Nonparametric Tests of Returns to Scale

Léopold Simar*

and

Paul W. Wilson**

June, 1998

*Institut de Statistique and CORE, Université Catholique de Louvain, Voie du Roman Pays 20, Louvain-la-Neuve, Belgium.

Research support from the contract "Projet d'Actions de Recherche Concertées" (PARC No. 93/98–164) of the Belgian Government is gratefully acknowledged.

**Department of Economics, University of Texas, Austin, Texas 78712, USA.

Research support from the Management Science Group, US Department of Veterans Affairs, is gratefully acknowledged.

ABSTRACT

This paper discusses various statistics for testing hypotheses regarding returns to scale in the context of nonparametric models of technical efficiency. In addition, the paper presents bootstrap estimation procedures which yield appropriate critical values for the test statistics. Evidence on the true sizes and power of the various proposed tests is obtained from Monte Carlo experiments. This paper is an extension of earlier work in Simar and Wilson (1998a).

Keywords: Returns to scale, data envelopment analysis, bootstrap, distance function, efficiency, frontier models.

1. INTRODUCTION

Nonparametric methods have been widely used in management science and economics to estimate the efficiency of production units.[1] These methods are based on definitions of technical and allocative efficiency by Debreu (1951) and Farrell (1957). A variety of nonparametric approaches for efficiency estimation have been proposed; *e.g.,* Charnes *et al.* (1978, 1979), Färe and Lovell (1978), Burley (1980), Zieschang (1984), Deprins *et al.* (1984), and Färe *et al.* (1985). The approaches which rely on convexity assumptions have been termed Data Envelopment Analysis (DEA), and typically rely on linear programming (LP) techniques for solution.

Whether the underlying technology exhibits increasing, constant, or decreasing returns to scale is a crucial question in any study of productive efficiency. Färe *et al.* (1985) suggested an approach for determining local returns to scale in the estimated frontier which involves comparing different DEA efficiency estimates obtained under the alternative assumptions of constant, variable, or nonincreasing returns to scale, but did not provide a formal statistical test of returns to scale. Banker (1996) proposed two rather *ad hoc,* semi-parametric statistical tests, although we see no particular reason why one should expect these test to have the correct size. In addition, Banker proposed using a Kolmogorov-Smirnov test to examine whether DEA efficiency estimates obtained under alternative assumptions regarding returns to scale have different distributions.

We propose a bootstrap procedure for testing hypotheses regarding returns to scale. Our procedure avoids the *ad hoc* assumptions of Banker (1996). In addition, we provide Monte-Carlo evidence indicating that our procedure yields tests whose sizes are much closer to the nominal size than Banker's tests. The development of a reliable test procedure for examining returns to scale is important for both economic and statistical reasons. The question of whether a technology exhibits constant returns to scale everywhere typically

[1] Seiford (1997) provides an extensive bibliography of this literature; see also the survey by Lovell (1993).

has important economic implications—if the technology does not exhibit constant returns to scale, then some production units may be found to be either too large or too small. Some authors have *a priori* imposed the rather restrictive assumption of constant returns to scale while using DEA methods; this may seriously distort measures of efficiency if the true technology displays nonconstant returns to scale. Indeed, as we discuss below, this scenario results in statistically inconsistent estimates of efficiency. Alternatively, if one assumes variable returns to scale when returns are actually constant everywhere, there may be a loss of statistical efficiency. Using the methods we propose below, researchers can first test whether returns to scale are constant, and then choose the appropriate estimator to measure efficiency.

DEA methods have frequently been characterized as *deterministic* in the literature (*e.g.,* Lovell, 1993), as if to suggest that nonparametric models of efficiency have no statistical underpinnings. Indeed, although there have been numerous applications of nonparametric efficiency models, they have invariably presented point estimates of inefficiency, with no measure and only slight or no discussion of uncertainty surrounding these estimates. Yet in the nonparametric approaches to efficiency estimation, efficiency is measured relative to an *estimate* of the boundary of the underlying *true* production set, conditional on observed data resulting from an underlying (and unobserved) data-generating process (DGP). Banker (1993) was one of the first to discuss the statistical nature of DEA estimators. Korostelev *et al.* (1995), Kneip *et al.* (1996), and Gijbels *et al.* (1996) examined the statistical properties of DEA estimators. Curiously, papers continue to appear in respectable journals which ignore the statistical nature of DEA problems. Simar and Wilson (1998a) provide a bootstrap methodology for estimating confidence intervals for individual production units' efficiency; this methodology is adapted here to the problem of testing hypotheses regarding returns to scale.

We proceed in the next section by discussing methods for measuring returns to scale in the true technology using distance functions. The third section discusses estimation of these distance functions. In section 4 we propose several statistics for testing hypotheses regarding returns to scale of the technology, and in section 5 we show how bootstrap methods can be used to estimate critical values for our statistics. Section 6 discusses Monte Carlo evidence on the true (versus nominal) sizes of our tests. Conclusions are given in the final section.

2. MEASURING RETURNS TO SCALE

To illustrate nonparametric estimation of technical efficiency in production, let $\boldsymbol{x} \in \mathbb{R}_+^p$ denote a vector of p inputs and $\boldsymbol{y} \in \mathbb{R}_+^q$ denote a vector of q outputs. Define the production set

$$\mathcal{P} \equiv \{(\boldsymbol{x}, \boldsymbol{y}) | \boldsymbol{x} \text{ can produce } \boldsymbol{y}\}, \tag{2.1}$$

which is merely the set of feasible combination of $\boldsymbol{x}$ and $\boldsymbol{y}$. The production set $\mathcal{P}$ is sometimes described in terms of its sections

$$\mathcal{Y}(\boldsymbol{x}) \equiv \{\boldsymbol{y} | (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}\} \tag{2.2}$$

and

$$\mathcal{X}(\boldsymbol{y}) \equiv \{\boldsymbol{x} | (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}\}, \tag{2.3}$$

which form the output feasibility and input requirement sets, respectively.

The boundary of $\mathcal{P}$ is sometimes referred to as the *technology* or, in the DEA literature, the *production frontier,* and is given by the intersection of $\mathcal{P}$ and the closure of its compliment, or

$$\mathcal{P}^B = \{(\boldsymbol{x}, \boldsymbol{y}) \mid (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P},\ (\boldsymbol{x}/\lambda, \boldsymbol{y}) \notin \mathcal{P}\ \forall\ \lambda > 1,\ (\boldsymbol{x}, \boldsymbol{y}/\theta) \notin \mathcal{P}\ \forall\ \theta < 1\}. \tag{2.4}$$

Knowledge of either $\mathcal{Y}(\boldsymbol{x})$ for all $\boldsymbol{x}$ or $\mathcal{X}(\boldsymbol{y})$ for all $\boldsymbol{y}$ is equivalent to knowledge of $\mathcal{P}$; $\mathcal{P}$ implies (and is implied by) both $\mathcal{Y}(\boldsymbol{x})$ and $\mathcal{X}(\boldsymbol{y})$. Thus, both $\mathcal{Y}(\boldsymbol{x})$ and $\mathcal{X}(\boldsymbol{y})$ inherit the properties of $\mathcal{P}$. Various assumptions regarding properties of the production set $\mathcal{P}$ are possible; we adopt those of Shephard (1970) and Färe (1988), although our bootstrap methodology does not depend on these assumptions. Typical assumptions, which we adopt, are: (i) $\mathcal{Y}(\boldsymbol{x})$ is convex, bounded, and closed for all $\boldsymbol{x} \in \mathbb{R}^p_+$, and $\mathcal{X}(\boldsymbol{y})$ is convex, bounded, and closed for all $\boldsymbol{y} \in \mathbb{R}^q_+$; (ii) $(\boldsymbol{x}, \boldsymbol{y}) \notin \mathcal{P}$ if $\boldsymbol{y} > 0$, $\boldsymbol{x} = 0$, *i.e.,* all production requires use of some inputs; and (iii) for $\widetilde{\boldsymbol{x}} \geq \boldsymbol{x}$, $\widetilde{\boldsymbol{y}} \leq \boldsymbol{y}$, if $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}$ then $(\widetilde{\boldsymbol{x}}, \boldsymbol{y}) \in \mathcal{P}$ and $(\boldsymbol{x}, \widetilde{\boldsymbol{y}}) \in \mathcal{P}$, *i.e.,* both inputs and outputs are strongly disposable.

Now define the set $\mathcal{V}$ as the convex cone (with vertex at the origin) spanned by $\mathcal{P}$; then $\mathcal{P} \subseteq \mathcal{V}$. If $\mathcal{P}^B$ exhibits constant returns to scale (CRS) everywhere, then the technology $\mathcal{P}^B$ implies a mapping $\boldsymbol{x} \to \boldsymbol{y}$ that is homogeneous of degree 1; *i.e.,* $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}^B$ implies $(\lambda\boldsymbol{x}, \lambda\boldsymbol{y}) \in \mathcal{P}^B$ for all $\lambda > 0$. In this case, $\mathcal{P} = \mathcal{V}$.

If $\mathcal{P}^B$ does not exhibit CRS everywhere, then $\mathcal{P} \subset \mathcal{V}$. In this case, $\mathcal{P}^B \bigcap \mathcal{V}$ gives the region of $\mathcal{P}^B$ that exhibits CRS in the sense that $\mathcal{X}(\theta\boldsymbol{y}) = \theta\mathcal{X}(\boldsymbol{y})$ for some (but not all) $\theta > 0$. The intersection $\mathcal{P}^B \bigcap \mathcal{V}$ may be as small as a single point, in which case $\mathcal{X}(\theta\boldsymbol{y}) = \theta\mathcal{X}(\boldsymbol{y})$ if and only if $\theta = 1$, but by construction the intersection cannot be the null set. Decreasing returns to scale in the neighborhood of a point $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}^B$ implies $(\lambda\boldsymbol{x}, \lambda\boldsymbol{y}) \notin \mathcal{P}$ for $\lambda > 1$. Similarly, increasing returns to scale in the neighborhood of a point $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}^B$ implies $(\lambda\boldsymbol{x}, \lambda\boldsymbol{y}) \notin \mathcal{P}$ for $\lambda < 1$. A technology $\mathcal{P}^B$ that exhibits increasing, constant, and decreasing returns to scale in different regions is one of variable returns to scale (VRS) in the sense of Afriat (1972).[2]

The Shephard (1970) output distance function provides a normalized measure of Euclidean distance from a point $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}$ to $\mathcal{P}^B$ in a direction orthogonal to $\boldsymbol{x}$, and may be defined as

$$D(\boldsymbol{x}, \boldsymbol{y}) \equiv \inf\{\theta > 0 | (\boldsymbol{x}, \theta^{-1}\boldsymbol{y}) \in \mathcal{P}\}. \tag{2.5}$$

Clearly, $D(\boldsymbol{x}, \boldsymbol{y}) \leq 1$. If $D(\boldsymbol{x}, \boldsymbol{y}) = 1$ then $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}^B$; *i.e.,* that the point $(\boldsymbol{x}, \boldsymbol{y})$ lies on the boundary of $\mathcal{P}$, and the firm is technically efficient. One can similarly define the Shephard (1970) input distance function, which provides a normalized measure of Euclidean distance from a point $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}$ to $\mathcal{P}^B$ in a direction orthogonal to $\boldsymbol{y}$. Alternatively, one could

[2] Grosskopf (1986) gives additional discussion of returns to scale.

employ measures of hyperbolic graph efficiency defined by Färe *et al.* (1985), directional distance functions as defined by Färe *et al.* (1996), or perhaps other measures. From a purely technical viewpoint, any of these can be used to measure technical efficiency; the only real difference is in the direction in which distance to the technology is measured. However, if one wishes to take account of behavioral implications, then this might influence the choice between the various possibilities. We present our methodology only in terms of the output orientation to conserve space, but our discussion can be adapted to the other cases by straightforward changes in our notation.

Analogous to (2.5), define

$$D^{crs}(\boldsymbol{x}, \boldsymbol{y}) \equiv \inf\{\theta > 0 | (\boldsymbol{x}, \theta^{-1}\boldsymbol{y}) \in \mathcal{V}\}. \tag{2.6}$$

Then $D^{crs}(\boldsymbol{x}, \boldsymbol{y}) \leq D(\boldsymbol{x}, \boldsymbol{y}) \leq 1$ gives a normalized measure of Euclidean distance from the point $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}$ to the boundary of $\mathcal{V}$ in a direction orthogonal to $\boldsymbol{x}$. If $D(\boldsymbol{x}, \boldsymbol{y}) > D^{crs}(\boldsymbol{x}, \boldsymbol{y})$, then necessarily $\mathcal{P} \neq \mathcal{V}$, and $\mathcal{P}^B$ does not exhibit CRS everywhere. If $D(\boldsymbol{x}, \boldsymbol{y}) = D^{crs}(\boldsymbol{x}, \boldsymbol{y})$, then the distance from $(\boldsymbol{x}, \boldsymbol{y})$ to the boundary of $\mathcal{V}$ is the same as the distance to the boundary of $\mathcal{P}$ along the same path, and the projection of $(\boldsymbol{x}, \boldsymbol{y})$ onto $\mathcal{P}^B$ along this path yields a point on $\mathcal{P}^B$ where $\mathcal{P}^B$ exhibits CRS in the sense defined above. The same may not be true for another point, however; *i.e.,* it does not follow that $\mathcal{P} = \mathcal{V}$.

Consider the set

$$\mathcal{V}^{nirs} = \mathcal{P} \bigcup \{(\boldsymbol{x}, \boldsymbol{y}) \mid (\boldsymbol{x}, \boldsymbol{y}) \notin \mathcal{P},\ (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{V},\ (\lambda\boldsymbol{x}, \lambda\boldsymbol{y}) \in \mathcal{P} \text{ for some } \lambda > 1\}. \tag{2.7}$$

Then $\mathcal{P} \subseteq \mathcal{V}^{nirs} \subseteq \mathcal{V}$. If $\mathcal{P} = \mathcal{V}^{nirs} \neq \mathcal{V}$, then $\mathcal{P}^B$ exhibits nonincreasing returns to scale (NIRS), *i.e.,* either constant or decreasing returns to scale. Analogous to (2.5), distance from a point $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}$ to the boundary of $\mathcal{V}^{nirs}$ in the direction orthogonal to $\boldsymbol{x}$ is given by

$$D^{nirs}(\boldsymbol{x}, \boldsymbol{y}) \equiv \inf\{\theta > 0 | (\boldsymbol{x}, \theta^{-1}\boldsymbol{y}) \in \mathcal{V}^{nirs}\} \leq 1. \tag{2.8}$$

The distance functions defined by (2.5) and (2.6) may be used to construct a measure of scale efficiency along the lines of Färe *et al.* (1985), namely

$$s(\boldsymbol{x}, \boldsymbol{y}) \equiv \frac{D^{crs}(\boldsymbol{x}, \boldsymbol{y})}{D(\boldsymbol{x}, \boldsymbol{y})} \leq 1. \tag{2.9}$$

Similarly, we may also define

$$\eta(\boldsymbol{x}, \boldsymbol{y}) \equiv \frac{D^{nirs}(\boldsymbol{x}, \boldsymbol{y})}{D(\boldsymbol{x}, \boldsymbol{y})} \leq 1. \tag{2.10}$$

¿From the preceding discussion, if $s(\boldsymbol{x}, \boldsymbol{y}) = 1$ then $\mathcal{P}^B$ exhibits CRS at the point where $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}$ is projected onto $\mathcal{P}^B$ in a direction orthogonal to $\boldsymbol{x}$. A firm operating at $(\boldsymbol{x}, \boldsymbol{y})$ is then said to be *output scale efficient.* On the other hand, if $s(\boldsymbol{x}, \boldsymbol{y}) < 1$ and the technology exhibits increasing, constant, and decreasing returns at different locations, then the firm at $(\boldsymbol{x}, \boldsymbol{y})$ operates under the decreasing returns portion of the technology if $\eta(\boldsymbol{x}, \boldsymbol{y}) = 1$, or under the increasing returns portion of the technology if $\eta(\boldsymbol{x}, \boldsymbol{y}) < 1$.

