 125:2, 252–267.

Ferrier, G., S. Grosskopf, K. Hayes, and S. Yaisawarng. (1993). "Economies of Diversification in the Banking Industry: A Frontier Approach." *Journal of Monetary Economics* 31, 229–249.

Ferrier, G.D. and J.G. Hirschberg. (forthcoming). "Bootstrapping DEA Efficiency Scores: With an Application to Italian Banks." *Journal of Productivity Analysis*.

Ferrier, G.D. and J.G. Hirschberg. (1995). "A Form of Stochastic Data Envelopment Analysis: Applying the Bootstrap to DEA." Mimeo.

Grosskopf, S. and K. Hayes. (1993). "Local Public Sector Bureaucrats and Their Input Choices." *Journal of Urban Economics* 33, 151–166.

Grosskopf, S., K. Hayes, and J. Hirschberg. (1995). "Fiscal Stress and the Production of Public Safety: A Distance Function Approach." *Journal of Public Econimics* 57, 277–296.

Grosskopf, S. and S. Yaisawarng. (1990). "Economices of Scope in the Provision of Local Public Services." *National Tax Journal* 43, 61–74.

Gstach, Dieter. (1994). "The Right Answers from the Wrong Model?" Mimeo, Vienna University for Economics and Business Administration.

Hall, Peter, Wolfgang Härdle, and Lèopold Simar. (1995). "Iterated Bootstrap with Applications to Frontier Models." *Journal of Productivity Analysis* 6:1, 63–76.

Hanoch, G. and M. Rothschild. (1972). "Testing the Assumptions of Production Theory: A Nonparametric Approach." *Journal of Political Economy* 89:4, 878–892.

Kittelsen, Sverre. (1995). "Monte Carlo Simulations of DEA Efficiency Measures and Hypothesis Tests." In *Using Data Envelopment Analysis to Measure Production Efficiency in the Public Sector.* Thesis for the degree of Dr. Polit. at the Department of Economics, University of Oslo, also presented at the Georgia Productivity Workshop, October 1994.

Korostelev, A.P., L. Simar, and A.B. Tsybakov. (1992). "Efficient Estimation of Monotone Boundaries." Discussion paper 9209, Institut de Statistique, Université Catholique de Louvain, Louvain-la-Neuve, to appear in *Annals of Statistics*.

Korostelev, A.P., L. Simar, and A.B. Tsybakov. (1995). "On Estimation of Monotone and Convex Boundaries." *Publications de l'Institut de Statistique de l'Université de Paris* 39:1, 3–18.

Land, K.C., C.A.K. Lovell, and S. Thore. (1988). "Chance-Constrained Efficiency Analysis." Working paper, Department of Economics, University of North Carolina, Chapel Hill, NC.

Land, K.C., C.A.K. Lovell, and S. Thore. (1993). "Chance-Constrained Data Envelopment Analysis." *Managerial and Decision Economics* 14:6, 541–554.

Lovell, C.A.K., L.C. Walters, and L.L. Wood. (1995). "Stratified Models of Education Production using Modified DEA and Regression Analysis." In A. Charnes, W.W. Cooper, A.Y. Lewin, and L.M. Seiford, (eds). *Data Envelopment Analysis: Theory, Methodology and Applications*, Boston, Kluwer, 329–352.

Olesen, O.B. and N.C. Petersen. (1995). "Chance Constrained Efficiency Evaluation." *Management Science* 41:3, 442–457.

Reifschneider, D. and R. Stevenson. (1991). "Systematic Departures from the Frontier: A Framework for the Analysis of Firm Inefficiency." *International Economic Review* 32:3, 715–724.

Schmidt, P. (1976). "On the Statistical Estimation of Parametric Frontier Production Functions." *Review of Economics and Statistics* May, 238–239.

Seaver, B.L. and K.P. Triantis. (1989). "The Implications of Using Messy Data to Estimate Production-Frontier-Based Technical Efficiency Measures." *The Journal of Business and Economic Statistics* 7, 49–59.

Seaver, B.L. and K. P. Triantis. (1992). "A Fuzzy Clustering Approach Used in Evaluating Technical Efficiency Measures in Manufacturing." *Journal of Productivity Analysis* 3:4, 337–363.

Sengupta, J. (1987). "Data Envelopment Analysis for Efficiency Measurement in the Stochastic Case." *Computers and Operations Research* 14:2, 117–129.

Shephard, R.W. (1970). *Theory of Cost and Production Functions*. Princeton, NJ: Princeton University Press.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Simar, L. (1992). "Estimating Efficiencies from Frontier Models with Panel Data: A Comparison of Parametric, Non-Parametric and Semi-Parametric Methods with Bootstrapping." *Journal of Productivity Analysis* 3:1/2, 171–191.

Simar, L., C.A.K. Lovell, and P. Vanden Eeckaut. (1994). "Stochastic Frontiers Incorporating Exogenous Influences on Efficiency." Discussion Paper 9403, Institut de Statistique, Université Catholique de Louvain, Belgium.

Simar, L. and P.W. Wilson. (1995). "Sensitivity Analysis of Efficiency Scores: How to Bootstrap in Nonparametric Frontier Models." Mimeo.

Sitter, R.R. (1992). "A Resampling Procedure for Complex Survey Data." *Journal of the American Statistical Association* 87, 755–765.

Thiry, B. and H. Tulkens. (1989). "Productivity, Efficiency and Technical Progress: Concepts and Measurement." *Annales de L'Economie Publique Sociale et Coopérative* 60:1, 9–42.

Thiry, B. and H. Tulkens. (1992). "Allowing for Technical Inefficiency in Parametric Estimates of Production Functions." *Journal of Productivity Analysis* 3:1/2, 45–66.

Thrall, R.M. (1988). "Classification Transitions Under Expansion of Inputs and Outputs in Data Envelopment Analysis." *Managerial and Decision Economics* 10, 159–162.

Timmer, C.P. (1971). "Using a Probabilistic Frontier Production Function to Measure Technical Efficiency." *Journal of Political Economy* 79, 776–794.

Valdmanis, V. (1992). "Sensitivity Analysis for DEA Models: An Empirical Example Using Public vs. NFP Hospitals." *Journal of Public Economics* 48, 185–205.

Varian, H. (1984). "The Nonparametric Approach to Production Analysis." *Econometrica* 52:3, May, 579–597.

Varian H. (1985). "Nonparametric Analysis of Optimizing Behavior with Measurement Error." *Journal of Econometrics* 30:1/2, 445–458.

Varian, H. (1990). "Goodness-of-Fit in Demand Analysis." *Journal of Econometrics* 46, 125–140.

Wilson, Paul W. (1993). "Detecting Outliers in Deterministic Nonparametric Frontier Models with Multiple Outputs." *Journal of Business & Economic Statistics*, 11:3, 319–323.

Wilson, Paul W. (1995). "Detecting Influential Observations in Data Envelopment Analysis." *Journal of Productivity Analysis* 6:1, 27–46.

Wilson, Paul W. and Lèopold Simar. (1995). "Bootstrap Estimation for Nonparametric Efficiency Estimates." Mimeo.

Yaisawarng, S. (1989). "Recovering Short-Run Price Efficiency: Theory and Application." Ph.D. dissertation, Southern Illinois University, Carbondale, IL.