STOCKHOLM SCHOOL
OF ECONOMICS
THE ECONOMIC RESEARCH INSTITUTE

Alternative Approaches to Estimate Returns to Scale in DEA-Models

Mickael Löthgren and Magnus Tambour

Working Paper No. 90
January 1996

Working Paper Series
in Economics and Finance

Alternative Approaches to Estimate Returns to Scale in DEA-Models

Mickael Löthgren[*] and Magnus Tambour[**]

Stockholm School of Economics

January 1996

Abstract: This paper presents a number of approaches to estimate returns to scale properties of general multiple-input multiple-output technologies. Both qualitative scale efficiency-properties and quantitative scale elasticity-properties are covered. The estimation approaches considered are based on data envelopment analysis (DEA)-models using primal production data on inputs and outputs. We propose an approach to estimate scale elasticity based on the primal efficiency problem. This alternative approach is derived as a direct approximation of the differentials of the efficiency measures in the scale elasticity definition. The different approaches to estimate scale efficiency and scale elasticity are exemplified and compared using a small dataset.

Keywords: Data Envelopment Analysis, Scale Efficiency, Scale Elasticity, Dual Linear Programming.

JEL-Classification: D24.

[*] Department of Economic Statistics, P.O. Box 6501, S-113 83 Stockholm, SWEDEN. *E-mail:* stml@hhs.se. *Phone:* +46 - 8 - 736 92 35. *Fax:* +46 - 8 - 34 81 91.

[**] Centre for Health Economics, P.O. Box 6501, S-113 83 Stockholm, SWEDEN. *E-mail:* hemt@hhs.se. *Phone:* +46 - 8 - 736 96 40. *Fax:* +46 - 8 - 30 21 15.

1. Introduction

In the economics literature two approaches to characterize scale economies are established (Färe *et al*. (1988), Forsund (1995)). The first is the neoclassical - production (cost) function based - approach (see, e.g., Frisch (1965)). The second is the axiomatic approach (see, e.g., Shephard (1953)). While the first approach generates quantitative scale measures, the second approach has mainly been devoted to give qualitative information of scale economies.

An estimation method closely related to the axiomatic approach is the non-parametric linear programming (LP) approach, dubbed as data envelopment analysis (DEA) by Charnes *et al.* (1978). In the recent works, efforts have been made to obtain measures of returns to scale that give quantitative information in addition to the well-established qualitative measures (Banker *et al.* (1994), Forsund (1995)).

In the DEA framework, three methods of obtaining qualitative information of scale economies have been proposed: the scale efficiency-method, the sum of intensity variables-method and the dual variable sign-method. These three methods involve scale measures that are defined in terms of LP specifications for Farrell-type (technical) efficiency measures. Two of the methods utilize the primal LP formulation, whereas the third is defined in terms of the dual LP formulation. Banker *et al.* (1994) established the equivalence of these measures.

Elasticity of scale is defined in terms of derivatives of distance functions, or equivalently in terms of Farrell efficiency measures. One method of estimating elasticity of scale in DEA-models has been outlined by Banker *et al.* (1984) and further developed in Banker and Thrall (1992)). This scale elasticity estimator is defined in terms of the dual efficiency LP problem, and is based on the ”slopes” of the separating hyperplanes that constitute the envelope of the data in DEA-models.

The purpose of this paper is to present various earlier established approaches to estimate scale characteristics of the technology using primal production data, as well as propose an alternative estimator of the scale elasticity based on the primal efficiency problem. Another purpose is to compare the different methods using an illustrative example.

The paper unfolds as follows: section 2 describes the production technology and alternative ways to characterize returns to scale properties. Section 3 contains a short description of DEA-models and section 4 presents established estimators of returns to scale in DEA-models. Section 5 continues with a derivation of an alternative scale elasticity estimator. The alternative approaches to estimate scale properties are compared using a small dataset in section 6. Section 7 concludes the paper with a summary.

2. Production technology and returns to scale

In this paper we consider measures of returns to scale for multiple-input, multiple-output technologies. Let $x \in R_+^N$ denote a vector of inputs used by a decision making unit (DMU) in the production of $y \in R_+^M$ outputs. In an input-based setting, where the output quantities are taken as given and inputs are the choice variables, the production technology can be described by the input requirement set, defined as

$$L(y) = \{x : x\ can\ produce\ y\}. \tag{1}$$

For the rest of the paper we assume that the technology satisfies the following set of axioms: 1) inactivity is allowed, 2) ”free lunch” is not allowed, 3) strong disposability of inputs and outputs and 4) the input requirement set is a compact and convex set. See, e.g., Färe (1988) for a thorough discussion of these axioms.

