School of Economic Sciences

Working Paper Series
WP 2006-8

Inferring the Latent Incidence of Inefficiency from DEA Estimates and Bayesian Priors

By

D. Friesner, R. Mittelhammer, and R. Rosenman

August 2006

WASHINGTON STATE UNIVERSITY
World Class. Face to Face.

Inferring the Latent Incidence of Inefficiency from DEA Estimates and Bayesian Priors

August 2006

Daniel Friesner [a]
Ron Mittelhammer [b*]
Robert Rosenman [b]

a School of Business Administration
Gonzaga University
502 E. Boone Ave.
Spokane, WA 99258

b School of Economic Sciences
Washington State University
P.O. Box 646210
Pullman, WA 99164-6210
Email: mittelha@wsu.edu
Telephone: (509) 335-1706
Fax: (509) 335-1173

* To whom correspondence should be addressed.

JEL Classification C11

Keywords: Data Envelopment Analysis, latent inefficiency, Bayesian inference, Beta priors, posterior incidence of inefficiency

Abstract

Data envelopment analysis (DEA) is among the most popular empirical tools for measuring cost and productive efficiency. Because DEA is a linear programming technique, establishing formal statistical properties for outcomes is difficult. We show that the incidence of inefficiency within a population of Decision Making Units (DMUs) is a latent variable, with DEA outcomes providing only noisy sample-based categorizations of inefficiency. We then use a Bayesian approach to infer an appropriate posterior distribution for the incidence of inefficient DMUs based on a random sample of DEA outcomes and a prior distribution on the incidence of inefficiency. The methodology applies to both finite and infinite populations, and to sampling DMUs with and without replacement, and accounts for the noise in the DEA characterization of inefficiency within a coherent Bayesian approach to the problem. The result is an appropriately up-scaled, noise-adjusted inference regarding the incidence of inefficiency in a population of DMUs.

Inferring the Latent Incidence of Inefficiency from DEA Estimates and Bayesian Priors

Abstract

Data envelopment analysis (DEA) is among the most popular empirical tools for measuring cost and productive efficiency. Because DEA is a linear programming technique, establishing formal statistical properties for outcomes is difficult. We show that the incidence of inefficiency within a population of Decision Making Units (DMUs) is a latent variable, with DEA outcomes providing only noisy sample-based categorizations of inefficiency. We then use a Bayesian approach to infer an appropriate posterior distribution for the incidence of inefficient DMUs based on a random sample of DEA outcomes and a prior distribution on the incidence of inefficiency. The methodology applies to both finite and infinite populations, and to sampling DMUs with and without replacement, and accounts for the noise in the DEA characterization of inefficiency within a coherent Bayesian approach to the problem. The result is an appropriately up-scaled, noise-adjusted inference regarding the incidence of inefficiency in a population of DMUs.

1.0 Introduction

Data envelopment analysis (DEA) is among the most popular tools for measuring productive and cost efficiency. Originally developed by Charnes *et al* (1978) to empirically quantify the distance functions posited by Debreu (1951) and Farrell (1957), DEA uses a linear programming algorithm to evaluate the efficiency of decision-making units (DMUs) on a 0 - 1 interval. An efficiency score of 0 implies that a particular DMU is completely inefficient, while a DMU with a score of 1 indicates that the DMU is producing at a point on the efficient frontier (whether the frontier is defined as an isoquant or a production possibilities frontier), and thus completely efficient.

Using DEA to characterize efficiency is advantageous for several reasons. First, DEA, in general, does not have stringent data requirements; the researcher need only collect data on the relevant inputs and outputs for each DMU. Additionally, DEA is “non-parametric” in that it does not require the researcher to make stringent functional assumptions about the technology underlying the production function. Instead, the technology set is inferred using the observed input-output relationships within the data. Third, unlike other alternatives, such as stochastic frontier analysis (SFA), DEA more easily generates efficiency estimates when DMUs produce multiple outputs. Finally, DEA does not force the researcher to make assumptions about the scale (either constant or variable returns to scale) of the production process. As with the technology, scale economies are deduced using observed input-output relationships.

Unfortunately, DEA also has at least two major drawbacks. First, DEA estimates tend to mismeasure efficiency in finite samples, and in particular to over-predict efficiency because the randomly collected samples may not contain enough truly efficient DMUs to accurately characterize the efficient production frontier.[1] As such, the frontier calculated by DEA will be above (below) the true efficient isoquant (production possibilities frontier), because the most efficient DMUs in the sample (against which the estimated frontier is calculated) may not lie on the true production frontier.

Several solutions to the small sample mismeasurement problem have been proposed. One "apparently obvious" solution suggested in the literature is to apply DEA only when the sample size is very large, especially relative to the number of inputs and outputs included in the analysis. This approach is often not practically viable because researchers frequently do not have access to extensive data sets. Moreover, we note as a consequence of the analysis we present ahead that even a census of the population may not suffice to solve the DEA mismeasurement problem, which appears to be an underappreciated property of DEA.

Secondly, and arguably the most important drawback to DEA, is that its statistical foundation is complex (Schmidt 1985). This is problematic for researchers aiming to determine the relationship between efficiency and a set of exogenous policy variables. A common use of DEA is in a two-stage empirical analysis, where the first step entails calculating DEA estimates, and the second step uses these estimates as outcomes of the dependent variable in a regression analysis to determine how exogenous policy variables influence efficiency (Simar and Wilson 2005). Unfortunately, the two stage approach

[1] In the DEA literature the mismeasurement is usually referred to as "bias" and the nature of DEA makes it possible to overestimate but not underestimate DMU efficiency. To avoid confusion with the statistical property of bias, we do not use bias to refer to DEA mismeasurement. The mismeasurement problem is exacerbated when the number of inputs and outputs is large relative to the sample size (the well-known "curse of dimensionality").

forces the researcher to make additional assumptions that may lead to detrimental statistical consequences. One possibility is to assume that DEA scores follow a particular distribution, which allows the researcher to employ standard maximum likelihood regression techniques. Past studies have made a number of different distributional assumptions, including the normal distribution (estimated via OLS) (Ray 1991; Chirkos and Sears 1994; Stanton 2002) the truncated normal, or Tobit distribution (Chilingerian 1995; Rosenman and Friesner 2004), the Beta distribution (Sengupta 1998) and the Beta-Binomial distribution (Sohn and Choi 2006). A drawback to this approach is specification bias; if one assumes an incorrect distribution, which appears rather likely in this complex statistical context, any coefficient estimates generated by this approach will be generally biased and inconsistent. Moreover, as Simar and Wilson (2005) note, because the efficient frontier is calculated relative to the DMUs in the data, DEA scores are serially correlated in an unknown and complicated manner. Thus, even if the distributional choice is correct, any coefficient estimates may be substantially inefficient unless the distributional choice accounts explicitly for this correlation.[2]

An alternative espoused by Simar and Wilson (2005) is to employ semiparametric regression techniques to estimate the second stage model. The authors suggest supplementing an MLE-based regression with a specific form of bootstrapping to adjust for any potential mismeasurement and serial correlation. However, one must still identify an appropriate distribution for the likelihood function, which allows for the possibility of

[2] Two studies (Hirschberg and Lloyd 2002; Xue and Harker 2002) adjust their regression estimates for serial correlation. However, Simar and Wilson (1999a, b) indicate that their approach still leads to inconsistent estimates.

specification bias. Moreover, as we already noted, a census of DMUs does not necessarily fully remove mismeasurement and thus bootstrap methods cannot solve these problems fully.[3]

Recent research has attempted to avoid these issues by deriving specific asymptotic distributions for DEA scores, as well as the asymptotic rate of convergence at which randomly sampled DEA scores converge to these distributions. Banker (1993) proved that a distribution of DEA scores exists, and established the consistency of DEA scores for the single output case. Kneip, Park and Simar (1998) identified the asymptotic rate of convergence for DEA estimates, while Gijbels *et al*. (1999) derived the asymptotic distribution for DEA scores involving a single output and a single input. Kneip, Simar and Wilson (2003) extended Gijbels' findings to the multiple input, multiple output case, while Jeong (2004) derived a more empirically tractable version of Kneip, Simar and Wilson's distribution. All of this work has depended on specific regularity conditions, with none of this work dealing with the empirically realistic case of relatively small finite population sizes and random sampling of DMUs without replacement.