3. ESTIMATING DISTANCE FUNCTIONS

Unfortunately, neither $\mathcal{P}$, $\mathcal{V}$, nor $\mathcal{V}^{nirs}$ are observed. Consequently, the distance functions $D(\boldsymbol{x}, \boldsymbol{y})$, $D^{crs}(\boldsymbol{x}, \boldsymbol{y})$, and $D^{nirs}(\boldsymbol{x}, \boldsymbol{y})$, as well as the scale efficiency measure $s(\boldsymbol{x}, \boldsymbol{y})$, are also not observed, and hence must be estimated from data. Given a sample of n observations of inputs and outputs of firms denoted by

$$\mathcal{S}_n = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) \mid i = 1, \ldots, n\}, \tag{3.1}$$

we can construct an estimator $\widehat{\mathcal{P}}$ of $\mathcal{P}$, which also implies an estimator $\widehat{\mathcal{P}}^B$ of $\mathcal{P}^B$. Substituting $\widehat{\mathcal{P}}$ for $\mathcal{P}$ in (2.5) yields a distance function estimator $\widehat{D}(\boldsymbol{x}, \boldsymbol{y})$ of $D(\boldsymbol{x}, \boldsymbol{y})$. Whereas $D(\boldsymbol{x}, \boldsymbol{y})$ measures distance in the output direction from the point $(\boldsymbol{x}, \boldsymbol{y})$ to $\mathcal{P}^B$, $\widehat{D}(\boldsymbol{x}, \boldsymbol{y})$ measures distance from the same point to $\widehat{\mathcal{P}}^B$, in the same direction.

Much of the DEA literature obscures the distinction between measures of *true* efficiency such as $D(\boldsymbol{x}, \boldsymbol{y})$, and *estimators* such as $\widehat{D}(\boldsymbol{x}, \boldsymbol{y})$ which, when computed from observed data, yield *estimates* of the true efficiency measures. Since the true production set is unobserved, all we can hope for is an estimate of efficiency, rather than any direct measure.

Several estimators of $\mathcal{P}$ are possible. For a single point $(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathcal{S}_n$, the free disposal hull is defined by

$$\mathcal{F}(\boldsymbol{x}_i, \boldsymbol{y}_i) \equiv \{(\boldsymbol{x}, \boldsymbol{y}) \mid \boldsymbol{x} \geq \boldsymbol{x}_i,\ 0 \leq \boldsymbol{y} \leq \boldsymbol{y}_i\}, \tag{3.2}$$

where the inequalities are understood to be element-wise. The free disposal hull of all of the points in $\mathcal{S}_n$ is merely the union of the free disposal hulls for each of the n points in $\mathcal{S}_n$, or

$$\widehat{\mathcal{P}}_n^{fdh} \equiv \left\{ \bigcup_{i=1}^{n} \mathcal{F}(\boldsymbol{x}_i, \boldsymbol{y}_i) \mid i = 1,\ \ldots\ ,\ n \right\}. \tag{3.3}$$

Deprins *et al.* (1984) proposed efficiency estimators based on using $\widehat{\mathcal{P}}_n^{fdh}$ to estimate $\mathcal{P}$; Korostelev *et al.* (1995a) prove that $\widehat{\mathcal{P}}_n^{fdh}$ is a consistent estimator of $\mathcal{P}$ and give rates of convergence.

Alternatively, let $\mathcal{C}_n$ denote the convex hull of $\mathcal{S}_n$. Then

$$\begin{aligned} \widehat{\mathcal{P}}_n^{vrs} &\equiv \mathcal{C}_n \bigcup \widehat{\mathcal{P}}_n^{fdh} \\ &= \{(\boldsymbol{x}, \boldsymbol{y}) \mid \boldsymbol{y} \leq \boldsymbol{Y}\boldsymbol{\tau},\ \boldsymbol{x} \geq \boldsymbol{X}\boldsymbol{\tau},\ \boldsymbol{i}\boldsymbol{\tau} = 1,\ \boldsymbol{\tau} \in \mathbb{R}_+^n\} \end{aligned} \tag{3.4}$$

provides another estimator of $\mathcal{P}$, where $\boldsymbol{Y} = [\boldsymbol{y}_1\ \ldots\ \boldsymbol{y}_n]$, $\boldsymbol{X} = [\boldsymbol{x}_1\ \ldots\ \boldsymbol{x}_n]$, with each $\boldsymbol{x}_i$, $\boldsymbol{y}_i$ $i = 1, \ldots, n$ denoting the $(p \times 1)$ and $(q \times 1)$ vectors of observed inputs and outputs (respectively), $\boldsymbol{i}$ is a $(1 \times n)$ vector of ones, and $\boldsymbol{\tau} = [\tau_1\ \ldots\ \tau_n]'$ is a $(n \times 1)$ vector of intensity variables. Korostelev *et al.* (1995b) prove that $\widehat{\mathcal{P}}_n^{vrs}$ is a consistent estimator of $\mathcal{P}$ and give rates of convergence.

The convex polyhedral cone (with vortex at the origin) spanned by $\mathcal{S}_n$ (or equivalently, by $\widehat{\mathcal{P}}_n^{fdh}$),

$$\widehat{\mathcal{V}}_n = \{(\boldsymbol{x}, \boldsymbol{y}) \mid \boldsymbol{y} \leq \boldsymbol{Y}\boldsymbol{\tau},\ \boldsymbol{x} \geq \boldsymbol{X}\boldsymbol{\tau},\ \boldsymbol{\tau} \in \mathbb{R}_+^n\}, \tag{3.5}$$

gives an estimator of $\mathcal{V}$. It is straightforward to extend the results of Korostelev *et al.* (1995b) to prove that $\widehat{\mathcal{V}}_n$ is a consistent estimator of $\mathcal{V}$, and therefore a consistent estimator of $\mathcal{P}$ if $\mathcal{P} = \mathcal{V}$, *i.e.,* if $\mathcal{P}^B$ exhibits CRS everywhere. Clearly,

$$\widehat{\mathcal{P}}_n^{vrs} \subseteq \widehat{\mathcal{V}}_n. \tag{3.6}$$

If $\mathcal{P}$ is convex, and $\mathcal{P}^B$ does not exhibit CRS everywhere, then $\widehat{\mathcal{P}}^{vrs}$ is a consistent estimator of $\mathcal{P}$, but $\widehat{\mathcal{V}}_n$ is not. We say that $\widehat{\mathcal{V}}_n$ is constrained relative to $\widehat{\mathcal{P}}_n^{vrs}$; *i.e.,* $\widehat{\mathcal{V}}_n$ is a more restrictive estimator of $\mathcal{P}$ than $\widehat{\mathcal{P}}_n^{vrs}$ in the sense that $\widehat{\mathcal{P}}_n^{vrs}$ will be consistent regardless of the shape of the convex set $\mathcal{P}$, while $\widehat{\mathcal{V}}_n$ will be consistent only if $\mathcal{P}^B$ exhibits constant returns to scale everywhere, *i.e.,* if $\mathcal{P} = \mathcal{V}$.

An estimator of $\mathcal{V}^{nirs}$ is provided by

$$\widehat{\mathcal{V}}_n^{nirs} = \{(\boldsymbol{x}, \boldsymbol{y}) \mid \boldsymbol{y} \leq \boldsymbol{Y\tau},\ \boldsymbol{x} \geq \boldsymbol{X\tau},\ \boldsymbol{i\tau} \leq 1,\ \boldsymbol{\tau} \in \mathbb{R}_+^n\}. \tag{3.7}$$

By construction,

$$\widehat{\mathcal{P}}_n^{vrs} \subseteq \widehat{\mathcal{V}}_n^{nirs} \subseteq \widehat{\mathcal{V}}_n. \tag{3.8}$$

Once again, results from Korostelev *et al.* (1995b) can be extended to prove that $\widehat{\mathcal{V}}_n^{nirs}$ will be a consistent estimator of $\mathcal{P}$ if $\mathcal{P}^B$ exhibits only constant or decreasing returns to scale (*i.e.,* nonincreasing returns to scale). Similar reasoning suggests that $\widehat{\mathcal{P}}_n^{vrs}$, $\widehat{\mathcal{V}}_n^{nirs}$, and $\widehat{\mathcal{V}}_n$ will consistently estimate $\mathcal{P}$ if $\mathcal{P}^B$ exhibits CRS everywhere, although $\widehat{\mathcal{V}}_n$ may have a higher convergence rate in this case. $\widehat{\mathcal{V}}_n^{nirs}$ and $\widehat{\mathcal{V}}_n$ will fail to estimate $\mathcal{P}$ consistently if $\mathcal{P}^B$ exhibits both increasing and decreasing returns to scale at different locations, but $\widehat{\mathcal{P}}_n^{vrs}$ remains consistent in this case.

Estimators of the distance functions defined by (2.5), (2.6), and (2.8) may be obtained by substituting the estimators $\widehat{\mathcal{P}}_n^{vrs}$, $\widehat{\mathcal{V}}_n$, and $\widehat{\mathcal{V}}_n^{nirs}$ for $\mathcal{P}$, $\mathcal{V}$, and $\mathcal{V}^{nirs}$. Doing so yields

$$\left[\widehat{D}_n^{vrs}(\boldsymbol{x}, \boldsymbol{y})\right]^{-1} = \max\left\{\theta \mid \theta\boldsymbol{y} \leq \boldsymbol{Y\tau},\ \boldsymbol{x} \geq \boldsymbol{X\tau},\ \boldsymbol{i\tau} = 1,\ \boldsymbol{\tau} \in \mathbb{R}_+^n\right\}, \tag{3.9}$$

$$\left[\widehat{D}_n^{crs}(\boldsymbol{x}, \boldsymbol{y})\right]^{-1} = \max\left\{\theta \mid \theta\boldsymbol{y} \leq \boldsymbol{Y\tau},\ \boldsymbol{x} \geq \boldsymbol{X\tau},\ \boldsymbol{\tau} \in \mathbb{R}_+^n\right\}, \tag{3.10}$$

and

$$\left[\widehat{D}_n^{nirs}(\boldsymbol{x}, \boldsymbol{y})\right]^{-1} = \max\left\{\theta \mid \theta\boldsymbol{y} \leq \boldsymbol{Y\tau},\ \boldsymbol{x} \geq \boldsymbol{X\tau},\ \boldsymbol{i\tau} \leq 1,\ \boldsymbol{\tau} \in \mathbb{R}_+^n\right\}, \tag{3.11}$$

respectively.

¿From (3.8), it follows that

$$\widehat{D}_n^{crs}(\boldsymbol{x},\boldsymbol{y}) \leq \widehat{D}_n^{nirs}(\boldsymbol{x},\boldsymbol{y}) \leq \widehat{D}_n^{vrs}(\boldsymbol{x},\boldsymbol{y}) \leq 1. \tag{3.12}$$

Moreover, if $\widehat{D}_n^{vrs}(\boldsymbol{x},\boldsymbol{y}) \neq \widehat{D}_n^{crs}(\boldsymbol{x},\boldsymbol{y})$, then necessarily either $\widehat{D}_n^{vrs}(\boldsymbol{x},\boldsymbol{y}) \neq \widehat{D}_n^{nirs}(\boldsymbol{x},\boldsymbol{y}) = \widehat{D}_n^{crs}(\boldsymbol{x},\boldsymbol{y})$ or $\widehat{D}_n^{vrs}(\boldsymbol{x},\boldsymbol{y}) = \widehat{D}_n^{nirs}(\boldsymbol{x},\boldsymbol{y}) \neq \widehat{D}_n^{crs}(\boldsymbol{x},\boldsymbol{y})$. Alternatively, if $\widehat{D}_n^{vrs}(\boldsymbol{x},\boldsymbol{y}) = \widehat{D}_n^{crs}(\boldsymbol{x},\boldsymbol{y})$, then $\widehat{D}_n^{vrs}(\boldsymbol{x},\boldsymbol{y}) = \widehat{D}_n^{nirs}(\boldsymbol{x},\boldsymbol{y}) = \widehat{D}_n^{crs}(\boldsymbol{x},\boldsymbol{y})$.

As shown in the next section, these relationships are useful in constructing tests of returns to scale.

4. TESTING HYPOTHESES REGARDING RETURNS TO SCALE

Aside from the economic significance of whether $\mathcal{P}^B$ exhibits CRS, there are also statistical issues relating to this question. If $\mathcal{P}^B$ exhibits CRS everywhere, then then both $\widehat{D}_n^{crs}(\boldsymbol{x},\boldsymbol{y})$ and $\widehat{D}_n^{vrs}(\boldsymbol{x},\boldsymbol{y})$ are consistent estimators of $D(\boldsymbol{x},\boldsymbol{y})$, but $\widehat{D}_n^{vrs}(\boldsymbol{x},\boldsymbol{y})$ may be less efficient in a statistical sense than $\widehat{D}_n^{crs}(\boldsymbol{x},\boldsymbol{y})$ due to slower convergence. On the other hand, if $\mathcal{P}^B$ exhibits nonconstant returns to scale at some locations, then $\widehat{D}_n^{crs}(\boldsymbol{x},\boldsymbol{y})$ will be an inconsistent estimator of $D(\boldsymbol{x},\boldsymbol{y})$.[3] Therefore, researchers need to know whether the technology is one of constant returns to scale before estimating technical efficiency. Moreover, even if $\mathcal{P}^B$ is known to exhibit nonconstant returns to scale, for policy purposes one might need to know whether a particular region of the technology displays increasing, constant, or decreasing returns to scale.