A scalar valued representation of the technology, in terms of the input requirement set, is given by the input distance function (Shephard (1953)),

$$D_i(x,y) = max\left\{\theta : \frac{x}{\theta} \in L(y)\right\}. \tag{2}$$

The input distance function is equal to the inverse of the technical efficiency measure proposed by Farrell (1957),

$$F_i(x,y) = min\{\lambda : \lambda x \in L(y)\}. \tag{3}$$

The efficiency measure takes on values less than or equal to unity for a feasible input bundle. Values equal to one indicate technical efficiency and values less than one indicates technical inefficiency. Technical efficiency is gauged relative to the isoquant of the input requirement set defined as

$$Isoq\ L(y) = \{x : x \in L(y), \mu x \notin L(y), \mu < 1\}. \tag{4}$$

Consequently, the (input) distance function takes on values greater than or equal to one, with an equivalent interpretation visa-vis the isoquant of the input requirement set.

Returns to scale

Global measures of returns to scale can be given in terms of $L(y)$ (see Färe *et al.* (1994: pp. 32)). If the technology exhibits constant returns to scale (CRS) then

$$L(\delta y) = \delta L(y), \delta > 0 \tag{5}$$

and for non decreasing returns to scale (NDRS) the following holds

$$\upsilon L(y) \subseteq L(\upsilon y), \upsilon > 1. \tag{6}$$

Finally, for non increasing returns to scale (NIRS) we have

$$L(\upsilon y) \subseteq \upsilon L(y), \upsilon > 1 \tag{7}$$

A local quantitative measure of returns to scale is given by the scale elasticity. In a single output, multiple input technology, the scale elasticity is defined as

$$\varepsilon(x) = \frac{\partial \ln f(\lambda x)}{\partial \ln \lambda}\bigg|_{\lambda=1} = \sum_{n=1}^{N} \frac{\partial f(x)}{\partial x_n} \frac{x_n}{f(x)} , \tag{8}$$

where $f(x)$ denotes a production *function*. This can be generalized to multiple-output technologies by the input or output distance function to define the scale elasticity. A multiple output, multiple input definition of scale elasticity was proposed by Panzar and Willig (1977), who also presented an input based (primal) measure of scale elasticity. They provided the following measure of elasticity of scale:

$$\varepsilon(x,y) = -\frac{\sum_{n=1}^{N} \frac{\partial \phi(x,y)}{\partial x_n} x_n}{\sum_{m=1}^{M} \frac{\partial \phi(x,y)}{\partial y_m} y_m} , \tag{9}$$

where $\phi(x,y)$ denotes a transformation function. Färe *et al.* (1988) define the transformation function as: $\phi(x,y) = D_i(y,x) - 1 = 0$. This substitution leads to a scale elasticity measure in terms of the input distance function, given by (Färe *et al.* (1986: p 178))

$$\varepsilon_i(y,x) = -\frac{D_i(x,y)}{\sum_{m=1}^{M} \frac{\partial D_i(x,y)}{\partial y_m} y_m} = -\frac{D_i(x,y)}{\nabla_y D_i(x,y) \cdot y} . \tag{10}$$

Since the input-based efficiency measure is the inverse of the distance function, which implies that $\frac{\partial D_i}{\partial y_m} = -\frac{1}{(F_i)^2} \frac{\partial F_i}{\partial y_m}$, the scale elasticity can be expressed in terms of the input-based efficiency measure as

$$\varepsilon_i(x,y) = \frac{F_i(x,y)}{\sum_{m=1}^{M} \frac{\partial F_i(x,y)}{\partial y_m} y_m} = \frac{F_i(x,y)}{\nabla_y F_i(x,y) \cdot y} . \tag{11}$$