Jeong's (2004) work not withstanding, the asymptotic distributions derived in past studies are not easily implemented because they do not belong to traditional parametric families, and nonparametric methods, numerical integration or other approximation techniques are generally required in order to construct confidence intervals for DEA scores. Another shortcoming is that the assumptions necessary to derive these distributions do not necessarily apply to a wide range of empirical DEA studies. For

[3] Wilson (2005) has released FEAR 1.0, a statistical package that not only calculates DEA scores, but also implements the bootstrapping mechanism outlined above.

example, virtually all studies mentioned above assume that the inputs and outputs are iid random vectors. However, if researchers are sampling from finite populations without replacement, this assumption will clearly not literally hold. Additionally, studies such as Gijbels *et al*. (1999) and Kneip, Simar and Wilson (2003) assume that the production frontier is smooth and twice continuously differentiable. In the absence of these assumptions the distributions and rates of convergence identified in these studies will generally not apply.

In this paper we treat the true incidence[4] of inefficiency in a population of DMUs as a latent variable and the sample DEA estimates as a collection of sample observations that provide useful but noisy information relating to the value of the latent variable. Together with a prior distribution relating to the incidence of inefficiency, which can be either informative or uninformative, we derive a posterior distribution for the incidence of inefficiency that accounts for inherent DEA noise, and provides a means for inferring the incidence of inefficiency as well as testing hypotheses relating to it. The approach places little *a priori* structure on the nature of the production process being studied so that inferences are applicable in very general problem contexts such as cases where the efficient frontier is not (twice) continuously differentiable or even in cases where the technology is not representable via a parametric functional form. Moreover, the geneses of the probability distributions used in the statistical model emanate directly from the

[4] One implication of our analysis is that it may be impossible to accurately characterize the distribution of DEA scores. Thus, our focus on the incidence of inefficiency may represent the most statistically coherent context in which an analysis can be performed.

sampling methods and the prior information employed, and thus errors in distributional specification are fully mitigated.

The remainder of this paper proceeds as follows. First, we identify an appropriate posterior probability distribution for the true incidence of inefficiency in a population of DMUs based on random samples of DEA estimates under the idealized assumption that sample outcomes categorize DMUs correctly as to whether they are truly inefficient. While the empirical relevance of this model is limited, analysis of this case provides results that are relevant to the existent DEA literature, and allows the analysis to focus initially on only the issue of sampling variability, and abstract from the issue of mismeasurement in the DEA information. A posterior distribution is derived using both a general and an uninformative Bayesian prior relating to the proportion of inefficiency in the population. We then turn to the empirically more relevant case where the DEA estimates are known to not necessarily characterize DMU inefficiency accurately. A posterior distribution for the incidence of inefficiency that accounts for both sampling variability and potential mismeasurement in the DEA estimates is derived. We then provide a brief numerical example to illustrate implementation of the methodology to the most empirically relevant case of random sampling, without replacement, from a finite population of DMUs. The penultimate section of the paper discusses implications of our results for the DEA literature, and we conclude the paper by discussing some limitations of our findings and as well as suggestions for future extensions of the work.

2.0 General Problem Context

Consider a population of DMUs all of which operate in the same industry. Each of these DMUs utilizes a $p \times 1$ vector of inputs to produce a $q \times 1$vector of outputs. Following the notation of Fare and Primont (1995) and Simar and Wilson (2005), let w $\in$ $R^p_{\geq 0}$ denote the vector of inputs and y $\in$ $R^q_{\geq 0}$ denote the vector of outputs. The conceptual set of feasible production possibilities, which subsumes all of the production technologies of the DMUs in the population under study, is defined by:

$$T = \{(w, y): w \in R^p_{\geq 0}, y \in R^q_{\geq 0}: w \text{ can produce } y\} \tag{1}$$

The definition of the set T is generally unknown and is to be approximated by sample observations on the behavior of DMUs in the population (Coelli *et al.*, p. 134, footnote 3).

Following Farrell (1957), we measure the amount of inefficiency using the (output-oriented) distance function, d: $R^p_{\geq 0}$ x $R^q_{\geq 0} \rightarrow R_{\geq 0} \cup \{+\infty\}$.[5] Thus, for any input-output combination (w, y), d can be defined as

$$d = d(w, y) \equiv \inf_{\theta}\{\theta \mid (w, \frac{y}{\theta}) \in T\} \tag{2}$$

Assuming non-zero input and output vectors, any input-output combination for which distance is equal to 1, i.e. $\theta = 1$, is efficient, while distances less than 1 indicate inefficiency with values closer to zero indicating increasing levels of inefficiency. Our goal is to draw inferences about the incidence of inefficiency in the population of DMUs

[5] While our focus is on output measures of efficiency, we note that analogous results can easily be obtained for input distance functions by re-defining the distance function in (2) to recover the minimum amount of inputs required to produce a given level of output.

through a statistical analysis of a randomly selected sample of n DMUs (either with or without replacement).

2.1 General Assumptions

Our analysis proceeds under a series of general assumptions that have close precedents in the literature (Kniep, Simar and Wilson 2003; Jeong 2004 and Simar and Wilson 2005).

Assumption 1. Each DMU chooses an input vector $w \in R^p_{\geq 0}$ such that w > 0, which results in an output vector $y \in R^q_{\geq 0}$ such that y > 0.

Assumption 2. If (w, y) $\in$ T and (w^*, y^*) $\in$ T, then (αw + (1-α)w^*, αy + (1-α)y^*) $\in$ T for all $\alpha \in [0,1]$.

Assumption 3. If (w, y) $\in$ T, then (w^*, y^*) $\in$ T for $w^* \geq$ w and $y^* \leq$ y.

Assumption 1 states that each DMU in the population being analyzed is operating at some level of production, using one or more inputs to produce one or more outputs. This assumption is only slightly stronger than Kneip, Simar and Wilson (2003) who assume that the input vector is non-empty (i.e., operation requires positive input usage, even if no outputs arise from that process). Assumption 2 is fundamental to the application of DEA methodology, and implies that the production possibilities set is convex. Assumption 3 ensures that the technology exhibits free disposability of outputs. The latter two assumptions are standard in the distance function literature (for example, see Fare and Primont 1995; and Jeong 2003).

2.2 Data Envelopment Analysis Overview

Data envelopment analysis (DEA) attempts to represent the true production frontier by first identifying those DMUs in the sample that produce the most output for a given set of inputs.[6] Figure 1 presents an illustrative example of an output-oriented, two-output production process (the outputs are y_1 and y_2). Two DMUs (A and B) produce maximal output for a given amount of inputs. These DMUs are characterized as "efficient" and assigned efficiency scores equal to one. DEA then "fills in" the (sample) technological frontier by examining all possible convex combinations of these DMUs (also known as "virtual DMUs"). For each DMU operating below the frontier, DEA projects a ray from the origin, through the inefficient DMU to the constructed frontier. The proportion of the ray length that lies between the inefficient DMU and the origin is the efficiency score of that DMU. By construction, DMU C's efficiency score is given by the ratio $0C / 0C_v^*$ where C^*_v is the virtual reference point for C, even if there is a more efficient possibility, for example C^+, not in the sample.

Computationally, the DEA sample frontier and efficiency scores can be defined through a linear programming algorithm. Define W as the $p \times n$ vector of inputs and Y as the $q \times n$ vector of outputs for the entire sample of n DMUs. Then for each DMU (and assuming variable returns to scale), DEA chooses λ and ρ to solve the following linear programming problem:

[6] We are adopting the output-oriented DEA method. One can also apply DEA based on an input-oriented method whereby a given level of outputs is produced with most efficient combination of inputs (Coelli *et al*. 1998).

$$\max\ \rho\ \ s.t.\ \left\{\begin{array}{c} -\rho y + Y\lambda \geq 0 \\ w - W\lambda \geq 0 \\ 1_n{}'\lambda = 1 \\ \lambda \geq 0 \end{array}\right\} \tag{3}$$

where 1_n is an $n \times 1$ vector of 1's; λ is an $n \times 1$ vector of convexity weights; and ρ is the inverse of the sample efficiency score. In this case the convex combinations of the n observations on output and input vectors, defined respectively by Yλ and Wλ, represent the vectors of outputs and inputs for the virtual DMU(s). An efficient DMU uses the same levels of inputs to produce the same levels outputs as a virtual DMU, with the first two inequality conditions in (3) actually holding with equality. Inefficient DMUs produce lesser levels of outputs using the same or higher levels of inputs, and thus one or both of the first two inequality constraints in (3) hold as strict inequalities.