In some instances, researchers have *a priori* assumed a CRS technology without investigating the possibility that returns to scale are not constant (*e.g.,* Charnes *et al.,* 1978, 1981). However, as we have demonstrated above, this incurs the risk of inconsistently estimating technical efficiency. In other instances, researchers have allowed for the possibility

[3] We state this without formal proof, although results from Kneip *et al.* (1998) can be extended to give a formal proof. The result is intuitively obvious. If $\mathcal{P}^B$ exhibits nonconstant returns to scale at some locations, then for points $(\boldsymbol{x},\boldsymbol{y})$ located below the nonconstant returns to scale regions of $\mathcal{P}^B$, $\widehat{D}_n^{crs}(\boldsymbol{x},\boldsymbol{y}) < D(\boldsymbol{x},\boldsymbol{y})$ even when the sample size n approaches infinity.

of a nonconstant returns to scale technology by using an estimator such as $\widehat{D}_n^{vrs}(\boldsymbol{x}, \boldsymbol{y})$. Given a sample $\mathcal{S}_n$, some researchers have analyzed returns to scale by computing both $\widehat{D}_n^{crs}(\boldsymbol{x}_i, \boldsymbol{y}_i)$ and $\widehat{D}_n^{vrs}(\boldsymbol{x}_i, \boldsymbol{y}_i)$, for each observation $i = 1, \ldots, n$ and then forming ratios $\widehat{D}_n^{crs}(\boldsymbol{x}_i, \boldsymbol{y}_i)/\widehat{D}_n^{vrs}(\boldsymbol{x}_i, \boldsymbol{y}_i) \leq 1$. If $\widehat{D}_n^{crs}(\boldsymbol{x}_i, \boldsymbol{y}_i)/\widehat{D}_n^{vrs}(\boldsymbol{x}_i, \boldsymbol{y}_i) = 1$, then the estimated technology is assumed to exhibit CRS at the point $(\boldsymbol{x}_i, \boldsymbol{y}_i/\widehat{D}_n^{vrs})$ (*i.e.,* where $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is projected onto $\widehat{\mathcal{P}}^B$ in the output direction); otherwise, the estimated technology is assumed to exhibit either increasing or decreasing returns to scale at this point. This type of analysis is typically performed for each observation in the sample, sometimes with further analysis (involving comparisons of $\widehat{D}_n^{vrs}(\boldsymbol{x}_i, \boldsymbol{y}_i)$ with $\widehat{D}_n^{nirs}(\boldsymbol{x}_i, \boldsymbol{y}_i)$) for observations where constant returns to scale are not found to determine whether returns are (locally) increasing or decreasing. Examples of this type of analysis include Färe and Grosskopf (1985), Byrnes *et al.* (1986), Grosskopf and Valdmanis (1987), Dusansky and Wilson (1994, 1995) and Ferrier (1994).

One problem with this type of analysis is that conclusions are drawn in terms of the estimated technology $\widehat{\mathcal{P}}^B$, rather than the true technology $\mathcal{P}^B$. Definitive statements regarding $\widehat{\mathcal{P}}^B$ are possible, but how these translate into inferences regarding $\mathcal{P}^B$ is unclear, due to the lack of a formal statistical testing procedure. In other words, if for a particular observation one finds $\widehat{D}_n^{crs}(\boldsymbol{x}_i, \boldsymbol{y}_i)/\widehat{D}_n^{vrs}(\boldsymbol{x}_i, \boldsymbol{y}_i) < 1$, then without a formal testing procedure, it is impossible to determine whether this is due to nonconstant returns to scale, or due to sampling variation. We attempt to redress this deficiency below.

¿From the viewpoint of statistical hypothesis testing, the first issue is whether $\mathcal{P}^B$ represents a CRS technology. If there is no evidence to the contrary, there seems to be little reason to examine scale efficiency. As discussed in the previous section, the assumption of CRS embodied in the estimator $\widehat{\mathcal{P}}_n^{crs}$ is more restrictive than the assumption of variable returns to scale embodied by $\widehat{\mathcal{P}}_n^{vrs}$. This suggests

$$\text{Test \#1:} \qquad H_0 : \mathcal{P}^B \text{ is globally CRS } \textit{versus } H_1 : \mathcal{P}^B \text{ is VRS.}$$

If H_0 cannot be rejected, we may choose to accept the null hypothesis of CRS. If H_0 is rejected, we may wish to perform another test before accepting H_1, namely

Test #2: $H_0' : \mathcal{P}^B$ is globally NIRS *versus* $H_1 : \mathcal{P}^B$ is VRS.

If both H_0 and H_0' are rejected, we may then choose to accept H_1. Otherwise, if H_0 is rejected but H_0' cannot be rejected, we may accept H_0', the null hypothesis of NIRS.

Given a finite sample $\mathcal{S}_n$, numerous test statistics are possible for testing the null hypothesis of constant returns to scale in Test #1. First, consider an estimator of the Färe *et al.* (1985) measure of scale efficiency $s(\boldsymbol{x}, \boldsymbol{y})$ defined in (2.9), namely

$$\widehat{s} = \widehat{D}_n^{crs}(\boldsymbol{x}, \boldsymbol{y}) \Big/ \widehat{D}_n^{vrs}(\boldsymbol{x}, \boldsymbol{y}). \tag{4.1}$$

This statistic in (4.1) gives a measure of the distance between the CRS and VRS estimated frontiers in the output direction at the point where $(\boldsymbol{x}, \boldsymbol{y})$ is projected onto the estimated frontiers. Evaluating this at each observation $(\boldsymbol{x}_i, \boldsymbol{y}_i) \in \mathcal{S}_n$ yields n estimates $\widehat{s}_i$, $i = 1, \ldots, n$, of scale efficiency corresponding to each observation in $\mathcal{S}_n$. Suppose that for each $i = 1, \ldots, n$ we compare $\widehat{s}_i$ with the appropriate critical value, rejecting the null hypothesis of CRS whenever $\widehat{s}_i$ falls below an appropriate critical value. Such a procedure amounts to n applications of Test #1; if each is independent of the other applications, then under the null H_0 in Test #1, the number of rejections of the null hypothesis H_0 over the n tests would be binomially distributed.

Now suppose we choose a nominal size equal to .05 for each application of the above test. Furthermore, suppose that in r cases (of n total cases) we reject the null hypothesis H_0. We wish to know whether r is large enough to cast doubt on the null hypothesis H_0 in Test #1. If the n individual tests are independent and the true size of the tests is α, then when H_0 is true, the probability of r or more rejections among the n individual tests is given by the incomplete beta function

$$I_\alpha(r, n-r+1) = \sum_{j=r}^{n} \binom{n}{j} \alpha^j (1-\alpha)^{n-j}. \tag{4.2}$$

With a nominal size of 5 percent, we would reject H_0 if $I_{.05}(r, n-r+1) < .05$.[4]

Alternatively, a binomial test can be constructed where we reject H_0 if only one (or more) of the n individual tests results in a rejection. This approach requires that the size of each of the n individual tests, α_ℓ, be chosen much smaller than the global or overall size of the test, α_g. Using (4.2), we have

$$\begin{aligned}\Pr(r \geq 1) &= 1 - \Pr(r = 0) \\ &= 1 - \binom{n}{0} \alpha_\ell^0 (1-\alpha_\ell)^n \\ &= 1 - (1-\alpha_\ell)^n.\end{aligned} \tag{4.3}$$

The null hypothesis H_0 will be rejected when this probability is less than or equal to α_g. Thus, setting $\Pr(r \geq 1) = \alpha_g$ yields

$$\alpha_\ell = 1 - (1-\alpha_g)^{1/n}. \tag{4.4}$$

For example, if $n = 20$ and the overall size of the test is set at $\alpha_g = .05$, then the size of the individual tests should be set at $\alpha_\ell = .00256$.

The binomial tests illustrate the fundamental importance of Test #1. It can only make sense to examine scale efficiency of individual firms if the underlying technology is one of nonconstant returns to scale along some regions. Yet, we do not know this—we must first test the null hypothesis H_0 in Test #1. If H_0 is rejected, one could then in principle use the statistic defined in (4.1) to test hypotheses regarding the scale efficiency of individual firms. Yet, any such tests would necessarily be conditional on the outcome in Test #1, which would affect how one might interpret the results of tests for scale efficiency of individual firms.

Other test statistics for Test #1 are also possible. An obvious candidate is the mean of ratios $\widehat{D}_n^{crs}(\boldsymbol{x}_i, \boldsymbol{y}_i)/\widehat{D}_n^{vrs}(\boldsymbol{x}_i, \boldsymbol{y}_i)$, *i.e,*

$$\widehat{S}_{1n}^{crs} = n^{-1} \sum_{i=1}^{n} \widehat{D}_n^{crs}(\boldsymbol{x}_i, \boldsymbol{y}_i) \Big/ \widehat{D}_n^{vrs}(\boldsymbol{x}_i, \boldsymbol{y}_i), \tag{4.5}$$

[4] If the n individual tests are not independent, then the binomial formula is incorrect and the binomial test is questionable. In particular, since each test involves the same estimate of the production set, this may be a crucial issue. The bootstrap methodology we propose in section 5 ensures independence among the individual tests.

which is an estimator of $S_1^{crs} = n^{-1}\sum_{i=1}^{n}[D_n^{crs}(\boldsymbol{x}_i,\boldsymbol{y}_i)/D_n^{vrs}(\boldsymbol{x}_i,\boldsymbol{y}_i)]$. If H_0 is true, then $D_n^{crs}(\boldsymbol{x}_i,\boldsymbol{y}_i) = D_n^{vrs}(\boldsymbol{x}_i,\boldsymbol{y}_i)$ for all $i = 1,\ldots,n$, and $S^{crs} = 1$; otherwise, $S^{crs} < 1$. By construction, $\widehat{S}_{1n}^{crs} \leq 1$; the null hypothesis H_0 will be rejected when $\widehat{S}_{1n}^{crs}$ is *significantly* less than unity.

Instead of using the mean of ratios as in (4.1), we might use a ratio of means as our test statistic in Test #1:

$$\widehat{S}_{2n}^{crs} = \sum_{i=1}^{n} \widehat{D}_n^{crs}(\boldsymbol{x}_i,\boldsymbol{y}_i) \Big/ \sum_{i=1}^{n} \widehat{D}_n^{vrs}(\boldsymbol{x}_i,\boldsymbol{y}_i). \tag{4.6}$$

Other possibilities include[5]

- the median of ratios:

$$\widehat{S}_{3n}^{crs} = \mathrm{Med}\left\{\widehat{D}_n^{crs}(\boldsymbol{x}_i,\boldsymbol{y}_i)/\widehat{D}_n^{vrs}(\boldsymbol{x}_i,\boldsymbol{y}_i)\right\}_{i=1}^{n}. \tag{4.7}$$

- a ratio of medians:

$$\widehat{S}_{4n}^{crs} = \mathrm{Med}\left\{\widehat{D}_n^{crs}(\boldsymbol{x}_i,\boldsymbol{y}_i)\right\}_{i=1}^{n} \Big/ \mathrm{Med}\left\{\widehat{D}_n^{vrs}(\boldsymbol{x}_i,\boldsymbol{y}_i)\right\}_{i=1}^{n}. \tag{4.8}$$

- the 10 percent trimmed mean of ratios:

$$\widehat{S}_{5n}^{crs} = \mathrm{Trim}_{10}\left\{\widehat{D}_n^{crs}(\boldsymbol{x}_i,\boldsymbol{y}_i)\Big/\widehat{D}_n^{vrs}(\boldsymbol{x}_i,\boldsymbol{y}_i)\right\}_{i=1}^{n}. \tag{4.9}$$

- a ratio of 10 percent trimmed means:

$$\widehat{S}_{6n}^{crs} = \mathrm{Trim}_{10}\left\{\widehat{D}_n^{crs}(\boldsymbol{x}_i,\boldsymbol{y}_i)\right\}_{i=1}^{n} \Big/ \mathrm{Trim}_{10}\left\{\widehat{D}_n^{vrs}(\boldsymbol{x}_i,\boldsymbol{y}_i)\right\}_{i=1}^{n}. \tag{4.10}$$

The mean of ratios in (4.5) has an intuitive geometric interpretation. To see this, consider the ratio $\widehat{D}_n^{crs}(\boldsymbol{x}_i,\boldsymbol{y}_i)/\widehat{D}_n^{vrs}(\boldsymbol{x}_i,\boldsymbol{y}_i)$. At the point where $(\boldsymbol{x}_i,\boldsymbol{y}_i)$ is projected onto

[5] Given a set of n scalars $A = \{a_1,\ldots,a_n\}$, we define $\mathrm{Med}\,A$ as the median of the elements in A. In addition, we define $\mathrm{Trim}_{10}\,A$ as the 10 percent trimmed mean of the elements in A obtained by arranging the elements of A in in algebraic order, deleting 5 percent of the elements from either end of the sorted array, and computing the mean of the remaining 90 percent of the elements of A. The median and trimmed mean are more robust measures of location than the sample mean when the underlying distribution is skewed, as in the present case.

the estimated frontiers in the output direction, this ratio gives the distance between the estimated VRS frontier and the estimated CRS frontier. If this distance is small on average, we would not want to reject the null hypothesis H_0 in Test #1. The ratio of means in (4.6) is a natural variation on this idea. The test statistics defined in (4.7)–(4.10) are analogous to those defined in (4.5)–(4.6), except that robust measures of location (namely, medians and trimmed means) are used rather than arithmetic means.

If, after employing one of the tests described above, the null hypothesis of global constant returns to scale is rejected, the researcher may wish to perform Test #2 before accepting H_1. For Test #2, test statistics $\widehat{\eta}_i$, $\widehat{S}_{1n}^{nirs}$, ..., $\widehat{S}_{6n}^{nirs}$ analogous to $\widehat{s}_i$, $\widehat{S}_{1n}^{crs}$, ..., $\widehat{S}_{6n}^{crs}$ can be defined by replacing $\widehat{D}_n^{crs}(\boldsymbol{x}_i, \boldsymbol{y}_i)$ with $\widehat{D}_n^{nirs}(\boldsymbol{x}_i, \boldsymbol{y}_i)$ in (4.1) and (4.5)–(4.10), respectively, where $\widehat{\eta}_i$ is an estimator of $\eta(\boldsymbol{x}_i, \boldsymbol{y}_i)$ defined in (2.10).

To employ either of these tests, we must obtain appropriate critical values or estimate p-values. To estimate the appropriate critical values, we extend the bootstrap methodology developed in Simar and Wilson (1998a) as discussed in the next section.

5. BOOTSTRAPPING THE TEST STATISTICS

In each of our testing problems, we have a univariate parameter represented by ω; for example, ω might represent $s(\boldsymbol{x}, \boldsymbol{y})$. In each of our tests, if the null hypothesis H_0 is true, then $\omega = \omega_0$. Otherwise, if the alternative hypothesis H_1 is true, then $\omega < \omega_0$. In addition, in each of our tests, we know $\omega_0 = 1$ when the null hypothesis is true, but ω is unknown. For each test we also have a consistent estimator of ω, $\widehat{\omega}$. For example, in the case of Test #1, ω might represent $S_1^{crs} = n^{-1} \sum_{i=1}^{n} [D_n^{crs}(\boldsymbol{x}_i, \boldsymbol{y}_i)/D_n^{vrs}(\boldsymbol{x}_i, \boldsymbol{y}_i)]$, which is estimated by $\widehat{S}_{1n}^{crs}$ defined in (4.5).

For each of our tests, we reject the null hypothesis if $\widehat{\omega} \leq \omega_0 - c_\alpha$, where for a test with size equal to α, $c_\alpha > 0$ is such that

$$\Pr(\widehat{\omega} \leq \omega_0 - c_\alpha \mid H_0) = \alpha. \tag{5.1}$$

Unfortunately, however, we do not know the distribution of our test statistic $\widehat{\omega}$ under H_0, which would allow us to determine the appropriate value for c_α. The bootstrap offers a solution to this problem.