3. Data envelopment analysis

We will restrict attention to piecewise linear technologies to analyze returns to scale for a set of decision making units (DMU). Consider *K* DMUs (observations) employing *N* inputs in the production of *M* outputs. Technical efficiency for the *K* DMUs can be estimated by solving *K* linear programming problems for each technology satisfying either CRS, NIRS or variable returns to scale (VRS). The estimated efficiency for DMU *k* when the technology is assumed to exhibit CRS, in addition to the axioms listed above, is obtained from the solution to the LP problem

$$\hat{F}_i\left(x_k, y_k | CRS\right) = \min_{\theta, z}\left\{\theta : y_k \leq Yz, \theta x_k \geq Xz, z \in R_+^K\right\}, \tag{12}$$

where Y is a (M×K) matrix of outputs, X is a (N×K) matrix of inputs and z is a K-vector of intensity variables. The same efficiency measure, but instead with the assumption of NIRS is obtained from the solution to the LP problem

$$\hat{F}_i\left(x_k, y_k | NIRS\right) = \min_{\theta, z}\left\{\theta : y_k \leq Yz, \theta x_k \geq Xz, 1'_K z \leq 1, z \in R_+^K\right\} \tag{13}$$

where 1_K is a K-vector of one's. Yet another efficiency estimate, which allows for variable returns to scale (VRS) is estimated by

$$\hat{F}_i\left(x_k, y_k | VRS\right) = \min_{\theta, z}\left\{\theta : y_k \leq Yz, \theta x_k \geq Xz, 1'_K z = 1, z \in R_+^K\right\}. \tag{14}$$

All the three efficiency measures are independent of the unit of measurement and the following nesting property also holds (see Färe *et al.* (1994: p. 73)):

$$0 < \hat{F}_i\left(x_k, y_k | CRS\right) \leq \hat{F}_i\left(x_k, y_k | NIRS\right) \leq \hat{F}_i\left(x_k, y_k | VRS\right) \leq 1\ , \tag{15}$$

with a value less than one indicating technical inefficiency relative to the isoquant of the input requirement set.

4. Returns to scale in DEA-models

This section presents reviews and presents already established estimators of scale properties in DEA-models. Two types of returns to scale measures will be considered. In section 4.1 we present measures that give qualitative information of returns to scale. That is, a DMU is determined as operating in a) an increasing returns to scale region, b) a decreasing returns to scale region, or c) the DMU is determined as being scale efficient. The next type of measures, considered in section 4.2, refers to estimators of the quantitative scale elasticity measure.

4.1. Qualitative measures

The scale efficiency method

One way to identify the nature of returns to scale for a DMU is to use the scale efficiency measure outlined in, e.g., Färe *et al.* (1994), chapter 3, for an input-based setting. A strength of this method is that scale economies can be inferred directly both for efficient as well as for inefficient DMUs. This method is not restricted to DEA estimations of the efficiency scores and can thus be employed in other settings as, e.g., parametric methods (see Forsund and Hjalmarsson (1979)). The input-based scale efficiency measure is defined as

$$S_{i1}(x,y) = \frac{F_i(x,y|\, CRS)}{F_i(x,y|\, VRS)} \tag{16}$$

Since $F_i(x,y|CRS) \leq F_i(x,y|VRS)$, S_{i1} satisfies $S_{i1} \leq 1$. S_{i1} = 1 (and $F_i(x,y|CRS) = 1$) indicates scale efficiency. A DMU with S_{i1} = 1 is scale efficient in the sense that the input-output mix is optimal and maximizes the average productivity. Furthermore, the input-output mix is equally efficient to the CRS as to the VRS technology.

Given S_{i1} < 1, the input-output mix is not scale efficient, and to determine in what region the DMU is operating in, another ratio has to be computed. This second ratio consists of the NIRS input-based efficiency measure and the equivalent CRS-measure. Thus, the ratio

$$S_{i2}(x,y) = \frac{F_i(x,y|\, CRS)}{F_i(x,y|\, NIRS)} \begin{cases} = 1 \Rightarrow IRS \\ < 1 \Rightarrow DRS \end{cases} \tag{17}$$

indicates whether the scale-inefficiency is due to a too small output (IRS) or a too large output (DRS). Following, e.g., Färe *et al.* (1994), we thus infer increasing returns to scale when $S_{i2}(x,y|S_{i1} < 1) = 1$, and decreasing returns to scale when $S_{i2}(x,y|S_{i1} < 1) < 1$.

The scale efficiency method is illustrated in *Figure 1* below.