2.3 Prior Distributions on Incidence of Inefficiency

The proportional incidence of inefficiency in the population of DMUs, which we henceforth represent by π, is contained *a priori* in the interval $\pi \in [0,1]$, which represents the maximal admissible support of any prior distribution on π. Under the idealization that the population of DMUs is infinite in size, any prior distribution is admissible that has a continuous support of $[0,1]$, or any subset thereof. However, if the population is finite, then the support for π cannot possibly be the continuum $[0,1]$ or any continuous subset, since the support of π is then clearly discrete in nature. Given that there are K inefficient DMUs contained within the finite population of N DMUs, and given that the value of π is clearly equal to $\pi = \frac{K}{N}$, the support for π is precisely equal to the finite set $\pi \in \left\{ \frac{\gamma}{N}, \, for \;\; \gamma = 0,1,...,N \right\}$, or any subset thereof.[7]

For concreteness henceforth, we adopt the Continuous and Discrete Beta distributions[8] for representing prior information on π when one is sampling from infinite and finite populations of DMUs, respectively. These distributions are well known to be highly flexible and capable of representing an extremely wide range of distributional patterns over the appropriate supports for the value of π. However, the general inference methodology that we outline in this and later sections applies equally well to whatever prior distributions an analyst wishes to utilize. We state our prior distribution assumptions below.

Assumption 4. Prior information on the incidence of inefficiency, π, in a population of N DMUs is represented by some member of the Continuous or Discrete Beta family of probability distribution functions, as follows:

$$N = \infty : \;\; f(\pi \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha-1}(1-\pi)^{\beta-1} \;\; for \;\; 0 \le \pi \le 1 \tag{4}$$

$$N < \infty : \;\; f_d(\pi \mid \alpha, \beta, N) = \tau^{-1} \int_{\pi-\frac{1}{2N}}^{\pi+\frac{1}{2N}} f(\pi_d \mid \alpha, \beta)\, d\pi_d \;\; for \;\; \pi \in \Omega = \left\{ \frac{\gamma}{N}, \, for \;\; \gamma = 1,...,N-1 \right\} \tag{5}$$

[7] Sohn and Choi (2006, p. 553) explicitly assume a finite population. As such, their use of the beta-binomial is not literally defensible.

[8] The Discrete Beta distribution has not appeared frequently in the Economics literature. Additional details relating to this distribution can be found in Mazzuchi and Soyer (1996) and Juang and Anderson (2004).

where α, β > 0, $\Gamma\left(\alpha\right)=\int_{0}^{\infty} z^{\alpha-1}e^{-z}dz$ is the gamma function, and $\tau = \int_{\frac{1}{2N}}^{1-\frac{1}{2N}} f(\pi_d \mid \alpha,\beta)\, d\pi_d$.

Note we have assumed that the probability of the events that all DMUs are efficient and all DMUs are inefficient is each *a priori* zero.[9] Following Pearson (1925) and Skellam (1948), if one were to assume no prior knowledge exists regarding the appropriate value of π, i.e. the prior belief is one of "ignorance", then the uniform distribution would be appropriate, which is given in the continuous and discrete cases by (4) and (5) with α = β = 1. The beta distribution is sufficiently flexible to allow for a myriad of different distributional shapes. For example, Beta densities are skewed to the left when α > β, skewed to the right when α < β, J-shaped when ((α-1)(β-1) < 0), and U shaped when $\alpha < 1$ and $\beta < 1$.

3.0 Inferring the Incidence of Non-Latent Inefficiency

In this section we operate within an idealized setting where sample DEA estimates correctly categorize a DMU as to whether it is efficient, and so the true proportion of inefficient DMUs in the sample is observable and non-latent. This is tantamount to the presumption that DEA estimates are being calculated with reference to the true production frontier. The functional form of the posterior distribution of π will depend on whether random sampling of DMUs is with or without replacement, and

[9] The only change that would be required in the discrete case is for one or both end points to be accommodated with nonzero probability in the support of the discretized beta distribution by reassigning interval probabilities appropriately. In the continuous case all elementary events have probability zero, and a mixed continuous-discrete distribution would need to be employed if probability masses were to be assigned to one or more of the elementary events.

whether the population of DMUs is assumed to be finite or infinite. We develop these cases ahead.

3.1 Sampling with Replacement

Assume that random sampling is with replacement, in which case one can view the random sample of efficient and inefficient DMUs as iid outcomes of a Bernoulli process with the probability of an inefficient observation being equal to π. This interpretation holds whether the population of DMUs is finite or infinite. The probability that X of these DMUs are inefficient (given π and n) is then necessarily determined by the binomial probability distribution:

$$f(x \mid \pi, n) = \binom{n}{x} \pi^{x} (1-\pi)^{n-x} \tag{6}$$

Utilizing Bayesian inferential methodology, we now seek the posterior distribution of π. Factoring the joint density of $X \text{ and } \pi$ into the product of a conditional and marginal distribution yields the following representation:

$$f(x, \pi \mid \alpha, \beta, n) = f(x \mid \pi, n)\, f(\pi \mid \alpha, \beta) \tag{7}$$

At this point, derivation of the posterior distribution of π is differentiated by whether the population is infinite or finite, and thus whether priors of the form (4) or (5) are used to represent the $f(\pi \mid \alpha, \beta)$ distributional component of (7).

3.1.1 $N = \infty$

Integrating out π from the joint distribution in (7) using the prior distribution in (4) yields the marginal probability distribution of X, as

$$f(x\,|\,\alpha,\beta,n)=\binom{n}{x}\frac{\Gamma(\alpha+\beta)\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(n+\alpha+\beta)}. \qquad (8)$$

which is recognized as the *beta-binomial distribution* (e.g., Schuckers, 2003 and Phillips, 2001). Its first two moments are given by:

$$E(X)=n\frac{\alpha}{\alpha+\beta} \qquad and \qquad Var(X)=n\left(\frac{\alpha}{\alpha+\beta}\right)\left(1-\frac{\alpha}{\alpha+\beta}\right)C \qquad (9)$$

where $C=\dfrac{\alpha+\beta+n}{\alpha+\beta+1}$ .The value of C is contained in the interval $(1,n)$ and depicts the variability in the observed frequency of inefficiency that is not accounted for in the pure binomial model. If C is close to 1, then over-dispersion is not a problem, and one can use the simpler binomial distribution as the model for the data generating process and the basis for a traditional binary model. Larger values indicate the appropriateness of a mixed continuous distribution representation such as the beta-binomial.

The posterior of π given x is given by the ratio of (7) over (8),

$$f\left(\pi\,|\,x,\alpha,\beta,n\right)=\frac{f\left(x,\pi\,|\,\alpha,\beta,n\right)}{f\left(x\,|\,\alpha,\beta,n\right)}=\frac{\Gamma\left(n+\alpha+\beta\right)}{\Gamma\left(x+\alpha\right)\Gamma\left(n-x+\beta\right)}\pi^{x+\alpha-1}(1-\pi)^{n-x+\beta-1} \qquad (10)$$

which is recognized as a $Beta\left(\alpha_*,\beta_*\right)$ distribution with parameters $\alpha_*=x+\alpha$ and $\beta_*=n-x+\beta$. The posterior expected value of π and the value of π associated with the highest posterior density (HPD) weighting are two alternative Bayesian estimates of π, the former being the minimum posterior quadratic risk (MPQR) estimator and the latter being the estimate that receives the maximum posterior density weighting. Given that the posterior distribution of π is Beta, the estimator that minimizes posterior expected quadratic loss is clearly

$$E\left(\pi\right)=\frac{x+\alpha}{\alpha+\beta+n} \tag{11}$$

while the HPD estimate, which maximizes (10), is:

$$\pi_{HPD}=\frac{x+\alpha-1}{\alpha+\beta+n-2} \tag{12}$$

Credible regions for the HPD can be generated based on the tails of the posterior in (10).