Efron's (1979) bootstrap involves replicating the data generating process which yielded the original sample $\mathcal{S}_n$ to generate B pseudosamples $\mathcal{S}^*_{bn}$, $b = 1, \ldots, B$, each with n observations. Then, the original estimation method is applied to each of the pseudo samples to yield bootstrap estimates $\widehat{\omega}^*_b$. This yields an observed empirical distribution for $(\widehat{\omega}^* - \widehat{\omega})$; the idea behind the bootstrap is to use this distribution to approximate the unknown distribution of $(\widehat{\omega} - \omega_0)$ under H_0.

The crucial point is that the pseudosamples $\mathcal{S}^*_{bn}$ must be generated in a way so that

$$(\widehat{\omega} - \omega)|H_0 \;\sim\; (\widehat{\omega}^* - \widehat{\omega})|H_0, \mathcal{S}_n; \tag{5.2}$$

since $\omega = \omega_0 = 1$ under the null hypothesis, (5.2) is equivalent to

$$(\widehat{\omega} - 1)|H_0 \;\sim\; (\widehat{\omega}^* - \widehat{\omega})|H_0, \mathcal{S}_n. \tag{5.3}$$

In particular, (5.2) requires that the bootstrap pseudosamples be generated under the null hypothesis. In addition, since for each of our tests the test statistic represented by $\widehat{\omega}$ is composed of distance function estimates, the pseudosamples must be generated so that the empirical bootstrap distributions of the bootstrap distance function estimates consistently estimate the sampling distributions of the original distance function estimates. As demonstrated in Simar and Wilson (1998a, 1998b), the naive bootstrap based on constructing pseudosamples by resampling from the empirical distribution of the $(\boldsymbol{x}, \boldsymbol{y})$ pairs does not accomplish this, but a smoothing procedure along the lines of Simar and Wilson (1998a) does.

Given the bootstrap values $\widehat{\omega}^*_b$, $b = 1, \ldots, B$, we can approximate c_α by c^*_α, defined by

$$\Pr\left(\widehat{\omega}^* \leq \widehat{\omega} - c^*_\alpha \mid H_0, \mathcal{S}_n\right) = \alpha, \tag{5.4}$$

which is the bootstrap analog of (5.1). The conditioning on $\mathcal{S}_n$ in (5.4) is due to the fact that the distribution of $(\widehat{\omega}^* - \widehat{\omega})$ in (5.2) is conditioned on $\mathcal{S}_n$. Mechanically, this involves sorting the values $(\widehat{\omega}_b^* - \widehat{\omega})$, $b = 1, \ldots, B$ by algebraic value, deleting $(1-\alpha)\times 100$ percent of the elements at the right-hand end of this sorted array, and then setting $-c_\alpha^*$ equal to the right-hand endpoint of the resulting (sorted) array.

Given (5.2)–(5.3), we obtain the bootstrap approximation

$$\Pr\left(\widehat{\omega} \leq \omega_0 - c_\alpha^* \mid H_0, \mathcal{S}_n\right) \approx \alpha \tag{5.5}$$

by substituting c_α^* for c_α in (5.1). Recalling that $\omega_0 = 1$ under the null hypothesis in our tests, we reject the null if

$$\widehat{\omega} \leq 1 - c_\alpha^* \tag{5.6}$$

for a test of size α.

Alternatively, the testing problem can be approached in terms of p-values (*e.g.,* see Efron and Tibshirani, 1993, chapters 15–16). Let $\widehat{\omega}_{obs}$ be the observed value of our test statistic $\widehat{\omega}$ in a particular application. Then the p-value for H_0 is

$$p = \Pr(\widehat{\omega} \leq \widehat{\omega}_{obs} | H_0); \tag{5.7}$$

if p were known, we would reject the null hypothesis H_0 when p is sufficiently small, say less than .05. Under H_0, (5.7) is equivalent to

$$p = \Pr(\widehat{\omega} - \omega_0 \leq \widehat{\omega}_{obs} - \omega_0 | H_0). \tag{5.8}$$

Using the bootstrap as described above, this can be approximated by

$$\widehat{p} = \Pr(\widehat{\omega}^* - \widehat{\omega} \leq \widehat{\omega}_{obs} - \omega_0 \mid H_0, \mathcal{S}_n). \tag{5.9}$$

Since conditionally on $\mathcal{S}_n$, $\widehat{\omega} = \widehat{\omega}_{obs}$, (5.9) is equivalent to

$$\widehat{p} = \Pr(\widehat{\omega}^* \leq 2\widehat{\omega}_{obs} - \omega_0 \mid H_0, \mathcal{S}_n), \tag{5.10}$$

which can be easily evaluated from the empirical bootstrap distribution obtained from the Monte Carlo simulation. Note that asymptotically, since $\widehat{\omega}$ is consistent, the probability statement in (5.10) is asymptotically equivalent to

$$\widehat{p} = \Pr(\widehat{\omega}^* \leq \widehat{\omega}_{obs} \mid H_0, \mathcal{S}_n). \tag{5.11}$$

This expression is used in Efron and Tibshirani (1993, chapter 16). In each of our Monte Carlo experiments, $\widehat{p}$ defined in (5.11) gives tests with size closer to the nominal size than does (5.10); consequently, all of our results in the next section use (5.11) rather than (5.10). If α is the nominal size of our test, we reject the null hypothesis H_0 when $\widehat{p} \leq \alpha$.

Each of the test statistics proposed in the previous section is a function of the distance function estimators defined in (3.9)–(3.11). Therefore, bootstrap estimates of the distance function estimators are needed to bootstrap the test statistics defined above. Given a pseudosample $\mathcal{S}_n^* = \{(\boldsymbol{x}_i^*, \boldsymbol{y}_i^*) \mid i = 1, \ldots, n\}$ for one of the B bootstrap replications, we can adapt (3.9)–(3.11) to measure the distance from each observation in the original sample $\mathcal{X}_n$ to the frontiers estimated from the pseudodata in $\mathcal{S}_n^*$. For the VRS estimator in (3.9), this amounts to solving

$$\widehat{D}_n^{vrs^*}(\boldsymbol{x}_i, \boldsymbol{y}_i) = \max\left\{\theta \mid \boldsymbol{y}_i \leq \boldsymbol{Y}^*\boldsymbol{\tau},\ \boldsymbol{x}_i \geq \boldsymbol{X}^*\boldsymbol{\tau},\ \boldsymbol{i\tau} = 1,\ \boldsymbol{\tau} \in \mathbb{R}_+^n\right\}, \tag{5.12}$$

where $\boldsymbol{Y}^* = [\boldsymbol{y}_1^* \ \ldots \ \boldsymbol{y}_n^*]$ and $\boldsymbol{X} = [\boldsymbol{x}_1^* \ \ldots \ \boldsymbol{x}_n^*]$.

Similarly, for the CRS and NIRS estimators in (3.10)–(3.11), we can compute

$$\widehat{D}_n^{crs^*}(\boldsymbol{x}_i, \boldsymbol{y}_i) = \max\left\{\theta \mid \boldsymbol{y}_i \leq \boldsymbol{Y}^*\boldsymbol{\tau},\ \boldsymbol{x}_i \geq \boldsymbol{X}^*\boldsymbol{\tau},\ \boldsymbol{\tau} \in \mathbb{R}_+^n\right\} \tag{5.13}$$

and

$$\widehat{D}_n^{nirs^*}(\boldsymbol{x}_i, \boldsymbol{y}_i) = \max\left\{\theta \mid \boldsymbol{y}_i \leq \boldsymbol{Y}^*\boldsymbol{\tau},\ \boldsymbol{x}_i \geq \boldsymbol{X}^*\boldsymbol{\tau},\ \boldsymbol{i\tau} \leq 1,\ \boldsymbol{\tau} \in \mathbb{R}_+^n\right\}, \tag{5.14}$$

respectively, to obtain the corresponding bootstrap estimators.

The remaining issues concern the creation of the pseudodata in $\mathcal{S}_n^*$. As discussed in Simar and Wilson (1998a, 1998b), resampling from the empirical distribution of the data (*i.e.,* drawing independently, with replacement from the set of original observations on inputs and outputs, or equivalently, from the set of original distance function estimates) to construct the pseudosamples $\mathcal{S}_n^*$ will lead to inconsistent bootstrap estimation in the present setting. The problem arises because we are, in effect, estimating the boundaries of sets. The empirical distribution of the observed data, for finite samples, places a positive probability mass at the boundary of the estimated production set, and thus provides a poor approximation of the underlying density, which we assume to be continuous.[6]

Using a smooth bootstrap procedure as in Simar and Wilson (1998a) overcomes this problem and yields consistent estimates. When bootstrapping distance function estimates from a single cross-section of data, this may be accomplished by using a univariate kernel estimator of the density of the original distance function estimates, and then drawing from this estimated density to construct the pseudosamples $\mathcal{S}_n^*$. Since for any of our tests we are interested in the distribution of an expression represented in symbolic form by $(\widehat{\omega} - \omega_0)$ *under the null hypothesis,* the pseudodata should be generated under the assumptions of the relevant null hypothesis.

To illustrate, consider Test #1 where the null hypothesis is that the technology exhibits CRS everywhere. Here we must draw values δ_i^*, $i = 1, \ldots, n$ from a consistent estimate of the density of the estimated values $\widehat{D}_n^{crs}(\boldsymbol{x}_i, \boldsymbol{y}_i)$, and then construct pseudo observations $(\boldsymbol{x}_i^*, \boldsymbol{y}_i^*)$ such that

$$\boldsymbol{x}_i^* = \boldsymbol{x}_i, \qquad \boldsymbol{y}_i^* = \delta_i^* \boldsymbol{y}_i / \widehat{D}_n^{crs}(\boldsymbol{x}_i, \boldsymbol{y}_i). \tag{5.15}$$

Since we are using the output orientation, $\boldsymbol{x}_i^*$ is merely $\boldsymbol{x}_i$; if we were using an input orientation, we would set $\boldsymbol{y}_i^* = \boldsymbol{y}_i$ and generate new input vectors.

The usual kernel density estimator for the density of the estimated values $\widehat{D}_n^{crs}(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is given by

$$\widetilde{f}_h(\delta) = n^{-1} h^{-1} \sum_{i=1}^{n} K\left(\frac{\delta - \widehat{D}_n^{crs}(\boldsymbol{x}_i, \boldsymbol{y}_i)}{h} \right), \tag{5.16}$$

where $K(\cdot)$ is a kernel function where $\int K(u)\,du = 1$, $K(u) = K(-u)$, $K(u)$ is piecewise continuous, and h is the bandwidth parameter which controls the amount of smooth-

[6] Even if the underlying density from which the data were drawn has a probability mass along the boundary of the production set, there is no reason to think that drawing from the empirical data distribution would provided an estimate of this mass which would lead to consistent bootstrap estimation. The naive bootstrap will lead to a constant amount of probability mass along the frontier, which would not necessarily reflect the actual mass along the frontier.

ing. Note, however, that by construction, the support of the underlying density of the $\widehat{D}_n^{crs}(\boldsymbol{x}_i, \boldsymbol{y}_i)$ is bounded on the right at unity. The density estimator in (5.16) can be shown to be biased and inconsistent in the presence of this boundary condition.[7]

To overcome this problem, we use the univariate reflection method described by Silverman (1986) and employed in Simar and Wilson (1998a). Let $\boldsymbol{\Delta} = [\delta_i]$ be an $(n \times 1)$ matrix whose elements are the distance function estimates for which we want to estimate the density. Since we are using output distance functions, $\delta_i \leq 1$ for all $i = 1, \ldots, n$. Then the reflected data are described by $\boldsymbol{\Delta}_R = [\delta_{Ri}]$, $\delta_{Ri} = 2 - \delta_i \geq 1$. Now we have an augmented set of $2n$ observations described by

$$\widetilde{\boldsymbol{\Delta}} = \begin{bmatrix} \boldsymbol{\Delta} \\ \boldsymbol{\Delta}_R \end{bmatrix} = [\widetilde{\delta}_j]. \tag{5.17}$$

Assuming the $2n$ observations in $\widetilde{\boldsymbol{\Delta}}$ represent draws from an unknown density $g(\cdot)$ with unbounded support on the real number line, a consistent kernel estimator of this density is given by

$$\widehat{g}_h(\delta) = \frac{1}{2nh} \sum_{i=1}^{n} \left[K\left(\frac{\delta - \delta_i}{h}\right) + K\left(\frac{\delta - \delta_{Ri}}{h}\right) \right] = \frac{1}{2nh} \sum_{j=1}^{2n} K\left(\frac{\delta - \widetilde{\delta}_j}{h}\right). \tag{5.18}$$

Truncating this on the right at unity yields a consistent estimator of the density of the observations in $\boldsymbol{\Delta}$, *i.e.*, the original distance function estimates:

$$\widehat{f}_h(\delta) = \begin{cases} 2\widehat{g}_h(\delta) & \text{if } \delta \in (0, 1], \\ 0 & \text{otherwise.} \end{cases} \tag{5.19}$$

Drawing samples from the density estimate $\widehat{f}_h(\delta)$ does not actually require computation of the density estimate itself; computational shortcuts as well as adjustments to ensure that the draws have the same first and second moments as the $\widehat{D}_n^{crs}(\boldsymbol{x}_i, \boldsymbol{y}_i)$ are discussed in Silverman (1986), Efron and Tibshirani (1993), and Simar and Wilson (1998a). Drawing samples from $\widehat{f}_h(\delta)$ does, however, require a value for the bandwidth parameter h. For

[7]The output distance function estimator is also bounded on the left at zero, but since we typically do not observe values near this boundary, the left boundary condition can be ignored without inducing an inconsistency problem.

the density estimator in (5.19) to be consistent, h must be of order $O(n^{-1/5})$ (Silverman, 1986). In the appendix, we discuss a method for selecting a value of h that minimizes approximate mean weighted integrated square error of the density estimate for a given sample $\mathcal{X}_n$.

Our bootstrap procedure for Test #1 using the statistic $\widehat{S}_{1n}^{crs}$ in (4.5) can now be summarized by the following steps:

Algorithm #1:

[1] For each firm $i = 1, \ldots, n$ in $\mathcal{S}_n$, compute $\widehat{D}_n^{vrs}(\boldsymbol{x}_i, \boldsymbol{y}_i)$ and $\widehat{D}_n^{crs}(\boldsymbol{x}_i, \boldsymbol{y}_i)$ from (3.9)–(3.10); compute $\widehat{S}_{1n}^{crs}$ from (4.1).

[2] Use a kernel estimator to obtain a consistent estimate $\widehat{f}$ of the density of $\left\{\widehat{D}_n^{crs}(\boldsymbol{x}_i, \boldsymbol{y}_i)\right\}_{i=1}^{n}$.

[3] Draw random deviates $\delta_i^* \ \forall\ i = 1, \ldots, n$ from $\widehat{f}$, making necessary adjustments to ensure that the deviates have the same first and second moments as the $\widehat{D}_n^{crs}(\boldsymbol{x}_i, \boldsymbol{y}_i)$ obtained in step [1], as discussed previously.

[4] Form a pseudo sample $\mathcal{X}_n^* = \{(\boldsymbol{x}_i^*, \boldsymbol{y}_i^*)\}_{i=1}^n$ where $\boldsymbol{x}_i^*$ and $\boldsymbol{y}_i^*$ are defined in (5.15).