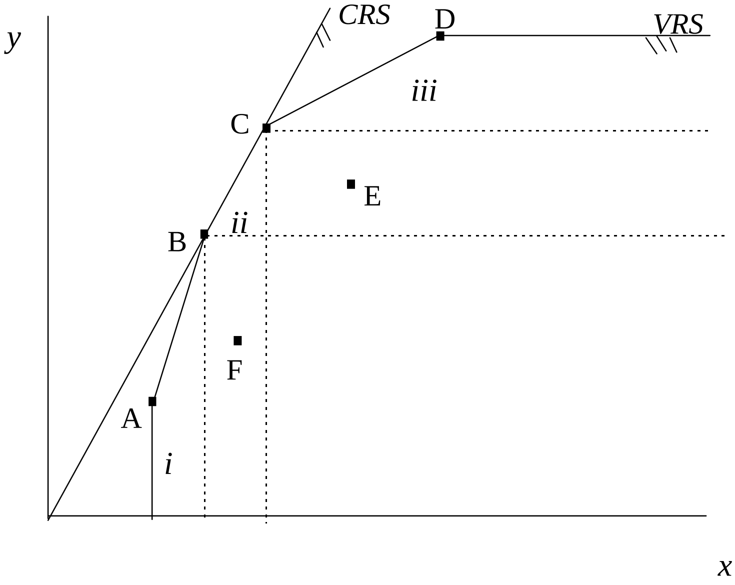

Figure 1. Single-input, single-output production technology

The figure contains input-, output- observations for six DMUs. DMUs *A*, *B*, *C* and *D* are all technically efficient relative to the VRS technology, although only *B* and *C* are scale efficient. Note that DMU *E* is also scale efficient, since it is equally inefficient to the VRS- as to the CRS technology. DMUs *A* and *F* are not scale efficient since $S_{i1} < 1$ and both are producing too small outputs, i.e. $S_{i2} = 1$. On the other hand DMU *D*, which is also scale inefficient is instead producing a too large output, i.e. $S_{i2} < 1$.

The figure also highlights the difference between choosing an input-based or an output-based scale efficiency measure. DMU *F* is not scale efficient in an input-based model, although it is scale efficient in an output-based model[1]. The opposite holds for DMU *E*, which is not scale efficient in an output-based setting. The differences occur because the input and outputs are treated differently in an input-based model compared to an output-based model. The input-based scale efficiency measure uses contracted input vectors and the scale properties are conditioned on outputs. For the output-based scale efficiency measure scale properties are given by expanded output vectors conditioned on inputs. The input-based and the output-based scale efficiency measures give the same qualitative result for observations located in regions like *i*, *ii* and *iii* in *Figure 1*. However, they will give different results for observations located in all other regions, as example for DMU *E* and *F* (c.f. Färe *et al.* (1994: pp122)).

The next two measures we consider are DEA specific returns to scale measures. The first measure is defined in terms of the (sum of) intensity variables in the primal CRS LP-problem. The second measure is defined in terms of a scalar variable in the dual VRS LP-problem.

The sum of intensity variables-method

An equivalent method to the scale efficiency method was introduced by Banker (1984). This method involves the sum of the intensity variables in the primal CRS LP-problem (12). The method was further scrutinized in Banker and Thrall (1992) and Banker *et al.* (1994). This further research was in part a response to the skepticism that was raised regarding the ability of determining returns to scale this way (see Chang and Guh (1991), Färe and Grosskopf (1994)).

According to Banker (1984) returns to scale can be determined locally by using the value of $\sum_{k=1}^{K} z_k^*$ in (12), where ”*” denotes an optimal value. The method was extended to alternative optimal solutions in Banker and Thrall (1992), leading to the following conditions:

$$\text{If } \sum_{k=1}^{K} z_k^* = 1 \text{ in any alternate optima, then CRS prevail.} \qquad (18)$$

$$\text{If } \sum_{k=1}^{K} z_k^* > 1 \text{for all alternate optima, then DRS prevail.} \qquad (19)$$

$$\text{If } \sum_{k=1}^{K} z_k^* < 1 \text{for all alternate optima, then IRS prevail.} \qquad (20)$$

[1] See Appendix 1 for output-based efficiency and scale efficiency measures.

This approach is computationally appealing since only one LP problem has to be solved, i.e., the CRS efficiency problem in (12).