If an uninformative prior is used, so that α = β = 1, the beta prior reduces to a continuous uniform density. Applying this restriction to (8) and noting that n and x are integers, the marginal distribution of X reduces to

$$f(x\,|\,1,1,n)=\begin{pmatrix} n \\ x \end{pmatrix}\frac{\Gamma(x+1)\Gamma(n-x+1)}{\Gamma(n+2)}=\begin{pmatrix} n \\ x \end{pmatrix}\frac{(x!)\left((n-x)!\right)}{(n+1)!}\,. \tag{13}$$

The posterior distribution for π can then be expressed as

$$\begin{aligned} f\left(\pi\,|\,x,1,1,n\right) &= \frac{\Gamma\left(n+2\right)}{\Gamma\left(x+1\right)\Gamma\left(n-x+1\right)}\pi^{x}(1-\pi)^{n-x}=\frac{(n+1)!}{x!(n-x)!}\pi^{x}(1-\pi)^{n-x} \\ &= \frac{\pi^{x}(1-\pi)^{n-x}}{\int\limits_{0}^{1}\pi^{x}(1-\pi)^{n-x}} \end{aligned} \tag{14}$$

from which the corresponding MPQR and HPD estimates for π can be derived yielding

$$E\left(\pi\right)=\frac{x+1}{2+n} \quad \mathit{and} \qquad \pi_{HPD}=\frac{x}{n} \tag{15}$$

3.1.2 $N < \infty$

We now consider analyzing the incidence of inefficiency within a finite population of firms when random sampling is with replacement. The probability that X of these DMUs are inefficient (given π and n), $f(x \mid n, \pi)$, continues to be the binomial probability distribution of (6). Given the finiteness of the population, we adopt the Discrete Beta distribution in (5) for the prior on π. A difficulty with (5) is that it does not lead to closed form algebraic solutions for the probabilities unless α *and* β are integer valued, in which case the integrand is a polynomial in x. In any case, the integrals in (5) are easily calculated numerically, and the numerical values of the probabilities are straightforward to define.

The general expression for the joint distribution of π *and* X can be defined via a factorization similar to (7), yielding

$$f(x, \pi \mid \alpha, \beta, n, N) = f(x \mid \pi, n) f_d(\pi \mid \alpha, \beta, N) \;\; \textit{for} \;\; \pi \in \Omega \qquad (16)$$

The posterior for π, given the data x, is then obtained by first deriving the marginal distribution of X from (16) as

$$f(x \mid \alpha, \beta, n, N) = \sum_{\pi \in \Omega} f(x \mid \pi, n) f_d(\pi \mid \alpha, \beta, N) \qquad (17)$$

and then forming the ratio of (16) and (17) to yield

$$f(\pi \mid x, \alpha, \beta, n, N) = \frac{f(x \mid \pi, n) f_d(\pi \mid \alpha, \beta, N)}{f(x \mid \alpha, \beta, n, N)} \;\; \textit{for} \;\; \pi \in \Omega \qquad (18)$$

To provide a concrete example analogous to that used in the previous subsection, we utilize the ignorance prior case of (5) where α=β=1,

$$f_d(\pi \mid 1,1,N) = \frac{1}{N-1} \; for\, \pi \in \Omega \quad . \tag{19}$$

which is the discrete uniform distribution. The joint distribution of X and π is then

$$f(x,\pi \mid 1,1,n,N) = f(X = x \mid \pi, n)\, f_d(\pi \mid 1,1,N) = \frac{1}{N-1}\binom{n}{x}(\pi)^x (1-\pi)^{n-x} \tag{20}$$

Summing over all possible values of π yields the marginal distribution of X

$$f(x \mid 1,1,n,N) = \frac{1}{N-1}\binom{n}{x}\sum_{\pi\in\Omega}(\pi)^x (1-\pi)^{n-x} \tag{21}$$

and the posterior of π given X is the ratio of (21) and (20), as

$$f(\pi \mid x,1,1,n,N) = \frac{(\pi)^x (1-\pi)^{n-x}}{\sum_{\pi\in\Omega}(\pi)^x (1-\pi)^{n-x}} \quad for\, \pi \in \Omega \tag{22}$$

which is recognized as the direct discrete analog to the continuous posterior distribution defined in (14). The MPQR estimator of π is equal to the mean of the distribution in (22), as

$$E(\pi) = \frac{\sum_{\pi\in\Omega}(\pi)^{x+1} (1-\pi)^{n-x}}{\sum_{\pi\in\Omega}(\pi)^x (1-\pi)^{n-x}} \tag{23}$$

Unconstrained maximization of (22) with respect to π yields $\pi_{\max} = \frac{x}{n}$, and it can be shown that (22) is increasing in π so long as $N\pi = K < N\left(\frac{x}{n}\right)$, and decreasing thereafter. Then defining

$$K^{-} = \left(\text{Largest Integer } \leq \tau\right) \text{ and } K^{+} = \left(\text{Smallest Integer } \geq \tau\right), \text{ where } \tau = N\left(\frac{x}{n}\right) \qquad (24)$$

the HPD estimate of π is given by

$$\pi_{HPD} = \max\left\{\frac{K^{-}}{N}, \frac{K^{+}}{N}\right\}. \qquad (25)$$

Credible regions for π can be generated in a straightforward manner using appropriate subsets of the support of (22) that have prescribed posterior probability.

Even though the preceding methodology provides a procedure for making inferences about π when sampling from finite populations with replacement most applications of DEA do not allow firms to be selected multiple times in the same sample. Thus, it is of interest to identify a parametric family of densities that characterize the number of inefficient and efficient firms when sampling without replacement from a finite population, holding other assumptions constant. The next section investigates this issue.

3.2 Random Sampling without Replacement

In analyzing the case of random sampling without replacement, we now focus exclusively on the empirically relevant case of a finite population. Given that random sampling is without replacement, the incidence of inefficiency is now characterized by the hypergeometric density. In particular, let K be the number of inefficient DMUs in the population, let x be the number of inefficient DMUs in a random sample of size n drawn from the population without replacement, and note that $\pi = \frac{K}{N}$. Then the probability of selecting x inefficient DMUs is given by:

$$f(x \mid N, \pi, n) = \left\{ \begin{matrix} \dfrac{\binom{N(1-\pi)}{n-x}\binom{N\pi}{x}}{\binom{N}{n}} & \max\left[0, n-(N-K)\right] \leq x \leq \min\left[n, K\right] \\ 0 & \textit{otherwise} \end{matrix} \right\} \tag{26}$$

While the support for π is contained in the unit interval, the support is not continuous, and we again adopt the Discrete Beta distribution (5) for the prior distribution on π. The joint distribution of X and π can be represented as

$$f(x, \pi \mid 1,1, N, n) = f(x \mid N, \pi, n) f_d(\pi \mid 1,1, N) = \frac{1}{N-1}\left(\frac{\binom{N(1-\pi)}{n-x}\binom{N\pi}{x}}{\binom{N}{n}} \right) \tag{27}$$

where for concreteness we have assumed an ignorance prior for π, and we are henceforth suppressing and leaving implicit the definition of the supports for the functional definitions of the probability distributions. Summing the joint density over π yields the marginal distribution of X. As before, this marginal distribution is a mixture (over π) of the conditional distributions of X, as

$$f(x \mid 1,1, N, n) = \left(\frac{1}{N-1}\right)\left(\frac{1}{\binom{N}{n}}\right)\sum_{\pi \in \Omega}\binom{N(1-\pi)}{n-x}\binom{N\pi}{x} \tag{28}$$

Equation (28) is a compound hypergeometric distribution.