[5] Compute bootstrap estimates $\widehat{D}_n^{vrs^*}(\boldsymbol{x}_i, \boldsymbol{y}_i)$ and $\widehat{D}_n^{crs^*}(\boldsymbol{x}_i, \boldsymbol{y}_i)$ for each firm i in the original sample by applying the original estimation methods to the pseudo sample as in (5.12)–(5.13).

[6] Repeat steps [3]–[5] B times to produce a set of B pairs of bootstrap estimates $\left\{\widehat{D}_{nb}^{vrs^*}(\boldsymbol{x}_i, \boldsymbol{y}_i),\ \widehat{D}_{nb}^{crs^*}(\boldsymbol{x}_i, \boldsymbol{y}_i)\right\}_{b=1}^{B}$ for each firm $i = 1, \ldots, n$.

Estimating the density of $\left\{\widehat{D}_n^{crs}(\boldsymbol{x}_i, \boldsymbol{y}_i)\right\}_{i=1}^{n}$ in step [2] and drawing from this density in step [3] ensures that the pseudo data are generated under the null hypothesis of CRS. Once the complete set of bootstrap estimates

$$\left\{\widehat{D}_{nb}^{vrs^*}(\boldsymbol{x}_i, \boldsymbol{y}_i),\ \widehat{D}_{nb}^{crs^*}(\boldsymbol{x}_i, \boldsymbol{y}_i) \mid i = 1, \ldots, n\right\}_{b=1}^{B}$$

has been obtained, it is straightforward to compute bootstrap estimates of the test statistic

$\widehat{S}_{1n}^{crs}$ for each $b = 1, \ldots, B$ as

$$\widehat{S}_{1nb}^{crs^*} = n^{-1} \sum_{i=1}^{n} \widehat{D}_{nb}^{crs^*}(\boldsymbol{x}_i, \boldsymbol{y}_i) \Big/ \widehat{D}_{nb}^{vrs^*}(\boldsymbol{x}_i, \boldsymbol{y}_i). \tag{5.20}$$

Then, following the procedures outlined by equations (5.1)–(5.6) or (5.7)–(5.11), we can use these bootstrap values, together with the original estimate $\widehat{S}_{1n}^{crs}$, to determine c_α^* or $\widehat{p}$.

While we have used the statistic in (4.5) to illustrate our bootstrap procedure, the algorithm can be extended to estimate critical values for the test statistics given by (4.6)–(4.10) by merely rewriting (5.20) to reflect the chosen statistic. For example, if the statistic of interest is $\widehat{S}_{2n}^{crs}$ in (4.6), (5.20) would become

$$\widehat{S}_{2nb}^{crs^*} = \sum_{i=1}^{n} \widehat{D}_{nb}^{crs^*}(\boldsymbol{x}_i, \boldsymbol{y}_i) \Bigg/ \sum_{i=1}^{n} \widehat{D}_{nb}^{vrs^*}(\boldsymbol{x}_i, \boldsymbol{y}_i). \tag{5.21}$$

If the null hypothesis of constant returns to scale in Test #1 is rejected, we may adapt the above bootstrap procedure for Test #2 via additional trivial changes in notation. For the null hypothesis of NIRS in Test #2, we would generate the pseudosamples $\mathcal{X}_n^*$ under the null by replacing $\widehat{D}_n^{crs}(\boldsymbol{x}_i, \boldsymbol{y}_i)$ in steps [1]–[3] in Algorithm #1 and in equation (5.15) with $\widehat{D}_n^{nirs}(\boldsymbol{x}_i, \boldsymbol{y}_i)$, and by replacing $\widehat{D}_{nb}^{crs^*}(\boldsymbol{x}_i, \boldsymbol{y}_i)$ in steps [5]–[6] in Algorithm #1 and equations (5.20)–(5.21) with $\widehat{D}_{nb}^{nirs^*}(\boldsymbol{x}_i, \boldsymbol{y}_i)$.

For the binomial test suggested by (4.2), more substantial changes in the bootstrap algorithm are required. For a given observation i, we can compute $\widehat{s}_i^{crs}$ from (4.1). We could then produce bootstrap values $\widehat{s}_{ib}^{crs^*}$, $b = 1, \ldots, B$ along the lines of Algorithm #1, and then use these bootstrap values to determine a critical value $\widehat{s}_i^{crs} + c_{\alpha,i}^*$ along the lines described above. However, the binomial probability in (4.2) requires that each application of this process over the n observations $i = 1, \ldots, n$ be independent of other applications. To meet this requirement, we must resample separately for each firm in the sample $\mathcal{S}_n$ as described below:

Algorithm #2:

Steps [1]–[4] are the same as in Algorithm #1.

[5] Compute bootstrap estimates $\widehat{D}_n^{vrs^*}(\boldsymbol{x}_i, \boldsymbol{y}_i)$ and $\widehat{D}_n^{crs^*}(\boldsymbol{x}_i, \boldsymbol{y}_i)$ for firm $i = 1$ in the original sample by applying the original estimation methods to the pseudo sample as in (5.12)–(5.13).

[6] Repeat [3]–[5] for firms $i = 2, \ldots, n$ in the original sample.

[7] Repeat [3]–[6] B times to produce a set of B pairs of bootstrap estimates $\left\{\widehat{D}_{nb}^{vrs^*}(\boldsymbol{x}_i, \boldsymbol{y}_i),\ \widehat{D}_{nb}^{crs^*}(\boldsymbol{x}_i, \boldsymbol{y}_i)\right\}_{b=1}^{B}$ for each firm $i = 1, \ldots, n$.

Step [6] in Algorithm #2 involves looping over the firms in the sample for each bootstrap replication in step [7]; this ensures the independence required for the binomial procedure. Once the complete set of bootstrap estimates

$$\left\{\widehat{D}_{nb}^{vrs^*}(\boldsymbol{x}_i, \boldsymbol{y}_i),\ \widehat{D}_{nb}^{crs^*}(\boldsymbol{x}_i, \boldsymbol{y}_i) \mid i = 1,\ \ldots,\ n\right\}_{b=1}^{B}$$

has been obtained, it is straightforward to compute bootstrap estimates of the test statistic $\widehat{s}_i$ for each $b = 1,\ \ldots,\ B$ and $i = 1,\ \ldots,\ n$ as

$$\widehat{s}_{ib}^* = \widehat{D}_{nb}^{crs^*}(\boldsymbol{x}_i, \boldsymbol{y}_i) \Big/ \widehat{D}_{nb}^{vrs^*}(\boldsymbol{x}_i, \boldsymbol{y}_i). \tag{5.22}$$

Then, following the procedure outlined by equations (5.1)–(5.6) or (5.7)–(5.11), we can use these bootstrap values, together with the original estimates $\widehat{s}_i$, to determine either a critical value or a p-value corresponding to each firm in the sample. This will yield r rejections of the null hypothesis of constant returns to scale in the case of Test #1, out of n trials. The binomial probability in (4.2) can then be used to determine if enough rejections have occurred to reject H_0, the hypothesis of global CRS. Or, the bootstrap values s_{ib}^* for each $i = 1, \ldots, n$ can be used to perform n individual tests of size α_ℓ determined from (4.4), and then H_0 can be rejected if these n test produce one or more rejections. As before, the binomial test and Algorithm #2 can be extended to Test #2 by trivial changes in notation along the lines described in discussing Algorithm #1.

6. BANKER'S TESTS

Banker (1996) proposed three statistics for testing the null hypotheses of CRS or NIRS against the alternative hypothesis of VRS. In terms of our notation, for the case of testing the null hypothesis of CRS, the first of Banker's statistics is given by

$$\widehat{F}_{1n}^{crs} = \sum_{i=1}^{n} \left[\left(\widehat{D}^{crs}(\boldsymbol{x}_i, \boldsymbol{y}_i) \right)^{-1} - 1 \right] \Big/ \sum_{i=1}^{n} \left[\left(\widehat{D}^{vrs}(\boldsymbol{x}_i, \boldsymbol{y}_i) \right)^{-1} - 1 \right]. \tag{6.1}$$

Banker's second statistic amounts to

$$\widehat{F}_{2n}^{crs} = \sum_{i=1}^{n} \left[\left(\widehat{D}^{crs}(\boldsymbol{x}_i, \boldsymbol{y}_i) \right)^{-1} - 1 \right]^2 \Big/ \sum_{i=1}^{n} \left[\left(\widehat{D}^{vrs}(\boldsymbol{x}_i, \boldsymbol{y}_i) \right)^{-1} - 1 \right]^2. \tag{6.2}$$

Banker argued that if $D(\boldsymbol{x}_i, \boldsymbol{y}_i) \sim$ exponential, then $\widehat{F}_{1n}^{crs} \overset{\text{asy.}}{\sim} F_{2n,2n}$. Alternatively, if one assumes $D(\boldsymbol{x}_i, \boldsymbol{y}_i) \sim$ half normal, then $\widehat{F}_{2n}^{crs} \overset{\text{asy.}}{\sim} F_{n,n}$.

Banker offered no guidance on how one might decide whether the true distance function values are distributed exponentially or half-normally across production units. Moreover, it is unclear why one should be willing to assume *any* distribution for purposes of hypothesis testing if the distance functions have been estimated nonparametrically. If the researcher were willing to make such an assumption for purposes of testing hypotheses, then why not make use of the distributional assumption in the estimation process? If the distributional assumption is correct, but not used in the estimation process, then presumably the resulting efficiency estimates will be statistically inefficient relative to an estimation procedure that makes use of the distributional information. On the other hand, if the distributional assumption is incorrect, then there is no reason to believe that tests based on the statistics in (6.1)–(6.2) would perform well in terms of size or power.

Kittelsen (1997) suggests two additional problems with Banker's appoach. The distance function estimators used to define the statistics in (6.1)–(6.2) are biased in finite samples, and have very low convergence rates in high dimensions. Thus, even if the distributional assumptions are correct, the resulting statistics may not be F distributed in finite samples. Moreover, the numerator is correlated with the denominator in each case, compounding the problem. Kittelsen presents Monte Carlo evidence showing that these tests perform poorly in terms of size and power with $p = q = 1$ and $n = 100$ even when the distributional assumptions are correct.

Banker's third statistic is given by

$$\widehat{K}_n^{crs} = \max \left[\left(\widehat{D}^{crs}(\boldsymbol{x}_i, \boldsymbol{y}_i) \right)^{-1} - \left(\widehat{D}^{vrs}(\boldsymbol{x}_i, \boldsymbol{y}_i) \right)^{-1} \;\middle|\; i = 1, \ldots, n \right], \tag{6.3}$$

which is the Kolmogorov-Smirnov test statistic. Kim and Jenrich (1973) give an algorithm for computing the exact sampling distribution for this statistic when $n^2 \leq 10^4$, as well as a very good approximation for cases where $n^2 > 10^4$.

Each of the statistics defined by (6.1)–(6.3) can be adapted to the case of Test #2, where the null hypothesis is NIRS, by merely replacing $\widehat{D}^{crs}(\boldsymbol{x}_i, \boldsymbol{y}_i)$ with $\widehat{D}^{nirs}(\boldsymbol{x}_i, \boldsymbol{y}_i)$. The approach suggested by the third statistic seems more reasonable than the first two, since no distributional assumptions are necessary. Moreover, using $\widehat{K}_n^{crs}$ has a potential advantage over those defined in section 4, since (possibly computationally burdensome) bootstrapping can be avoided. Unfortunately, however, the distance function estimators in (6.3) are not independent, as observed earlier, violating a key assumption in deriving the sampling distribution for the Kolmogorov-Smirnov statistic. Our Monte-Carlo experiments, as well as Kittelsen's (1997), indicate that tests based on this statistic have grossly incorrect size.

7. EVIDENCE FROM MONTE CARLO EXPERIMENTS

Although the bootstrap procedures discussed in the section 5 are consistent, errors (either type I or type II) may occur in finite samples due to sampling variation in the distance function estimators as well as additional noise introduced by the resampling process itself.[8] As a result, the true sizes of the tests described in section 5 may differ from the nominal value α. In order to examine the performance of our bootstrap tests of returns to scale, we ran a series of Monte Carlo experiments.

[8] In particular, kernel estimators, while consistent, are slow to converge. Resampling from kernel estimates of the density of distance function estimates might be a significant source of noise in the bootstrap process.

For each experiment, we used $B = 2000$ bootstrap replications, and performed 1000 Monte Carlo trials.[9] For each Monte Carlo trial, after generating the input and output data, we computed the relevant test statistics. Then, we applied the bootstrap procedures outlined in section 5 to obtain critical values for each test statistic at nominal significance levels $\alpha = .3$, .25, .2, .15, .1, .05, .02, and .01. On each Monte Carlo trial, we used a standard normal kernel function and with a bandwidth that minimized mean weighted integrated square error as discussed in the Appendix. We estimated the true size of each test for each experiment by recording the number of times the null hypothesis was rejected at each significance level over all Monte Carlo trials, then dividing by the number of Monte Carlo trials. In each of our experiments, the number of outputs, q, was set equal to one to simplify the data-simulation process and to reduce the computational burden. All input data ($\boldsymbol{x}_i$) were simulated by generating identically, independently distributed (iid) pseudo random uniform deviates on the interval (1,9).[10]

We first examined the performance of the bootstrap in the context of Test #1. We ran twelve experiments, with one output ($q = 1$), either one or two inputs ($p \in \{1, 2\}$), and $n \in \{20, 40, 60\}$. In six experiments we resampled for the bootstrap as described by Algorithm #1 to obtain results for the statistics $\widehat{S}_{1n}^{crs}, \ldots, \widehat{S}_{6n}^{crs}$ defined in section 4, as well as the statistics $\widehat{F}_{1n}^{crs}$, $\widehat{F}_{2n}^{crs}$, and $\widehat{K}_{n}^{crs}$ defined in section 6. In the other six experiments, we resampled as described by Algorithm #2 to obtain results for the binomial test. To simulate the output data (y_i) under the null hypothesis that $\mathcal{P}^B$ exhibits CRS everywhere, we first generated iid pseudo random standard normal deviates v_i. Then the output for

[9] In effect, our bootstrap procedures are attempting to estimate the tails of sampling distributions. Researchers typically set $B \approx 100$ when computing variances, and $B \approx 1000$ when computing confidence intervals. Since estimating the tail of a distribution constitutes a more difficult problem, we chose $B = 2000$ to ensure adequate coverage. For the experiment represented in Table 1 with $n = 20$, we also used 4000 and then 10,000 replications to check whether the choice of B would influence our results for the statistics $\widehat{S}_{1n}^{crs}, \ldots, \widehat{S}_{6n}^{crs}$. In both instances, we obtained estimates of true size very close to those reported in Table 1 with $B = 2000$.

[10] Pseudo random uniform (0,1) deviates were generated using the multiplicative congruential generator with modulus $(2^{31} - 1)$ and multiplier 16807 (see Lewis *et al.*, 1969). Pseudo random normal deviates were generated via the Box-Muller method (see Press *et al.*, 1986, pp. 202–203).

the ith simulated firm was computed as

$$y_i = \prod_{j=1}^{p} x_{ij}^{1/p} e^{-0.1|v_i|}, \qquad i = 1, \ldots, n. \tag{7.1}$$

The Monte Carlo results for Test #1, with one input ($p = 1$), appear in Table 1. Similarly, Monte Carlo results for Test #1 with two inputs ($p = 2$) appear in Table 2.