Dual variable-sign-method

A third method of determining returns to scale (qualitatively) was proposed by Banker *et al.* (1984). This method exploits the dual to the VRS-efficiency formulation in (14). Returns to scale properties are characterized in terms of the sign of the variable corresponding to the intensity variables restrictions in the primal formulation . The dual LP problem to (14) is the following,

$$\hat{F}_i\left(x_k, y_k|VRS\right) = \max_{u,v,u_0}\left\{u'y_k + u_0 : v'x_k = 1, Y'u - X'v + 1_K u_0 \le 0_K, u \in R_+^M, v \in R_+^N\right\}, (21)$$

where *u* is a M-vector of virtual output prices, *v* is a N-vector of virtual input prices and u_0 is an unconstrained scalar. Banker *et al*. (1984) showed that for a radial efficient observation with a unique supporting hyperplane, u_0 indicates IRS if $u_0 > 0$, DRS if $u_0 < 0$ and finally CRS if $u_0 = 0$. In general we have several observations that are not radially efficient and in that case the LP problem in (21) cannot be used directly to infer returns to scale from the scalar variable, u_0. To obtain the relevant results for u_0 in a general setting with inefficiencies, the following LP problem should be used

$$\max_{u,v,u_0}\left\{u'y_k + u_0 : v'\hat{x}_k^f = 1, Y'u - \left(\hat{X}^f\right)'v + 1_K u_0 \le 0_K, u \in R_+^M, v \in R_+^N\right\}, \qquad (22)$$

where $\hat{X}^f$ denotes the transformed input matrix. The inputs are transformed so that $\hat{x}_k^f = \left(\hat{F}_i\left(x_k, y_k|VRS\right)x_k\right) \in Isoq\, L\left(y_k\right)\ \forall k$. This transformation is used since the reference technology, $Isoq\, L\left(y_k\right)$, is conditioned on the observed output levels (which are left untransformed) and the inputs are contracted by the estimated input-based efficiency measure.

The method is generalized to handle multiple optimal solutions in Banker and Thrall (1992). The generalization is the following: reformulate the LP problem in (22) by adding one more constraint and keep only the scalar variable in the objective function. This extension gives two LP problems to be solved to handle the multiple solutions:

$$\max_{u,v,u_0^+}\left\{u_0^+ : u'y_k + u_0^+ = 1, Y'u - \left(\hat{X}^f\right)'v + 1_K u_0^+ \le 0_K, v'\hat{x}_k^f = 1, u \in R_+^M, v \in R_+^N\right\}, \quad (23)$$

$$\min_{u,v,u_0^-}\left\{u_0^- : u'y_k + u_0^- = 1, Y'u - \left(\hat{X}^f\right)'v + 1_K u_0^- \le 0_K, v'\hat{x}_k^f = 1, u \in R_+^M, v \in R_+^N\right\}. \quad (24)$$

Using the solutions to these two LP problems, returns to scale can be determined by invoking the theorem proved in Banker and Thrall (1992):

$$\text{CRS prevails iff } u_0^+ \ge 0 \ge u_0^- \qquad (25)$$

DRS prevails iff $0 > u_0^+ \geq u_0^-$ (26)

IRS prevails iff $u_0^+ \geq u_0^- > 0$ (27)

4.2. Elasticity of scale

The hyperplane-approach

Using the dual LP formulation to the ”envelopment” problem, Banker *et al*. (1984) indicated how the separate variable, ”*u₀*”, could be used to measure elasticity of scale in DEA models. They did not explicitly show how elasticity of scale could be estimated by the scalar, u_0. The elasticity of scale estimator was (explicitly) provided in Banker and Thrall (1992), were they proposed that returns to scale could be quantitatively measured by

$$\hat{\varepsilon}_i\left(\hat{x}_k^f, y_k\right) = \frac{1}{u'y_k} = \frac{1}{1-u_0}. \qquad (28)$$

The connection between the Panzar and Willig-definition and the dual DEA-approach can be seen by inspecting the LP problem in (22) and the elasticity of scale measure in (9) and (28). Considering the *u* and *v* - vectors as output and input shadow prices, these prices will define the supporting hyperplane for DMU *k* (Banker *et al*. (1984)):

$$HP = \left\{\left(\hat{x}_k^f, y_k\right): u'y_k - v'\hat{x}_k^f + u_0 = 0\right\}. \qquad (29)$$