The conditional posterior of π given X can be found by taking the ratio of (27) to (28):

$$f\left(\pi \mid x,1,1,N,n\right)=\frac{f\left(x,\pi \mid 1,1,N,n\right)}{f\left(x \mid 1,1,N,n\right)}=\frac{\binom{N\left(1-\pi\right)}{n-x}\binom{N\pi}{x}}{\sum_{\pi\in\Omega}\binom{N\left(1-\pi\right)}{n-x}\binom{N\pi}{x}} \qquad (29)$$

The expectation, and thus the MPQR estimator of π is defined by calculating

$$E\left(\pi\right)=\frac{\sum_{\pi\in\Omega}\pi\binom{N\left(1-\pi\right)}{n-x}\binom{N\pi}{x}}{\sum_{\pi\in\Omega}\binom{N\left(1-\pi\right)}{n-x}\binom{N\pi}{x}} \qquad (30)$$

The posterior density for π can be maximized to determine the HPD estimate of π. This is equivalent to maximizing the numerator of (29) as

$$\pi_{HPD}=\arg\max_{\pi\in\Omega}\left\{\binom{N\left(1-\pi\right)}{n-x}\binom{N\pi}{x}\right\} \qquad (31)$$

Some tedious but straightforward algebra demonstrates that the bracketed expression in (31) is increasing in π so long as $N\pi=K<\left[\left(N+1\right)\frac{x}{n}-1\right]$. Then defining

$K^{-}=\left(\text{Largest Integer }\leq\tau\right)$ and $K^{+}=\left(\text{Smallest Integer }\geq\tau\right)$, where $\tau=\left[\left(N+1\right)\frac{x}{n}-1\right]$ (32)

the HPD estimate of π is given by

$$\pi_{HPD}=\max\left\{\frac{K^{-}}{N},\frac{K^{+}}{N}\right\} \qquad (33)$$

Credible regions for π can be generated in a straightforward fashion using appropriate subsets of the support of (29) having prescribed posterior probability.

4.0 Inferring the Incidence of Latent Inefficiency

In empirical analyses, the production frontier is generally unknown, and using the sample-based DEA procedure to categorize DMUs as inefficient results in a potentially downward-biased count of inefficient firms in the sample which can then bias the estimate of the proportion of inefficient firms in the population.[10] However, it is possible to extend the analysis in the previous section to provide an appropriate Bayesian adjustment to the posteriors that accounts for the uncertainty and bias and leads to corrected inferences relating to the incidence of inefficiency.

We concentrate here on the case of random sampling without replacement from a finite population since it is this case that is most relevant for empirical work. We begin by illustrating the extension for a simple production technology in which outputs are produced in a single fixed proportion. We then extend the analysis to more complicated contexts, and it will be seen that the definitions of the posterior distributions remain similar to the case of the simpler technology, except that the supports of the distributions become more involved.

[10] Firms on the true efficiency frontier can never be falsely categorized as inefficient, but firms below the frontier can be falsely categorized as efficient if more efficient firms are not in the sample. Thus, the *count* of inefficient firms can only be biased downward. What this means for the estimate of the *proportion* of inefficient firms depends on how many efficient firms are missed in the sample and how many inefficient firms are falsely categorized as efficient.

4.1 Sampling without Replacement for a Simple Technology

Assume that the finite population of N DMUs produces only two outputs y_1 and y_2 in a single fixed ratio. Then retaining all other assumptions for the case of random sampling without replacement from a finite population, the population of DMUs all lie on a single ray through the origin, as illustrated in Figure 2. Conceptually, we assume that N-K of these DMUs are truly efficient, and thus lie on the single point G at the end of the ray, and the remaining K DMUs are dispersed at various points between the origin and the N-K efficient DMUs.

Now suppose that a random sample of n DMUs is drawn from the population, and DEA is used to estimate the true efficient frontier (which in this case is, of course, a single point). The fundamental problem that needs to be addressed is the “incomplete ray problem”, i.e., the true end point of the ray is unknown and not necessarily revealed when analyzing a random sample of DMUs from the population. If at least one of the N-K truly efficient DMUs producing at point G are included in the sample, then the sample DEAs correctly reflect the true efficient frontier, and all DMUs are classified correctly as to whether they are inefficient or efficient. On the other hand, if none of the N-K efficient DMUs appear in the sample, then any DMUs categorized as inefficient are, in fact, inefficient, but any DMUs categorized as efficient are incorrectly categorized. For example, if no DMUs producing at G are in the sample, but at least one DMU producing at H is in the sample, the DMUs at H will be mischaracterized as efficient while all other DMUs in the sample, at points on the ray below H, are correctly characterized as inefficient.

Let the number of truly inefficient DMUs in the sample be represented by $x_* = x + e$ where x continues to represent the number of DMUs categorized as inefficient based on the sample DEA methodology, and e represents the unobserved error in categorization. In this framework, x_* is an unobservable *latent* variable, and x is an observable sample-based estimate related to the latent variable as $x_* \geq x$. If one or more of the N-K truly efficient DMUs appear in the random sample, then $e = 0$ and $x_* = x$ whereas if the sample contains only inefficient DMUs, then the most efficient DMUs in the sample will be incorrectly categorized as efficient, and all other DMUs will be correctly categorized as inefficient so that $e = n_E$ and $x_* > x$, where n_E denotes the number of DMUs incorrectly categorized as efficient by the sample DEA analysis. In this simple technology, e is, by definition, a binary random variable that can take either the value 0 or n_E.

Our objective is to derive an appropriate posterior distribution for the unknown π that depends on only *observable* data. We begin with the joint probability distribution of the unobservable x_* and π, and for simplicity we assume at the outset that the prior on π is the uniform ignorance prior so that

$$f(x_*, \pi \mid N, n) = \frac{1}{N-1}\left(\frac{\binom{N(1-\pi)}{n - x_*}\binom{N\pi}{x_*}}{\binom{N}{n}} \right) \tag{34}$$

where we have suppressed the notation indicating $\alpha = \beta = 1$ in the discrete Beta density. This joint density is, of course, simply a copy of (27) except that it is now the unobservable x_*, rather than the observable x, that represents the number of truly inefficient DMUs in the sample.

Now suppose that we observe $\{x, n_E\}$ as the outcome of the sample DEA analysis, so there are x DMUs categorized as inefficient, and there are n_E DMUs located at the end point of the sample ray and thus categorized as efficient by the DEA analysis. Note that conditional on observing $\{x, n_E\}$, it must necessarily be the case that either $x_* = x$ or $x_* = x + n_E$. Then the joint probability distribution of $\{x_*, \pi\}$, *conditional* on $\{x, n_E\}$, can be defined as

$$f(x_*, \pi | N, n, x, n_E) = \frac{f(x_*, \pi | N, n)}{\sum\limits_{x_* \in \Psi} \sum\limits_{\pi \in \Omega} f(x_*, \pi | N, n)} = \frac{\binom{N(1-\pi)}{n - x_*}\binom{N\pi}{x_*}}{\sum\limits_{x_* \in \Psi} \sum\limits_{\pi \in \Omega} \binom{N(1-\pi)}{n - x_*}\binom{N\pi}{x_*}} \; for \{x_*, \pi\} \in \Psi \times \Omega \quad (35)$$

where $\Psi = \{x, x + n_E\}$ and Ω is as defined in (5).

The posterior distribution of π, conditioning on only observables by marginalizing out the unobservable latent variable x_* from (34), is given by

$$f(\pi | N, n, x, n_E) = \sum_{x_* \in \Psi} f(x_*, \pi | N, n, x, n_E) = \frac{\sum\limits_{x_* \in \Psi} \binom{N(1-\pi)}{n - x_*}\binom{N\pi}{x_*}}{\sum\limits_{x_* \in \Psi} \sum\limits_{\pi \in \Omega} \binom{N(1-\pi)}{n - x_*}\binom{N\pi}{x_*}} \quad (36)$$

This posterior could then be used to derive the MPQR estimate, $E(\pi)$, or the HPD estimate, $\arg\max\limits_{\pi} \{ f(\pi | N, n, x, n_E)\}$ of the true proportion of inefficient DMUs in the population, π, as well as generate credible regions for the value of π.

4.2 Sampling without Replacement for More General Technologies

Allowing for more general technologies complicates the relationship between x and x_*. In this case the DMUs can lie on a number of rays, and the number of ways in which x can misrepresent x_* expands concomitantly. We first extend the results to the two ray case, which is illustrative of some of the additional complications that arise in the multiple ray case. We then proceed to generalize the results to an arbitrary finite number of rays.

4.2.1 The Two Ray Case

We now assume that the outputs can be produced in two different fixed ratios, as was illustrated in Figure 1, where for now we ignore ray $0C^+$. We do not assume that additional information is available beyond what has already been assumed, e.g., we do not assume that it is known how many DMUs in the population reside on each ray.[11] All of the N DMUs lie on one of two rays 0A or 0B through the origin. There are conceptually N-K of these DMUs that are truly efficient, and these DMUs lie on one of the two ray endpoints, and the remaining K DMUs are dispersed at various points, for example D, E and F, along the two rays between the origin and the N-K efficient DMUs.