The first binomial test performs quite poorly in terms of its size, which is larger than the size of any of the other tests in Tables 1–2 in almost every instance. The second binomial test also performs quite poorly, but its size is far smaller than the nominal size in every instance, with the true size becoming smaller (for a given nominal size) as the sample size n increases. This seemingly perverse behavior is explained by considering the expression in (4.4). With $n = 60$ and an overall nominal test size of .1, the individual tests comprising the second binomial test procedure should have size .00175—apparently requiring a higher degree of accuracy in estimating the tail of the sampling distribution than is acheived with $B = 2000$ bootstrap replications. The problem is worse when the overall nominal test size is .05 or less. While the second binomial test might be more accurate if a much larger number of bootstrap replications were used, doing so would substantially increase the computational burden of the test, and make our Monte Carlo experiments infeasible.

Banker's (1996) test statistics defined in (6.1)–(6.2) also perform poorly. This is to be expected for the first two of Banker's statistics, since the DGP in (7.1) used for our Monte Carlo experiments in Tables 1–2 do not reflect the distributional assumptions underlying these tests.[11] Although $\widehat{F}_{1n}^{crs}$ results in a test with true size close to the nominal size for nominal sizes in the range .1–.01 and for $n = 20$ in Table 1, its performance deteriorates in Table 1 as the sample size is increased. In Table 2, the true sizes of tests based on this statistic are much too large in every instance. The second of Banker's statistics, $\widehat{F}_{2n}^{crs}$, yields tests with true size that is near zero for small nominal sizes. But, the true size of these tests jumps suddenly to values far in excess of the nominal size as the nominal size is

[11] The specification in (7.1) ensures that, for $p = 1$, we have $D(\boldsymbol{x}_i, \boldsymbol{y}_i) = x_i e^{-0.1|v_i|}/x_i = e^{-0.1|v_i|}$, and hence $\log D(\boldsymbol{x}_i, \boldsymbol{y}_i) \sim$ half normal.

increased. This behaviour again merely reflects the misspecification of the distributional assumptions underlying these tests. While our Monte Carlo experiments are designed in such a way that ensures that tests based on $\widehat{F}_{1n}^{crs}$ and $\widehat{F}_{2n}^{crs}$ will have incorrect size, we would expect similar results in real applications since there is no convincing reason why distance function values should be distributed either exponentially or half-normally. The Kolmororov-Smirnov statistic in (6.3) suggested by Banker avoids the distributional assumptions underlying the other tests suggested by Banker, but this test also performs poorly in our Monte Carlo experiments, yielding tests with true size that is either much too large or too small relative to nominal size.

Of the remaining statistics, $\widehat{S}_{2n}^{crs}$ consistently performs best in terms of size. In Table 1 with one input and one output, the true size is quite close to the nominal size with only 40 observations. In Table 2 where we consider two inputs, the true size is larger, but not as large as the other tests. Moreover, as sample size increases, we see that the size of tests based on $\widehat{S}_{2n}^{crs}$ decreases toward the nominal size more rapidly than for other tests we consider. The increased size in Table 2, relative to Table 1, merely reflects the curse of dimensionality inherent in nonparametric estimators of technical efficiency; with more than 60 observations, we would expect the size of our bootstrap tests to improve, but it is computationally infeasible to demonstrate this in Monte Carlo experiments.

Among the statistics $\widehat{S}_{1n}^{crs}, \ldots, \widehat{S}_{6n}^{crs}$, $\widehat{S}_{4n}^{crs}$ performs most poorly in both Tables 1 and 2, although its performance improves with sample size. The next-poorest performance in Tables 1–2 is by $\widehat{S}_{6n}^{crs}$. There are no compelling differences in the performance of $\widehat{S}_{1n}^{crs}$, $\widehat{S}_{3n}^{crs}$, and $\widehat{S}_{5n}^{crs}$ in Tables 1–2, but these are dominated in almost every instance by the performance of $\widehat{S}_{2n}^{crs}$.

In evaluating the performance of our test statistics and bootstrap procedure for Test #2, we must simulate the data under the null hypothesis of NIRS. We simulate the input data as before. For the case of one input, we compute the output for the ith simulated firm as

$$y_i = \begin{cases} x_i e^{-0.1|v_i|}, & \text{if } x_i \le 5; \\ (\frac{5}{2} + \frac{1}{2}x_i)e^{-0.1|v_i|} & \text{otherwise.} \end{cases} \tag{7.2}$$

For the case of two inputs, we compute the output for the ith simulated firm as

$$y_i = \begin{cases} \sqrt{x_{i1}}\sqrt{x_{i2}}e^{-0.1|v_i|} & \text{if } x_{i1} + x_{i2} \le 11 \\ \Bigg\{ \sqrt{\frac{x_{i2}}{x_{i1}}}(11)\left(\frac{x_{i2}}{x_{i1}}+1\right)^{-1} + & \\ \quad \left[x_{i1} - \left(\frac{x_{i2}}{x_{i1}}\right)(11)\left(\frac{x_{i2}}{x_{i1}}+1\right)^{-1}\right]^{1/4} \times & \\ \quad \left[x_{i2} - (11)\left(\frac{x_{i2}}{x_{i1}}+1\right)^{-1}\right]^{1/4} \Bigg\} e^{-0.1|v_i|} & \text{otherwise.} \end{cases} \tag{7.3}$$

This ensures continuity in the simulated technology. We ran twelve experiments, analogous to those for Test #1. Results for the case of one input are shown in Table 3, while results for the case of two inputs are shown in Table 4.

The results in Tables 3–4 are qualitatively similar to those in Tables 1–2. The binomial tests perform poorly in every case, as do Banker's tests. The binomial tests and Banker's exponential and half-normal tests do not improve with sample size; the Kolmogorov-Smirnov test improves only very slightly with sample size. Among the remaining statistics, $\widehat{S}_{2n}^{nirs}$ has estimated true size closer to nominal size than the other statistics in almost every case. Surprisingly, in view of the previous results, $\widehat{S}_{4n}^{nirs}$ does not stand out as having exceptionally bad performance in Table 3, but $\widehat{S}_{3n}^{nirs}$ does. Note that in Table 4, as the sample size increases from $n = 20$ to $n = 60$, the estimated sizes diverge from the nominal sizes for the first six test statistics in many cases. Statisticsl consistency does not necessarily imply any type of monotonic convergence; the phenomena in Table 4 merely reflects the curse of dimensionality. Apparently, for the test of the null hypothesis of NIRS, with $p + q = 3$, even $n = 60$ is a very small sample size.

There is, of course, a trade-off between size and power in hypothesis testing. Therefore, we also examined the power of tests based on the statistics $\widehat{S}_{1n}^{crs}$, ..., $\widehat{S}_{6n}^{crs}$, which performed best in our Monte Carlo experiments in terms of size. Since it is typically impossible to analytically derive power functions when either the null or alternative hypotheses are composite, examining the power of our tests requires additional Monte Carlo experiments. Since these experiments are computationally burdensome, we focus only on Test #1 for the case of one input and one output, and sample size $n = 20$.

To examine the power function for our test, data must be generated such that the null hypothesis of CRS is false. Let $\varphi \in (\frac{\pi}{2}, \pi]$. We simulated inputs x_i by generating iid pseudo random uniform deviates on the interval $\left(-5[1 - \tan(\frac{3\pi}{4} - \frac{\varphi}{2})]/\tan(\frac{3\pi}{4} - \frac{\varphi}{2}),\ 9\right)$. Output data (y_i) were simulated by first generating iid pseudo random standard normal deviates v_i as before, and then setting

$$y_i = \begin{cases} \left\{5\left[1 - \tan(\frac{3\pi}{4} - \frac{\varphi}{2})\right] + \tan(\frac{3\pi}{4} - \frac{\varphi}{2})x_i\right\} e^{-0.1|v_i|} & \text{if } x < 5 \\ \left\{5\left[1 - \tan(\frac{\varphi}{2} - \frac{\pi}{4})\right] + \tan(\frac{\varphi}{2} - \frac{\pi}{4})x_i\right\} e^{-0.1|v_i|} & \text{otherwise,} \end{cases} \tag{7.4}$$

for $i = 1, \ldots, n$. For $\varphi = \pi$, (7.4) reduces to $y_i = x_i e^{-0.1|v_i|}$, where the technology exhibits CRS everywhere. For $\varphi < \pi$, the technology has a kink at the point (5,5) in (x, y)-space, so that the technology and a ray from the origin tangent to the technology form angles of $\frac{\pi - \varphi}{2}$ radians on either side of the point (5,5). The technology itself forms a right angle at (5,5) when $\varphi = \frac{\pi}{2}$. By varying the value of φ in our Monte Carlo experiments, we can control the degree of departure from the null hypothesis of CRS.

Using the DGP represented by (7.4), we performed Monte Carlo experiments with $\varphi \in \{\pi - \kappa\frac{\pi}{2} \mid \kappa = 0.01, 0.02, \ldots, 0.09, 0.1, 0.2, \ldots, 0.9\}$ to estimate the power function $\mathcal{P}(\varphi)$ for each statistic $\widehat{S}_{1n}^{crs}, \ldots, \widehat{S}_{6n}^{crs}$. Each experiment consisted of $M = 1000$ Monte Carlo trials, with $B = 2000$ bootstrap replications on each of these trials. For nominal size equal to .05, $n = 20$, and $p = q = 1$, this yielded the results shown in Figure 1. These results indicate that tests based on $\widehat{S}_{2n}^{crs}$ have good power even for small departures from the null hypothesis of CRS. For instance, with $\kappa = 0.01$, $\mathcal{P}(\varphi) = 0.619$; for $\kappa = 0.02$, $(\varphi) = 0.866$, and for $\kappa = 0.03$, we find $\mathcal{P}(\varphi) = 0.913$. By contrast, for tests based on $\widehat{S}_{2n}^{crs}$, the power is much lower as indicated by Figure 1. Tests based on $\widehat{S}_{4n}^{crs}$ also have low power.

8. CONCLUSIONS

As with most Monte Carlo studies, the scope of our experiments is rather limited in terms of the number of cases considered due to the computational burden. Each experiment requires solution of $[2 \times M \times n \times (B+1)]$ LPs, where M is the number of Monte Carlo trials, with the time required for each LP increasing with the number of observations and the dimensionality of the input/output space. Nonetheless, our results provide some insight on the likely performance of nonparametric tests regarding returns to scale in small-sample applications.

In most instances, we find that the true sizes of our tests are reasonably close to the nominal sizes, or that the difference is comparable to corresponding differences that have been reported for other statistics used by econometricians in testing, for example, hypotheses regarding unit roots in time series data. Indeed, since there seem to be no other formal statistical tests of regarding returns to scale in the context of DEA models at this time, our tests seem to offer the best alternative. We find that the true sizes of our tests typically exceed the nominal sizes, which would tend to cause excessive type-I errors.

Whether this is a great problem may depend on the purpose of the tests; if the purpose is to decide which estimator of the production set to use in estimating distance functions, then our results indicate that the choice will be too often in favor of estimators that do not assume constant returns to scale. this means that distance function estimates may be less statistically efficient than they could be when the true technology is CRS. But, if the true technology is not CRS everywhere and the CRS estimator of the production set is used, distance function estimates will be statistically inconsistent. So, for this purpose, type-I errors may be less serious than type-II errors, since statistical inconsistency is a worse problem than statistical inefficiency. If, on the other hand, the researcher is trying to infer returns to scale along the technology to answer economic policy questions, then armed with the information provided by this study, careful researchers should perhaps choose a smaller nominal size than they otherwise would, particularly as the dimensionality of the input/output space increases for a given number of observations.

APPENDIX

For standard Gaussian kernel functions $K(\cdot)$ used to estimate the density of independently, normally distributed data with variance σ^2, the bandwidth

$$\widetilde{h} = (4/3)^{1/5}\sigma n^{-1/5} \tag{A.1}$$

can be shown to minimize the asymptotic mean integrated square error (MISE) of the density estimate (Silverman, 1986). Distance function estimates, however, are clearly nonnormally distributed, so the bandwidth in (A.1) will be inappropriate.

Least-squares cross validation provides an approach for minimizing MISE for a given sample $\mathcal{S}_n$. In the present setting, there are two complications. First, we construct $\widehat{f}_h(\delta)$ in (5.19) by truncating the density estimate $\widehat{g}_h(\delta)$ on the right at unity. The usual method for obtaining a computable approximation to MISE relies on the convolution theorem (see Silverman, 1986, pp. 49–50), but this requires an unrestricted domain for the density estimate. Fortunately, minimizing MISE with respect to h for $\widehat{f}_h(\delta)$ yields the same optimal value for h as minimizing MISE for $\widehat{g}_h(\delta)$. To demonstrate this, note that the MISE for $\widehat{f}_h(\delta)$ is defined by

$$\begin{aligned} \mathrm{MISE}_{\widehat{f}}(h) &\equiv E\left[\int_{-\infty}^{1}\left(\widehat{f}_h(\delta)-f(\delta)\right)^2 d\delta\right] \\ &= E\left[\int_{-\infty}^{1}\widehat{f}_h(\delta)^2\,d\delta - 2\int_{-\infty}^{1}\widehat{f}_h(\delta)f(\delta)\,d\delta + \int_{-\infty}^{1} f(\delta)^2\,d\delta\right]. \end{aligned} \tag{A.2}$$

The last term on the second line of $(A.2)$ does not depend on h; hence minimizing $\mathrm{MISE}_{\widehat{f}}(h)$ with respect to h is equivalent to minimizing

$$R_{\widehat{f}}(h) \equiv E\left[\int_{-\infty}^{1}\widehat{f}_h(\delta)^2\,d\delta - 2\int_{-\infty}^{1}\widehat{f}_h(\delta)f(\delta)\,d\delta\right]. \tag{A.3}$$

Substituting from (5.19),

$$
\begin{aligned}
R_{\widehat{f}}(h) &= E\left[\int_{-\infty}^{1} [2\widehat{g}_h(\delta)]^2 \; d\delta - 2\int_{-\infty}^{1} [2\widehat{g}_h(\delta)]\,[2g(\delta)] \; d\delta\right] \\
&= E\left[4\int_{-\infty}^{1} \left[\widehat{g}_h(\delta)^2\right] \; d\delta - 8\int_{-\infty}^{1} [\widehat{g}_h(\delta)]\,[g(\delta)] \; d\delta\right] \\
&= E\left[2\int_{-\infty}^{+\infty} \left[\widehat{g}_h(\delta)^2\right] \; d\delta - 4\int_{-\infty}^{+\infty} [\widehat{g}_h(\delta)]\,[g(\delta)] \; d\delta\right] \\
&= 2R_{\widehat{g}}(h),
\end{aligned}
\tag{A.4}
$$

where $R_{\widehat{g}}(h)$ is defined analogously to $R_{\widehat{f}}(h)$ in $(A.3)$. The second line of $(A.3)$ results from the symmetry of $\widehat{g}(\cdot)$ and $g(\cdot)$ about unity. Thus minimizing $R_{\widehat{f}}(h)$ with respect to h is equivalent to minimizing $R_{\widehat{g}}(h)$ with respect to h; either will yield the same optimal value for h. The approximation to $R_{\widehat{g}}(h)$ given by equation (3.39) of Silverman (1986) can be used; for very large n, the Fourier series idea discussed in Silverman (1986) could be used.