Let the transformation function $\phi\left(\hat{x}_k^f, y_k\right)$ be represented by the hyperplane restriction in (29), i.e., $\phi\left(\hat{x}_k^f, y_k\right) = u'y_k - v'\hat{x}_k^f + u_0 = 0$ (Forsund (1995)). This gives:

$$\nabla_{y_k}\phi\left(\hat{x}_k^f, y_k\right) = u \quad \text{and} \quad \nabla_{\hat{x}_k^f}\phi\left(\hat{x}_k^f, y_k\right) = -v\ . \qquad (30)$$

If we use (30) in (9) we obtain:

$$\hat{\varepsilon}_i\left(\hat{x}_k^f, y_k\right) = -\left(\frac{-v'\hat{x}_k^f}{u'y_k}\right) = \frac{1}{u'y_k} = \frac{1}{1-u_0}. \qquad (31)$$

The second equality holds since the sum in the nominator is restricted to be equal to unity in the LP problem (22). The expression on the far right hand side is derived from the objective function in (22). Remember that we are considering radially efficient observations, which means that the value of the objective function is also equal to one, i.e., $u'_k\, y + u_0 = 1$.

Note that (28) is ”valid” estimate of scale elasticity only for a unique optimal solution, and for radially efficient observations. In the case of multiple solutions, upper and lower bounds have to be obtained. Banker and Thrall (1992) extended the method to the cases of non-unique solutions. By using the same LP problems as in (23) and (24), they proposed the following procedure to estimate the upper and lower bounds of the scale elasticity:

Upper bound: $\hat{\varepsilon}_i^+\left(\hat{x}_k^f, y_k\right) = \frac{1}{1-u_0^+}$, where u_0^+ is obtained from (23).

Lower bound: $\hat{\varepsilon}_i^-\left(\hat{x}_k^f, y_k\right) = \frac{1}{1-u_0^-}$, where u_0^- is obtained from (24).

Note again that the upper and lower bounds $\left(u_0^+, u_0^-\right)$ are obtained for radially efficient combinations of *x* and *y*. Hence, $\hat{x}_k^f \in IsoqL\left(y_k\right)$ is the radially efficient input for DMU *k*. To obtain the bounds for all DMUs, 3·K LP problems have to be solved. In the extreme cases $u_0^+ \to 1$, which gives $\hat{\varepsilon}_i^+ \to \infty$ and $u_0^- \to -\infty$, which gives $\hat{\varepsilon}_i^- \to 0$.

5. An alternative scale elasticity estimator

This section presents an alternative approach to estimate the scale elasticity as compared to the approach presented in section 4.2 above. Based on the definition of the scale elasticity in (11) we propose an estimator of the elasticity based on differential approximations. The idea of the approximation is based on the fact that the elasticity definition can be reformulated in terms of the differential of the efficiency measure. Given this, a discrete difference is used to approximate the differential of the efficiency measure.

Considering the inputs as fixed, the total differential of the input efficiency measure is given by

$$dF_i(x,y) = \sum_{m=1}^{M} \frac{\partial F_i}{\partial y_m} dy_m = \nabla_y F_i(x,y) \cdot dy \, . \qquad (32)$$

Let $dy_m = \delta y_n$, *n=1,...,N*, for δ a small positive scalar, the differential can be expressed as

$$dF_i(x,y) = \delta \sum_{m=1}^{M} \frac{\partial F_m}{\partial y_m} y_m = \delta \nabla_y F_i(x,y) \cdot y \, . \qquad (33)$$

Hence, the inner product is given by $\nabla_y F_i(x,y) \cdot y = \frac{dF_i(x,y)}{\delta}$. Substituting this in (11) the scale elasticity measure can be expressed as

$$\varepsilon_i(x,y) = \frac{F_i(x,y)}{\nabla_y F_i(x,y) \cdot y} = \frac{\delta F_i(x,y)}{dF_i(x,y)} \, . \qquad (34)$$

An approximation of the differential dF_i can be obtained by the DEA-based difference of the efficiency measure as $dF_i(x,y) \approx \Delta\hat{F}_i(x,y) = \hat{F}_i(x,(1+\delta)y) - \hat{F}_i(x,y)$, where $\Delta\hat{F}_i(x,y)$ is the DEA-based difference using the VRS estimates of the efficiency measure in (14). We need to consider two approximations of the scale elasticity measure depending on whether increases or decreases in outputs are considered. This leaves us with two approximations of the differential dF_i :

$$d^+F_i(x,y) \approx \Delta^+\hat{F}_i\left(x,y,\delta^+\right) = \hat{F}_i\left(x,\left(1+\delta^+\right)y\right) - \hat{F}_i(x,y) \qquad (35a)$$

$$d^-F_i(x,y) \approx \Delta^-\hat{F}_i\left(x,y,\delta^-\right) = \hat{F}_i(x,y) - \hat{F}_i\left(x,\left(1-\delta^-\right)y\right)\ . \qquad (35b)$$

Note that different relative changes in the outputs are allowed for the right (δ^+) and the left (δ^-) approximations.