The sample outcome of DEA can be such that one ray is dominated by the other in the sample representation of the production frontier, or else both rays contribute to the

[11] A refinement in the analysis could be pursued if such information were in fact available whereby the proportions of inefficient DMUs residing on each ray, say π_i, could be considered in the analysis.

definition of the frontier boundary.[12] If one of the sample rays is dominated, then the single ray analysis in the previous subsection applies based on the dominating ray, where all of the DMUs sampled from the dominated ray would be included in the total number, *x*, of DMUs designated as inefficient by the sample DEA. Letting n_E denote the number of DMUs categorized as efficient by DEA, and thus residing at the endpoint of the dominating ray, the actual number of inefficient DMUs is either $x_* = x$ or $x_* = x + n_E$, as in the single ray case.

If neither ray is dominated, and if one or more truly efficient DMUs are at the endpoints of each of the two rays, then all of the sample DMUs are categorized correctly by DEA regarding whether they are inefficient or not, and thus $e = 0$ and $x_* = x$. If the sample contains only truly inefficient DMUs, then the most efficient of these will be incorrectly categorized by DEA as efficient, and all other DMUs will be correctly categorized as inefficient so that $e = n_E$ and $x_* = x + n_E$.

There are two additional possibilities regarding errors in categorizing truly inefficient firms as efficient. In order to delineate these events, let $n_E^A \; and \; n_E^B$ be the number of sampled DMUs residing on the endpoints of rays 0A and 0B, respectively, where $n_E = n_E^A + n_E^B$. It could be the case that the n_E^A DMUs at the end of ray 0A are classified incorrectly as efficient while the n_E^B DMUs at the end of ray 0B are correctly

[12] For example, if the sample include DMUs at B, E and F but not DMUs at D or A only DMUs at B would be deemed efficient and only ray 0B contributes to the definition of the efficient frontier. If the sample also included DMUs at D, both rays 0A and 0B would contribute to the definition of the efficient frontier even though sample ray 0D would be an incomplete ray.

classified as efficient, or vice versa. The set of possible events regarding the relationship between x and x_* is then

$$x_* = x + e,\ \ e \in Unique\left\{0, n_E^A, n_E^B, n_E\right\} \tag{37}$$

where $Unique\left\{\bullet\right\}$ is the uniqueness operator returning only the unique items within any listing $\left\{\bullet\right\}$. Note that if $n_E^A = n_E^B$, then there are only three distinct error values that are possible, and one of the redundant values will be removed from the list of possibilities by this operator.

Continuing to assume an ignorance prior, the joint distribution of $\left\{x_*, \pi\right\}$ is given by (34) as before. Once the outcome $\left\{x, n_E^A, n_E^B\right\}$ of the DEA analysis is observed one can proceed to define the joint density of $\left\{x_*, \pi\right\}$, conditional on $\left\{x, n_E^A, n_E^B\right\}$, i.e. $f\,(x_*, \pi \middle| N, n, x, n_E^A, n_E^B)$, by the right hand sides of (35), except the support of that density in the x_* dimension is now given by $\Psi = Unique\left\{x, x + n_E^A, x + n_E^B, x + n_E\right\}$. Finally, the posterior distribution of π, conditioned entirely on observable data, can be defined as in (36) but again using the immediately preceding definition of the support Ψ. Estimators and credible regions could be defined accordingly.

4.2.2 The Multiple Ray Case

We now consider the case where the technology is characterized by $m > 2$ rays, and we label them sequentially as $\left\{1, 2, ..., m\right\}$. We continue to make the same assumptions as previously, and in particular do not assume that any additional information is available about the problem other than the fact that DMUs can potentially reside on any of m rays. Of the population of N DMUs, N-K of these DMUs are truly efficient, and thus lie on one of the endpoints of the m rays, and the remaining K DMUs are dispersed at various points along the m rays, and between the origin and the N-K efficient DMUs.

The joint distribution of $\{x_*, \pi\}$, continuing to assume an ignorance prior, is again given by (34) as before. Once the sample outcome of the DEA analysis is observed, it is revealed which of the m rays have observations that place them on the sample production frontier. Let $I_E \subset \{1, 2, ..., m\}$ be the index set identifying the rays whose endpoints lie on that frontier, yielding the observations $\left\{x \text{ and } n_E^i,\ i \in I_E\right\}$, where n_E^i denotes the number of DMUs on ray i that are designated as efficient by the sample DEA. One can then proceed to define the joint density of $\{x_*, \pi\}$, conditional on $\left\{x \text{ and } n_E^i, i \in I_E\right\}$, i.e., $f(x_*, \pi | N, n, x, \text{ and } n_E^i, i \in I_E)$, by the right hand sides of (35), except that now the support of that density in the x_* dimension is given by

$$\Psi = Unique\left\{x + \sum_{i \in J} n_E^i\,,\ \ for\ \ J \subset I_E\right\} \tag{38}$$

Note that Ψ contains *x* as before, since one of the index subsets $J \subset I_E$ is $J = \varnothing$. Finally, the posterior distribution of π, conditioned entirely on observable data, would be defined by (36) using the preceding definition of the support Ψ. Estimators and credible regions could be defined accordingly.

5. An Illustrative Numerical Example

As an illustration of how the preceding methodology can be applied, assume there are $N = 150$ DMUs in a population under study, the DMUs produce two outputs, and the technology for producing those outputs is categorized by 8 distinct rays. Characteristics of the population, including the number of efficient and inefficient firms along each ray, are displayed in Table I.

Table I. Illustrative Population of N = 150 DMUs

Ray	1	2	3	4	5	6	7	8	Total
# DMUs	20	15	19	16	25	12	22	21	150
# inefficient	15	13	17	14	21	10	17	15	122
Proportion Inefficient	.75	.867	.895	.875	.840	.833	.773	.714	.813

.

Random samples, without replacement, of $n = 25 \textit{ and } 50$ DMUs were extracted (via simulation) from the population described in Table I, resulting in the DEA categorizations given in Table II. When n=25 five of the DMUs (from rays 2, 4, 5, 6 and 7) measured as efficient were actually inefficient. In the sample of 50, only 2 DMUs (from rays 6 and 7) were mischaracterized as efficient.

Table II: Simulated Outcomes of DEA Analysis

	n=25		n=50	
Ray	Efficient	Inefficient	Efficient	Inefficient
1	1	4	2	5
2	1	4	1	8
3	2	3	5	5
4	1	0	1	5
5	1	0	1	3
6	1	2	1	2
7	1	2	1	6
8	2	0	2	2
TOTAL	10	15	14	36

Now consider inferring the incidence of inefficiency from the observed DEA sample outcomes. Simply using the unadjusted sample means from the DEA analysis would produce the estimates $\hat{\pi} = \frac{15}{25} = .6$ and $\hat{\pi} = \frac{36}{50} = .72$ for the samples of 25 and 50 DMUs, respectively, which represent sample DEA estimates for the true underlying proportion of inefficiency, .813.

For the sake of illustration, we adopt a uniform ignorance prior on the incidence of inefficiency in the population, and we consider the various Bayesian approaches for inferring π. Assuming (incorrectly) that the population is infinite and that sampling was with replacement, and treating the incidence of inefficiency as non-latent, the estimates in (15) would be employed, whereby the MPQR estimates would be $\hat{\pi} = \frac{16}{27} = .593$ and $\hat{\pi} = \frac{37}{52} = .712$, and the HPD estimates would be identical in numerical value to the sample mean DEA estimates of .6 and .72 for sample sizes 25 and 50, respectively. If one instead proceeds in the context of the population being finite, but continues to mistakenly assume that sampling was done with replacement, and also assumes that the incidence of inefficiency is non-latent, the MPQR and HPD estimates would be given by (23) and (25), respectively. The symmetry of the posterior distributions (14) and (22) under the current problem conditions leads to MPQR and HPD estimates that are identical to those above, these being .593 and .712, and .60 and .72, respectively for samples of sizes 25 and 50.

Next consider proceeding in the context of a finite population, and random sampling without replacement, while continuing to assume that the incidence of inefficiency is non-latent. The calculation of the MPQR and HPD estimates would then be based on (30) and (31). The MPQR estimates are .594 and .718 for n = 25 and 50, respectively. The HPD estimates are given by .60 and .72.