There is, unfortunately, a second complication. As Silverman (1986) discusses, the least-squares cross validation procedure may fail to yield a suitable bandwidth when applied to discretized data. In our setting, many of the distance function estimates will typically equal unity; if the percentage of values equal to unity is large (*i.e.*, larger than about 55 percent), then the least squares cross validation function will approach $-\infty$ as $h \to 0$.

One solution to this problem is to restrict the choice of the optimal h over some range such as $0.25\widetilde{h}$ to $1.5\widetilde{h}$, where $\widetilde{h}$ is given in (A.1). Another solution, which we adopt, is to choose h to minimize an estimator of the mean weighted integrated square error (MWISE) of $\widehat{g}(\delta)$:

$$
\begin{aligned}
\mathrm{MWISE}_{\widehat{g}}(h) &\equiv E\left\{\int w(\delta)\,[g(\delta) - \widehat{g}_h(\delta)]^2 \; d\delta\right\} \\
&= E\left[\int w(\delta)\widehat{g}_h^2(\delta)\,d\delta - 2\int w(\delta)g(\delta)\widehat{g}_h(\delta)\,d\delta + \int w(\delta)g^2(\delta)\,d\delta\right]
\end{aligned}
\tag{A.5}
$$

where $w(\delta)$ is some predefined weight function. As before, the last term on the right-hand side (RHS) of the second line of (A.5) does not depend on h, and so for purposes of choosing h, only the first two RHS terms on the second line need be considered. As shown by Silverman (1986), an unbiased estimator of these first two terms is given by

$$\mathrm{CV}(h) = \int w(\delta)\widehat{g}_h^2(\delta)\,d\delta - \frac{1}{2n}\sum_{i=1}^{2n} w(\widetilde{\delta}_i)\widehat{g}_{h,-i}(\widetilde{\delta}_i) \tag{A.6}$$

where $\widehat{g}_{h,-i}(\delta_i)$ is the leave-one out estimator of $g(\delta_i)$ based on all the original observations except δ_i, with bandwidth h.

Since the discretization problem is due to the cluster of observations at unity, we choose the weight function

$$w(\delta) = \begin{cases} 0 & \text{if } \delta \in [1-\epsilon,\ 1+\epsilon]; \\ 1 & \text{otherwise,} \end{cases} \tag{A.7}$$

where $\epsilon > 0$ is chosen to be close to zero.[12] All $2n$ (reflected) observations are used in computing the density estimate $\widehat{g}_h(\delta)$, but including the weight function defined in (A.7) in $\mathrm{CV}(h)$ means that observations $\widetilde{\delta}_i = 1$ are ignored for purposes of selecting the bandwidth.

Define Ω as the open interval $(1-\epsilon, 1+\epsilon)$ on $\mathbb{R}$. To derive a computable expression for $\mathrm{CV}(h)$, note that the first term on the RHS of (A.6) can be written as

$$\int_{-\infty}^{+\infty} w(\delta)\widehat{g}_h^2(\delta)\,d\delta = \int_{-\infty}^{+\infty} \widehat{g}_h^2(\delta)\,d\delta - \int_{\delta\in\Omega} \widehat{g}_h^2(\delta)\,d\delta. \tag{A.8}$$

The first term on the RHS of (A.8) becomes

$$\begin{aligned} \int_{-\infty}^{+\infty} \widehat{g}_h^2(\delta)\,d\delta &= \frac{1}{4n^2h^2}\int_{-\infty}^{+\infty}\sum_{i=1}^{2n}\sum_{j=1}^{2n} K\left(\frac{\delta-\widetilde{\delta}_i}{h}\right) K\left(\frac{\delta-\widetilde{\delta}_j}{h}\right) d\delta \\ &= \frac{1}{4n^2h}\sum_{i=1}^{2n}\sum_{j=1}^{2n}\int_{-\infty}^{+\infty} K\left(h^{-1}\widetilde{\delta}_i - u\right) K\left(u - h^{-1}\widetilde{\delta}_j\right) du \\ &= \frac{1}{4n^2h}\sum_{i=1}^{2n}\sum_{j=1}^{2n} K^{(2)}\left(\frac{\widetilde{\delta}_i-\widetilde{\delta}_j}{h}\right) \end{aligned} \tag{A.9}$$

where $K^{(2)}(\cdot)$ is the convolution of $K(\cdot)$ with itself; *i.e.*, if $K(\cdot)$ is $N(0,1)$, then $K^{(2)}(\cdot)$ is $N(0,2)$.

[12] The precise value of ϵ does not qualitatively affect the results, provided a value is chosen that is much smaller than $1-\max\{\delta_i \mid \delta_i \neq 1,\ i=1,\ \ldots,\ n\}$. We set $\epsilon = 10^{-6}$ in each of our Monte Carlo experiments.

For the second term on the RHS of $(A.8)$, we have

$$\int_{\delta\in\Omega} \widehat{g}^2(\delta)\,d\delta = \frac{1}{4n^2h^2}\int_{\delta\in\Omega}\sum_{i=1}^{2n}\sum_{j=1}^{2n} K\left(\frac{\delta-\widetilde{\delta}_i}{h}\right) K\left(\frac{\delta-\widetilde{\delta}_j}{h}\right) d\delta. \tag{A.10}$$

This can be computed using Gauss-Legendre formulas (*e.g.*, see Press *et al.*, 1986), and presents no special problems.[13]

Now consider the second term on the RHS of $(A.6)$. We have

$$\widehat{g}_{h,-i}(\delta) = \frac{1}{(2n-1)h}\sum_{\substack{j=1\\ j\neq i}}^{2n} K\left(\frac{\delta-\widetilde{\delta}_j}{h}\right). \tag{A.11}$$

Then

$$\frac{1}{n}\sum_{i=1}^{2n} w(\widetilde{\delta}_i)\widehat{g}_{h,-i}(\widetilde{\delta}_i) = \frac{1}{n(2n-1)h}\sum_{i=1}^{2n} w(\widetilde{\delta}_i)\sum_{\substack{j=1\\ j\neq i}}^{2n} K\left(\frac{\widetilde{\delta}_i-\widetilde{\delta}_j}{h}\right). \tag{A.12}$$

This term is also easily computed.

Substituting these results into the expression for $\mathrm{CV}(h)$ in $(A.6)$ yields

$$\begin{aligned}\mathrm{CV}(h) =& \frac{1}{4n^2h}\sum_{i=1}^{2n}\sum_{j=1}^{2n} K^{(2)}\left(\frac{\widetilde{\delta}_i-\widetilde{\delta}_j}{h}\right)\\ &- \frac{1}{4n^2h^2}\int_{\delta\in\Omega}\sum_{i=1}^{2n}\sum_{j=1}^{2n} K\left(\frac{\delta-\widetilde{\delta}_i}{h}\right) K\left(\frac{\delta-\widetilde{\delta}_j}{h}\right) d\delta\\ &- \frac{1}{n(2n-1)h}\sum_{i=1}^{2n} w(\widetilde{\delta}_i)\sum_{\substack{j=1\\ j\neq i}}^{2n} K\left(\frac{\widetilde{\delta}_i-\widetilde{\delta}_j}{h}\right).\end{aligned} \tag{A.13}$$

Note that if $\epsilon = 0$, then $w(\delta) = 1\ \forall\ \delta$, $\Omega = \emptyset$, and $\mathrm{CV}(h)$ reduces to the least-squares cross-validation function given in equation (3.39) of Silverman (1986, p. 50).

In our Monte Carlo experiments, we minimize $\mathrm{CV}(h)$ using a one-dimensional golden section search. Note, however, that due to the nature of kernel functions, the expressions comprising $\mathrm{CV}(h)$ will be easier to compute for small values of h than for large values. In a similar problem, Fan and Gijbels (1996) suggest evaluating $\mathrm{CV}(h)$ for h equal to the lower

[13] In our Monte Carlo experiments, we use 10-point Gauss-Legendre integration to solve the integral in $(A.10)$

bound of the range being considered, then successively increasing h by small increments, evaluating $\mathrm{CV}(h)$ each time. The process is terminated after some number of successive increases in $\mathrm{CV}(h)$ have been observed, and the value of h that yielded the smallest value for $\mathrm{CV}(h)$ is then chosen as the optimum.

REFERENCES

Afriat, S. (1972), Efficiency estimation of production functions, *International Economic Review* 13, 568–598.

Banker, R. D. (1993), Maximum likelihood, consistency and data envelopment analysis: A statistical foundation, *Management Science* 39, 1265–1273.

Banker, R. D. (1996), Hypothesis tests using data envelopment analysis, *Journal of Productivity Analysis* 7, 139–159.

Burley, H. T. (1980), Productive efficiency in U. S. manufacturing: a linear programming approach, *Review of Economics and Statistics* 62, 619–622.

Byrnes, P., S. Grosskopf, and K. Hayes (1986), Efficiency and ownership: Further evidence, *Review of Economics and Statistics* 68, 337–341.

Charnes, A., W. W. Cooper, and E. Rhodes (1978), Measuring the efficiency of decision making units, *European Journal of Operational Research* 2, 429–444.

Charnes, A., W. W. Cooper, and E. Rhodes (1979), Measuring the efficiency of decision making units, *European Journal of Operational Research* 3, 339.

Debreu, G. (1951), The coefficient of resource utilization, *Econometrica* 19, 273–292.

Deprins, D., L. Simar, and H. Tulkens (1984), "Measuring labor inefficiency in post offices," in *The Performance of Public Enterprises: Concepts and Measurements,* ed. by M. Marchand, P. Pestieau, and H. Tulkens, North-Holland, Amsterdam, 243–267.

Dusansky, R., and P. W. Wilson (1994), Technical efficiency in the decentralized care of the developmentally disabled, *Review of Economics and Statistics* 76, 340–345.

Dusansky, R. and P. W. Wilson, (1995), On the relative efficiency of alternative modes of producing public sector output: The case of the developmentally disabled, *European Journal of Operational Research* 80, 608–618.

Efron, B. (1979), Bootstrap methods: another look at the jackknife, *Annals of Statistics* 7, 1–16.

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

Fan, J. and I. Gijbels (1996), *Local Polynomial Modelling and Its Applications,* London: Chapman and Hall.

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

Färe, R., and S. Grosskopf (1985), A nonparametric cost approach to scale efficiency, *Journal of Economics* 87, 594–604.

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

Färe, R., and C. A. K. Lovell (1978), Measuring the technical efficiency of production, *Journal of Economic Theory* 19, 150–162.

Färe, R., S. Grosskopf, and P. Roos (1997), "Profit Productivity and Quality: A Directional Distance Function Approach," unpublished working paper, Department of Economics, Sourtern Illinois University, Carbondale, IL 62901.

Farrell, M. J. (1957), The measurement of productive efficiency, *Journal of the Royal Statistical Society A* 120, 253–281.

Ferrier, G. D. (1994), "Ownership Type, Property Rights, and Relative Efficiency," in *Data Envelopment Analysis: Theory, Methodology and Applications,* ed. by Abraham Charnes, William Cooper, Arie Y. Lewin, and Lawrence M. Seiford, Kluwer Academic Publishers, Boston, pp. 273–283.

Ferrier, G. D., and J. G. Hirschberg (1997), Bootstrapping confidence intervals for linear programming efficiency scores: With an illustration using Italian bank data, *Journal of Productivity Analysis* 8, 19–33.

Gijbels, I., E. Mammen, B. U. Park, and L. Simar (1996), "On Estimation of Monotone and Concave Frontier Functions," Discussion Paper #9611, Institut de Statistique, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Grosskopf, S. (1986), The role of the reference technology in measuring productive efficiency, *The Economic Journal* 96, 499–513.

Grosskopf, S., and V. Valdmanis (1987), Measuring hospital performance: A nonparametric approach, *Journal of Health Economics* 6, 89–107.

Kim, P. J., and R. I. Jennrich (1973), "Tables of the exact sampling distribution of the two sample Kolmogorov-Smirnov criterion D_{mn} $(m < n)$," in *Selected Tables in Mathematical Statistics, Volume 1,* ed. by H. L. Harter and D. B. Owen, American Mathematical Society, Providence, Rhode Island.

Kittelsen, S. A. C. (1997), "Monte Carlo Simulations of DEA Efficiency Measures and Hypothesis Tests," unpublished working paper, SNF Oslo, Gaustadalleen 21, N–0371 Oslo, Norway.

Kneip, A., Park, B. U., and L. Simar (1998), A note on the convergence of nonparametric DEA efficiency measures, *Econometric Theory,* forthcoming.

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, 3–18.

Lewis, P. A., A. S. Goodman, and J. M. Miller (1969), A pseudo-random number generator for the System/360, *IBM Systems Journal* 8, 136–146.

Lovell, C. A. K. (1993), "Production Frontiers and Productive Efficiency," in *The Measurement of Productive Efficiency: Techniques and Applications,* ed. by Hal Fried, C. A. Knox Lovell, and Shelton S. Schmidt, Oxford University Press, Inc., Oxford, pp. 3–67.

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling (1986), *Numerical Recipes,* Cambridge University Press, Cambridge.

Seiford, L. M. (1997), A bibliography for data envelopment analysis (1978–1996), *Annals of Operations Research* 73, 393–438.

Simar, L. and P. W. Wilson (1998a), Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models, *Management Science* 44, 49–61.

Simar, L. and P. W. Wilson (1998b), Some problems with the Ferrier/Hirschberg bootstrap idea, *Journal of Productivity Analysis,* forthcoming.

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

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

Zieschang, K. O. (1984), An extended Farrell technical efficiency measure, *Journal of Economic Theory* 33, 387–396.