Finally, consider the most appropriate approach in which the correct assumptions are made that the population is finite, random sampling is without replacement, and observations on inefficiency are *latent*. We continue to assume an ignorance prior on π. In this case, the posterior distribution of π defined in (36) is appropriate, and the MPQR and HPD estimates of π are defined based on that posterior distribution. The MPQR estimates for sample sizes n = 25 and 50, i.e., the expected values of the posterior in (36), are given by .778 and .847, respectively, which are both closer to the true value of .813 than any of the other estimates examined heretofore. The HPD estimates that maximize (36) are given by .933 and .906 for sample sizes n = 25 and n = 50, respectively.

6. Implications for the DEA Literature

The results of this analysis have important implications for the DEA literature that can be separated into two main categories – those dealing with the statistical foundation of DEA for estimating the incidence of inefficiency among DMUs and those dealing with the empirical implementation of DEA for measuring inefficiency among DMUs.

6.1 Implications for the Statistical Foundation of DEA

Regarding the statistical foundation of DEA, first note that the functional definition of the posterior distribution of the proportion of inefficient DMUs need not be assumed, ad hoc, as is done in Sohn and Choi (2006). Instead, Bayesian methods provide an internally consistent and defensible statistical specification. A related contribution is the delineation of specific problem conditions under which a beta-binomial distributional assumption, which has recently been employed in the DEA literature, is appropriately implemented in the analysis. Unfortunately, under most empirical circumstances, the conditions would seem to preclude the use of the distribution, despite its flexibility. The assumption that the population is infinite is clearly false in any empirical application and given that most empirical studies do not allow the same DMU (in a given time period) to appear multiple times in a data set and sample randomly *without* replacement[13] the observations cannot be viewed as iid outcomes of a Bernoulli process. Thus it is literally inappropriate to assign binomial probabilities to the number of inefficient firms observed. This, in turn, contradicts the assumptions of the beta-binomial, since it is, by definition, a mixture of a beta and a binomial density.

Perhaps the most important contribution with respect to providing a statistical foundation for the incidence of inefficiency in DEA estimates is that we provide an

[13] Some researchers argue that binomial probabilities can be used to approximate random sampling without replacement (whose probabilities are given by the hypergeometric distribution) as long as N is infinite or as long as n <<N. While the former is true in theory, in reality an instance where N is large enough to be considered practically infinite is quite rare. Additionally, the latter assumption is problematic because it prevents the researcher from using the usual limiting arguments to derive the consistency of one’s estimates. In particular, for a sample’s estimates to be consistent, one must demonstrate that the sample estimate converges to the population parameter in probability as n approaches N. However, as n approaches N the binomial distribution is a very poor approximation of the hypergeometric.

appropriate characterization of the nature of the posterior distribution of inefficiency when sampling is without replacement from a finite population, which is in fact the most relevant case for empirical work. Unlike previous studies, such as Gijbels *et al*. (1999), Jeong (2004) and Simar and Wilson (2005), the results are obtained without making specific assumptions about the smoothness of the production function. More importantly, we are able to account for the mismeasurement (DEA bias which results from the latency of the incidence) in the incidence of inefficiency that is likely to occur when DEA is applied to small samples. The method becomes computationally more involved as one allows for more general production technologies and/or as one allows the support for the error in the relationship between the latent number of truly inefficient firms and the number indicated by DEA to expand. However, any such computation remains relatively straightforward on a computer.

We note that while the approach makes probabilistically coherent use of available sample and prior information when making inferences about the true proportion of inefficient firms, the distribution of the DEA scores themselves has not been defined. Our results therefore do not directly address the critique of Simar and Wilson (2005) who argue that the majority of the two-stage DEA literature, in which DEA scores are regressed on explanatory factors, has mischaracterized the distribution of DEA scores, and thus generated biased and inconsistent parameter estimates of the effects of explanatory factors. At the same time, we do provide a proper statistical basis for tests relating to the proportion of efficient and inefficient DMUs based on posterior odds ratios.

6.2 Implications for the Implementation of DEA

Our analysis of the small sample efficiency mismeasurement problem of DEA provides an additional perspective on the potential for DEA to eventually (in the limit, as sample size increases) generate fully accurate estimates of the incidence of inefficiency, as well as fully accurate efficiency scores. In particular, our conceptualization suggests that even if the researcher has a *census* of the population of DMUs, DEA bias may still persist, which is a perspective that appears to have been underappreciated in the DEA literature. Consider Figure 1 once again, this time including the full ray $0C^+$, so the production context is one of a three-ray technology. Assume that DMU A is the truly efficient firm on the first ray, DMU B is the truly efficient firm on the second ray, while C^+ is the theoretically efficient outcome for a DMU operating on the middle ray, but we assume that no firm in the population of DMUs has actually achieved that level of efficiency.

First recall that even if all potential production technologies have DMUs in the population that achieve true efficiency, some DEA scores will always be in error (regarding both incidence of inefficiency, and efficiency score) unless the researcher has a large enough sample to ensure that all of the endpoints for the rays that compose the true efficiency frontier are represented in the sample. For example, if a sample of n DMUs does not include point C^+, but instead indicates that the end of the ray is at some point below C^+, then the estimate for the number of inefficient DMUs (as well as the magnitude of the efficiency scores for firms operating along the ray) generated by DEA will be in error. This result is a primary reason why the literature suggests only applying

DEA to large samples. However, the ability to collect and analyze even a *census* of data may not be sufficient to eliminate the error. If no DEA in the population operates at C^+ (i.e., no firm along this ray is truly efficient) all DEA scores along this ray will be inflated because DEA will either compare them to the DMU furthest from the origin (if it lies between C^*_v and C^+), or construct a linear combination of DMU A and DMU B to identify a "virtual" efficient point for comparison. Thus, either all DMUs along this ray will be categorized as inefficient relative to C^*_v, or at least one (between C^*_v and C^+) will be falsely categorized as efficient, in both cases causing errors in both the number of inefficient DMUs in the sample, and the efficiency *scores*. It follows that in such cases there can be an upper bound to the accuracy of the results (whether incidence or scores) generated by DEA, regardless of sample size.[14]

Summarizing conditions under which DEA produces correct *sample* proportions of inefficiency and correct efficiency scores, all sampled DMUs must lie on rays that have *true* production frontier endpoints defined by DMUs in the sample. In order for increasing sample size $n \to N$ to be sufficient for DEA to generate, in the limit, the correct (true) proportion of inefficient DMUs in the population, as well as to generate accurate efficiency scores, the previous condition must hold *and* there must be a DMU in the population that resides on the true efficient point of every ray in the production possibilities set.

[14] The magnitude of this error is increased if both of these scenarios occur; that is, when the production frontier has multiple rays, and where one ray's true endpoint is excluded from the sample (but not the population), and the other ray's endpoint is not identified in the population or the sample.

7. Conclusions

In this paper, we used Bayesian methods to derive posterior distributions for the incidence of inefficient DMUs based on information calculated by DEA when sampling from either an infinite or a finite population. Our findings suggest that the functional forms of these distributions are quite sensitive to the sampling procedure (with or without replacement) employed. Moreover, it is through interpreting the true incidence of inefficiency as a latent variable that the noise and potential mismeasurement of efficiency inherent in the DEA scores can be effectively integrated, and properly accounted for, in the definition of posterior distributions.

Our analysis extrapolates to the implication that not only is it generally incorrect to use DEA estimates of inefficiency as dependent variables in regression analyses, but also that bootstrapping or nonparametric methods do not necessarily mitigate DEA mismeasurement problems. The failure derives from viewing the actual incidence of inefficiency as a latent variable susceptible to mismeasurement by DEA, and noting that DEA calculations may be wrong *even if one applies DEA to the entire population of DMUs*. Moreover, our analysis provides some initial insights into how to characterize and adjust DEA-based calculations for this error, and offers an approach that can improve the accuracy of the estimate of incidence. The distribution for the proportion of truly inefficient DMUs based on DEA information can provide the researcher with additional information that can be incorporated into a subsequent analysis of institutional and market factors that might influence the incidence of a firm being efficient or not. The authors are currently researching this extension of the current work.

References

Andersen, P., and N. Petersen, “A Procedure for Ranking Efficient Units in Data Envelopment Analysis,” *Management Science* 39 (1993), 1261-1264.

Banker, R., “Maximum Likelihood, Consistency and Data Envelopment Analysis: A Statistical Foundation,” *Management Science* 39 (1993), 1265-1273.

Charnes, A., W. Cooper, and E. Rhodes, “Measuring the Efficiency of Decision Making Units,” *European Journal of Operational Research* 2 (1978), 429-444.

Chilingerian, J., “Evaluating Physician Efficiency in Hospitals: A Multivariate Analysis of Best Practices,” *European Journal of Operational Research* 80 (1995), 548-574.

Chirkos, T., and A. Sears, “Technical Efficiency and the Competitive Behavior of Hospitals,” *Socio-Economic Planning Sciences* 28 (1994), 219-227.

Coelli,T., D.S.P. Rao, and G.E. Battese, *An Introduction to Efficiency and Productivity Analysis* (Boston, MA; Kluwer Academic Publishers, 1998).

Debreu, G., “The Coefficient of Resource Utilization,” *Econometrica* 19: (1951), 273-292.

Fare, R., and D. Primont, *Multi-Output Production and Duality: Theory and Applications* (Boston, MA: Kluwer Academic Publishers, 1995).

Farrell, M., “The Measurement of Productive Efficiency,” *Journal of the Royal Statistical Society, Series A* 120 (1957), 253-281.

Gijbels, I., E. Mammen, B. Park and L. Simar, “On Estimation of Monotone and Concave Frontier Functions,” *Journal of the American Statistical Association* 94 (1999), 220-228.

Hirschberg, J., and P. Lloyd, “Does the Technology of Foreign-Invested Enterprises Spill over to Other Enterprises in China?” in *Modeling the Chinese Economy*, ed. by P. Lloyd and X. Zang, (London: Edward Elgar Press, 2002).

Hogg, R., and A. Craig, *Introduction to Mathematical Statistics* (Englewood Cliffs, NJ: Prentice Hall, 1995).

Jeong, S., “Asymptotic Distribution of DEA Efficiency Scores,” Institut de Statistique, Universite Catholique de Louvain Working Paper No. 0425, 2004.

Jeong, S., and B. Park, “Limit Distribution of Convex-Hull Estimators of Boundaries,” Institut de Statistique, Universite Catholique de Louvain Working Paper No. 0424, 2004.

Juang, M., and G. Anderson, “A Bayesian Method on Adaptive Preventive Maintenance Problem,” *European Journal of Operational Research* 155 (2004), 455-473.

Kneip, A., B. Park and L. Simar, “A Note on the Convergence of Nonparametric DEA Estimators for Production Efficiency Scores,” *Econometric Theory* 14 (1998), 783-793.

Kneip, A., L. Simar and P. Wilson, “Asymptotics for DEA Estimators in Non-Parametric Frontier Models,” Institut de Statistique, Universite Catholique de Louvain Working Paper No. 0317, 2003.

Lothgren, M., and M. Tambour, “Testing Scale Efficiencies in DEA Models: A Bootstrapping Approach,” *Applied Economics* 31:10 (1999), 1231-1237.

Lovell, C., and A. Rouse, “Equivalent Standard DEA Models to Provide Super-Efficiency Scores,” *Journal of the Operational Research Society* 54 (2003), 101-108.

Mazzuchi, T., and R. Soyer, “A Bayesian Perspective on Some Replacement Strategies,” *Reliability Engineering and System Safety* 51 (1996), 295-303.

Pearson, E., “Bayes’ Theorem, Examined in the Light of Experimental Sampling,” *Biometrika* 17 (1925), 388-442.

Phillips, K., “Testing Microbiologic Response to Anti-Infective Medications with Incomplete Data,” *Journal of Biopharmaceutical Statistics* 11:4 (2001), 237-252.

Ray, S., “Resource-Use Efficiency in Public Schools: A Study of Connecticut Data,” *Management Science* 37 (1991), 1620-1628.

Rosenman, R., and D. Friesner, “Scope and Scale Inefficiencies in Physician Practices,” *Health Economics* 13 (2004), 1091-1116.

Schmidt, P., “Frontier Production Functions,” *Econometric Reviews* 4:2 (1985).

Schuckers, M., “Using the Beta-Binomial Distribution to Assess Performance of a

Biometric Identification Device," *International Journal of Image and Graphics* 3:3 (2003), 523-529.

Sengupta, J., "The Efficiency Distribution in a Production Cost Model," *Applied Economics* 30 (1998), 125-132.

Simar, L., and P. Wilson, "Some Problems with the Ferrier/Hirschberg Bootstrap Idea," *Journal of Productivity Analysis* 11 (1999a), 67-80.

_______, "Of Course We Can Bootstrap DEA Scores! But Does It Mean Anything? Logic Trumps Wishful Thinking," *Journal of Productivity Analysis* 11 (1999b), 93-97.

_______,"Estimation and Inference in Two-Stage, Semiparametric Models of Production Processes," *Journal of Econometrics,* (2006), forthcoming.

Skellam, J., "A Probability Distribution Derived from the Binomial Distribution by Regarding the Probability of Success as a Variable between the Sets of Trials," *Journal of the Royal Statistical Society Series B* 10 (1948), 257-261.

Sohn, S., and H. Choi, "Random Effects Logistic Regression Model for Data Envelopment Analysis with Correlated Decision Making Units," *Journal of the Operational Research Society* 57 (2006), 552-560.

Stanton, K., "Trends in Relationship Lending and Factors Affecting Relationship Lending Efficiency," *Journal of Banking and Finance* 26 (2002), 127-152.

Steck, G., and W. Zimmer, "The Relationship between Neyman and Bayes Confidence Intervals for the Hypergeometric Parameter," *Technometrics* 10 (1968), 199-203.

Xue, M., and P. Harker, "Overcoming the Inherent Dependency of DEA Efficiency Scores: A Bootstrap Approach," Wharton Financial Institutions Center Working Paper, 2002.

Wilson, P., "FEAR 1.0: A Software Package for Frontier Efficiency Analysis with R," unpublished manuscript, Department of Economics, University of Texas, 2005.

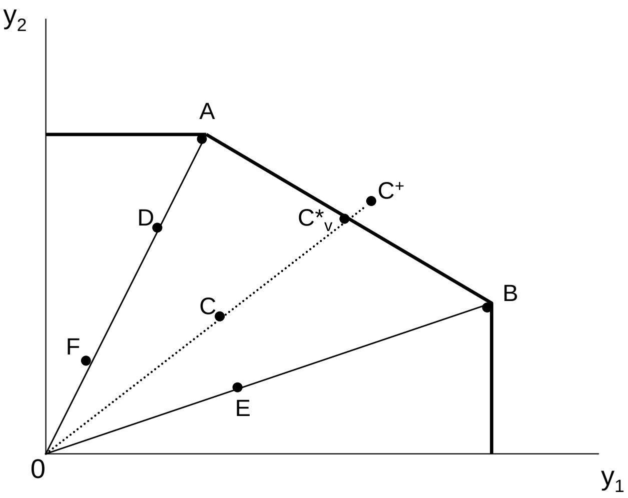

Figure 1: An Illustration of DEA

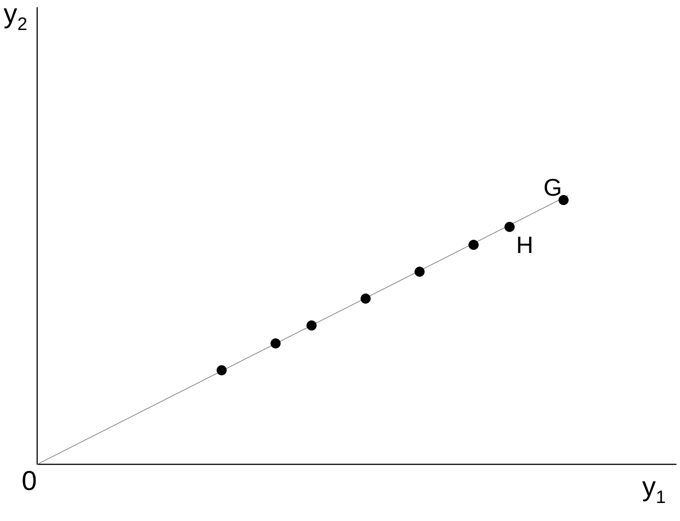

Figure 2: DEA with a Single Ray Production Technology