TABLE 1

Monte Carlo Results for Test #1 ($H_0 : CRS$)
One Input, One Output ($p = 1,\ q = 1$)

	Nominal Size							
Statistic	.3	.25	.2	.15	.1	.05	.02	.01
$n = 20$								
$\widehat{S}_{1n}^{crs}$	0.384	0.324	0.260	0.204	0.144	0.070	0.026	0.015
$\widehat{S}_{2n}^{crs}$	0.368	0.305	0.249	0.195	0.130	0.070	0.025	0.011
$\widehat{S}_{3n}^{crs}$	0.357	0.298	0.227	0.179	0.117	0.061	0.021	0.012
$\widehat{S}_{4n}^{crs}$	0.562	0.504	0.441	0.363	0.279	0.176	0.090	0.062
$\widehat{S}_{5n}^{crs}$	0.378	0.323	0.251	0.195	0.138	0.066	0.023	0.011
$\widehat{S}_{6n}^{crs}$	0.522	0.465	0.383	0.304	0.223	0.124	0.054	0.025
binomial #1	0.504	0.492	0.458	0.361	0.291	0.203	0.101	0.045
binomial #2	0.165	0.136	0.104	0.077	0.049	0.028	0.013	0.008
$\widehat{F}_{1n}^{crs}$	0.690	0.551	0.415	0.252	0.127	0.043	0.008	0.004
$\widehat{F}_{2n}^{crs}$	0.999	0.041	0.016	0.001	0.001	0.000	0.000	0.000
$\widehat{K}_{n}^{crs}$	0.389	0.389	0.389	0.184	0.184	0.076	0.028	0.007
$n = 40$								
$\widehat{S}_{1n}^{crs}$	0.320	0.268	0.220	0.171	0.110	0.045	0.014	0.010
$\widehat{S}_{2n}^{crs}$	0.313	0.256	0.210	0.160	0.102	0.040	0.012	0.010
$\widehat{S}_{3n}^{crs}$	0.323	0.281	0.224	0.168	0.121	0.060	0.024	0.007
$\widehat{S}_{4n}^{crs}$	0.471	0.400	0.343	0.281	0.206	0.109	0.049	0.025
$\widehat{S}_{5n}^{crs}$	0.325	0.276	0.223	0.168	0.110	0.054	0.012	0.006
$\widehat{S}_{6n}^{crs}$	0.477	0.416	0.340	0.264	0.191	0.106	0.035	0.014
binomial #1	0.513	0.433	0.392	0.351	0.273	0.172	0.070	0.041
binomial #2	0.067	0.058	0.046	0.032	0.024	0.007	0.006	0.000
$\widehat{F}_{1n}^{crs}$	0.597	0.404	0.271	0.138	0.053	0.011	0.002	0.001
$\widehat{F}_{2n}^{crs}$	0.676	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$\widehat{K}_{n}^{crs}$	0.321	0.186	0.186	0.186	0.092	0.013	0.007	0.004
$n = 60$								
$\widehat{S}_{1n}^{crs}$	0.294	0.237	0.192	0.141	0.098	0.049	0.020	0.005
$\widehat{S}_{2n}^{crs}$	0.278	0.229	0.182	0.138	0.095	0.046	0.015	0.004
$\widehat{S}_{3n}^{crs}$	0.312	0.267	0.213	0.162	0.106	0.051	0.025	0.010
$\widehat{S}_{4n}^{crs}$	0.411	0.359	0.300	0.253	0.184	0.095	0.038	0.022
$\widehat{S}_{5n}^{crs}$	0.290	0.245	0.188	0.140	0.092	0.051	0.017	0.009
$\widehat{S}_{6n}^{crs}$	0.435	0.376	0.315	0.235	0.166	0.092	0.046	0.018
binomial #1	0.466	0.453	0.404	0.346	0.276	0.181	0.083	0.052
binomial #2	0.057	0.045	0.037	0.029	0.022	0.009	0.000	0.000
$\widehat{F}_{1n}^{crs}$	0.518	0.300	0.156	0.064	0.016	0.005	0.001	0.000
$\widehat{F}_{2n}^{crs}$	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$\widehat{K}_{n}^{crs}$	0.166	0.088	0.088	0.053	0.027	0.011	0.003	0.002

TABLE 2

Monte Carlo Results for Test #1 ($H_0 : CRS$)
Two Inputs, One Output ($p = 2,\ q = 1$)

Statistic	.3	.25	.2	.15	.1	.05	.02	.01
$n = 20$								
$\widehat{S}_{1n}^{crs}$	0.689	0.639	0.561	0.477	0.379	0.243	0.133	0.083
$\widehat{S}_{2n}^{crs}$	0.660	0.589	0.516	0.431	0.330	0.202	0.113	0.070
$\widehat{S}_{3n}^{crs}$	0.698	0.631	0.558	0.454	0.352	0.218	0.122	0.075
$\widehat{S}_{4n}^{crs}$	0.963	0.938	0.919	0.894	0.839	0.701	0.485	0.345
$\widehat{S}_{5n}^{crs}$	0.697	0.637	0.568	0.487	0.386	0.240	0.130	0.090
$\widehat{S}_{6n}^{crs}$	0.754	0.702	0.629	0.544	0.437	0.286	0.153	0.095
binomial #1	0.657	0.616	0.585	0.390	0.325	0.224	0.144	0.072
binomial #2	0.441	0.381	0.334	0.264	0.195	0.100	0.042	0.020
$\widehat{F}_{1n}^{crs}$	0.975	0.945	0.882	0.795	0.669	0.482	0.294	0.202
$\widehat{F}_{2n}^{crs}$	1.000	1.000	1.000	1.000	0.517	0.004	0.000	0.000
$\widehat{K}_{n}^{crs}$	0.614	0.614	0.614	0.357	0.357	0.164	0.069	0.014
$n = 40$								
$\widehat{S}_{1n}^{crs}$	0.580	0.513	0.447	0.375	0.287	0.174	0.094	0.045
$\widehat{S}_{2n}^{crs}$	0.526	0.459	0.399	0.324	0.234	0.141	0.065	0.033
$\widehat{S}_{3n}^{crs}$	0.607	0.551	0.476	0.403	0.312	0.188	0.099	0.060
$\widehat{S}_{4n}^{crs}$	0.893	0.867	0.829	0.785	0.702	0.546	0.411	0.318
$\widehat{S}_{5n}^{crs}$	0.603	0.539	0.478	0.404	0.309	0.179	0.090	0.060
$\widehat{S}_{6n}^{crs}$	0.705	0.651	0.590	0.495	0.399	0.269	0.153	0.086
binomial #1	0.732	0.614	0.580	0.549	0.490	0.367	0.193	0.154
binomial #2	0.334	0.281	0.225	0.194	0.134	0.076	0.042	0.000
$\widehat{F}_{1n}^{crs}$	0.995	0.985	0.948	0.877	0.721	0.463	0.207	0.123
$\widehat{F}_{2n}^{crs}$	1.000	1.000	1.000	1.000	0.761	0.001	0.001	0.001
$\widehat{K}_{n}^{crs}$	0.873	0.747	0.747	0.573	0.573	0.224	0.132	0.065
$n = 60$								
$\widehat{S}_{1n}^{crs}$	0.593	0.484	0.417	0.329	0.235	0.154	0.076	0.040
$\widehat{S}_{2n}^{crs}$	0.509	0.427	0.349	0.291	0.193	0.115	0.060	0.036
$\widehat{S}_{3n}^{crs}$	0.640	0.564	0.487	0.386	0.284	0.174	0.088	0.072
$\widehat{S}_{4n}^{crs}$	0.861	0.841	0.800	0.749	0.632	0.479	0.313	0.244
$\widehat{S}_{5n}^{crs}$	0.625	0.552	0.459	0.365	0.261	0.143	0.099	0.060
$\widehat{S}_{6n}^{crs}$	0.788	0.729	0.647	0.566	0.438	0.265	0.142	0.087
binomial #1	0.697	0.682	0.572	0.584	0.425	0.293	0.169	0.091
binomial #2	0.180	0.180	0.157	0.123	0.078	0.033	0.000	0.000
$\widehat{F}_{1n}^{crs}$	1.000	0.998	0.970	0.895	0.709	0.387	0.157	0.070
$\widehat{F}_{2n}^{crs}$	1.000	1.000	0.998	0.987	0.000	0.000	0.000	0.000
$\widehat{K}_{n}^{crs}$	0.902	0.784	0.784	0.644	0.500	0.351	0.142	0.888

TABLE 3

Monte Carlo Results for Test #2 ($H_0 : NIRS$)
One Input, One Output ($p = 1,\ q = 1$)

	Nominal Size							
Statistic	.3	.25	.2	.15	.1	.05	.02	.01
$n = 20$								
$\widehat{S}_{1n}^{nirs}$	0.481	0.436	0.393	0.330	0.264	0.164	0.078	0.050
$\widehat{S}_{2n}^{nirs}$	0.475	0.429	0.384	0.323	0.258	0.154	0.076	0.049
$\widehat{S}_{3n}^{nirs}$	0.990	0.976	0.943	0.879	0.741	0.506	0.261	0.136
$\widehat{S}_{4n}^{nirs}$	0.622	0.539	0.454	0.380	0.285	0.174	0.077	0.045
$\widehat{S}_{5n}^{nirs}$	0.561	0.505	0.445	0.374	0.294	0.190	0.100	0.059
$\widehat{S}_{6n}^{nirs}$	0.516	0.470	0.424	0.371	0.301	0.195	0.100	0.056
binomial #1	0.995	0.991	0.982	0.778	0.393	0.288	0.233	0.127
binomial #2	0.704	0.665	0.598	0.525	0.391	0.186	0.078	0.040
$\widehat{F}_{1n}^{nirs}$	0.221	0.142	0.083	0.040	0.012	0.000	0.000	0.000
$\widehat{F}_{2n}^{nirs}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$\widehat{K}_{n}^{nirs}$	0.006	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$n = 40$								
$\widehat{S}_{1n}^{nirs}$	0.457	0.404	0.336	0.270	0.204	0.116	0.053	0.031
$\widehat{S}_{2n}^{nirs}$	0.448	0.391	0.320	0.253	0.188	0.112	0.049	0.030
$\widehat{S}_{3n}^{nirs}$	0.997	0.985	0.958	0.916	0.836	0.641	0.421	0.290
$\widehat{S}_{4n}^{nirs}$	0.549	0.465	0.384	0.311	0.219	0.142	0.070	0.036
$\widehat{S}_{5n}^{nirs}$	0.537	0.488	0.427	0.374	0.282	0.173	0.076	0.044
$\widehat{S}_{6n}^{nirs}$	0.542	0.493	0.432	0.358	0.279	0.164	0.086	0.048
binomial #1	1.000	1.000	1.000	0.969	0.572	0.477	0.375	0.348
binomial #2	0.801	0.764	0.700	0.597	0.395	0.150	0.084	0.000
$\widehat{F}_{1n}^{nirs}$	0.142	0.071	0.027	0.005	0.001	0.000	0.000	0.000
$\widehat{F}_{2n}^{nirs}$	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$\widehat{K}_{n}^{nirs}$	0.006	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$n = 60$								
$\widehat{S}_{1n}^{nirs}$	0.403	0.348	0.285	0.236	0.185	0.108	0.057	0.033
$\widehat{S}_{2n}^{nirs}$	0.389	0.331	0.271	0.230	0.175	0.099	0.051	0.031
$\widehat{S}_{3n}^{nirs}$	0.994	0.990	0.975	0.950	0.886	0.723	0.481	0.344
$\widehat{S}_{4n}^{nirs}$	0.466	0.395	0.346	0.277	0.198	0.119	0.054	0.026
$\widehat{S}_{5n}^{nirs}$	0.472	0.424	0.373	0.307	0.236	0.144	0.075	0.040
$\widehat{S}_{6n}^{nirs}$	0.510	0.448	0.394	0.325	0.254	0.155	0.090	0.048
binomial #1	1.000	1.000	1.000	0.987	0.557	0.491	0.446	0.409
binomial #2	0.844	0.792	0.697	0.556	0.344	0.107	0.000	0.000
$\widehat{F}_{1n}^{nirs}$	0.091	0.035	0.013	0.002	0.000	0.000	0.000	0.000
$\widehat{F}_{2n}^{nirs}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$\widehat{K}_{n}^{nirs}$	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000

TABLE 4

Monte Carlo Results for Test #2 ($H_0 : NIRS$)
One Input, One Output ($p = 2,\ q = 1$)

Statistic	.3	.25	.2	.15	.1	.05	.02	.01
$n = 20$								
$\widehat{S}_{1n}^{nirs}$	0.755	0.698	0.642	0.577	0.495	0.366	0.246	0.172
$\widehat{S}_{2n}^{nirs}$	0.740	0.682	0.625	0.553	0.460	0.336	0.222	0.151
$\widehat{S}_{3n}^{nirs}$	0.919	0.843	0.721	0.559	0.397	0.256	0.183	0.127
$\widehat{S}_{4n}^{nirs}$	0.859	0.833	0.798	0.765	0.715	0.597	0.444	0.329
$\widehat{S}_{5n}^{nirs}$	0.812	0.766	0.697	0.621	0.557	0.416	0.260	0.173
$\widehat{S}_{6n}^{nirs}$	0.792	0.748	0.696	0.615	0.530	0.402	0.262	0.187
binomial #1	0.907	0.752	0.367	0.088	0.044	0.009	0.007	0.000
binomial #2	0.156	0.123	0.097	0.067	0.033	0.018	0.009	0.003
$\widehat{F}_{1n}^{nirs}$	0.731	0.623	0.527	0.411	0.294	0.160	0.082	0.053
$\widehat{F}_{2n}^{nirs}$	1.000	0.999	0.417	0.024	0.001	0.000	0.000	0.000
$\widehat{K}_{n}^{nirs}$	0.081	0.081	0.081	0.016	0.016	0.004	0.002	0.000
$n = 40$								
$\widehat{S}_{1n}^{nirs}$	0.847	0.806	0.755	0.689	0.580	0.426	0.298	0.211
$\widehat{S}_{2n}^{nirs}$	0.820	0.776	0.722	0.635	0.533	0.389	0.255	0.168
$\widehat{S}_{3n}^{nirs}$	0.984	0.959	0.910	0.808	0.669	0.547	0.466	0.380
$\widehat{S}_{4n}^{nirs}$	0.902	0.875	0.852	0.779	0.689	0.538	0.376	0.268
$\widehat{S}_{5n}^{nirs}$	0.911	0.887	0.854	0.800	0.709	0.562	0.391	0.292
$\widehat{S}_{6n}^{nirs}$	0.893	0.865	0.838	0.773	0.695	0.546	0.374	0.281
binomial #1	1.000	0.946	0.385	0.238	0.245	0.065	0.012	0.003
binomial #2	0.158	0.102	0.075	0.051	0.023	0.012	0.006	0.000
$\widehat{F}_{1n}^{nirs}$	0.830	0.715	0.582	0.409	0.256	0.093	0.021	0.008
$\widehat{F}_{2n}^{nirs}$	1.000	0.996	0.001	0.001	0.001	0.000	0.000	0.000
$\widehat{K}_{n}^{nirs}$	0.180	0.076	0.076	0.035	0.035	0.003	0.000	0.000
$n = 60$								
$\widehat{S}_{1n}^{nirs}$	0.840	0.818	0.773	0.729	0.674	0.542	0.353	0.287
$\widehat{S}_{2n}^{nirs}$	0.818	0.784	0.740	0.707	0.619	0.443	0.342	0.199
$\widehat{S}_{3n}^{nirs}$	0.994	0.994	0.960	0.872	0.784	0.707	0.674	0.619
$\widehat{S}_{4n}^{nirs}$	0.895	0.862	0.796	0.752	0.653	0.553	0.344	0.278
$\widehat{S}_{5n}^{nirs}$	0.961	0.950	0.939	0.873	0.807	0.686	0.498	0.443
$\widehat{S}_{6n}^{nirs}$	0.950	0.939	0.884	0.818	0.774	0.663	0.520	0.397
binomial #1	1.000	1.000	0.551	0.276	0.234	0.147	0.008	0.000
binomial #2	0.098	0.089	0.035	0.017	0.009	0.009	0.000	0.000
$\widehat{F}_{1n}^{nirs}$	0.902	0.775	0.617	0.437	0.238	0.076	0.016	0.003
$\widehat{F}_{2n}^{nirs}$	1.000	0.978	0.616	0.000	0.000	0.000	0.000	0.000
$\widehat{K}_{n}^{nirs}$	0.263	0.144	0.144	0.073	0.032	0.014	0.002	0.000

Columns .3 through .01 are under the spanning header "Nominal Size".

FIGURE 1

Power Functions ($\alpha = .05,\ n = 20,\ p = q = 1$)