Given the two differential-approximations, the input based scale elasticity estimators are defined as[2]:

$$\hat{\varepsilon}_i^-\left(x,y,\delta^+\right) = \frac{\delta^+\hat{F}_i(x,y)}{\Delta^+\hat{F}_i\left(x,y,\delta^+\right)} = \frac{\delta^+}{\Delta^+\hat{F}_i\left(\hat{x}^f,y,\delta^+\right)} \qquad (36a)$$

and

$$\hat{\varepsilon}_i^+\left(x,y,\delta^-\right) = \frac{\delta^-\hat{F}_i(x,y)}{\Delta^-\hat{F}_i\left(x,y,\delta^-\right)} = \frac{\delta^-}{\Delta^-\hat{F}_i\left(\hat{x}^f,y,\delta^-\right)}\ . \qquad (36b)$$

For this elasticity of scale estimator, the extreme cases are$\Delta^+\hat{F}_i\left(x,y,\delta^+\right) \to \infty$, which implies that $\hat{\varepsilon}_i^- \to 0$ and $\Delta^-\hat{F}_i\left(x,y,\delta^-\right) \to 0$, which implies that $\hat{\varepsilon}_i^+ \to \infty$.

The equivalence of the alternative and the dual hyperplane approach

It exists a clear connection between the proposed alternative and the above presented dual-hyperplane approach to estimate scale elasticity. Using the dual formulation in (22) the differential approximations can be expressed in terms of the dual variables as

$$\Delta^+\hat{F}_i\left(\hat{x}^f,y,\delta^+\right) = u'\left(1+\delta^+\right)y + u_0 - \left(u'y+u_0\right) = \delta^+u'y \qquad (37a)$$

$$\Delta^-\hat{F}_i\left(\hat{x}^f,y,\delta^-\right) = u'y + u_0 - \left(u'\left(1-\delta^-\right)y+u_0\right) = \delta^-u'y\ . \qquad (37b)$$

Using this in the scale elasticity expressions (36a) and (36b) implies that the scale elasticity estimators can be expressed as

$$\hat{\varepsilon}_i^-\left(x,y,\delta^+\right) = \hat{\varepsilon}_i^+\left(x,y,\delta^-\right) = \frac{\delta^{\cdot}}{\delta^{\cdot}u'y} = \frac{1}{u'y} = \frac{1}{1-u_0}\ . \qquad (38)$$

[2] See Appendix 2 for a derivation of the corresponding output based scale elasticity estimators.

Hence the upper and lower bound of the scale elasticity estimate coincide. Note that this result depends on two important implicit assumptions: 1) A unique solution for the dual formulation in (22) exists, and 2) The deltas used in the differential approximations must not be ”too large”. I.e., the discrete approximations must be based on efficiency measures relative to the same, and unique, supporting hyperplane of the envelope of the data.

If these two conditions are not satisfied, then (38) is invalid. If no unique solution in (22) exists, the same argument as in the previous section on the dual hyperplane approach applies. That is, the two dual problems specified in (23) and (24) give the same lower and upper bound for the alternative approach and the hyperplane approach. I.e., the lower bound of the scale elasticity in (36a) can alternatively be expressed in terms of dual solutions as $\hat{\varepsilon}_i^- = 1/(1-u_0^-)$. The upper bound in (36b) can equivalently be expressed as $\hat{\varepsilon}_i^+ = 1/(1-u_0^+)$.

Endogenous estimators of the scale elasticity

One problem with the proposed alternative approach is that its accuracy depends crucially on the choice of the approximation parameter δ. To avoid this we describe below possible approaches in specifying endogenous estimators, obtained by introducing δ as a parameter in a formal specification of the scale elasticity optimization problem.

At first, note that the scale elasticity estimators in (36a) and (36b) are monotone functions of the approximation parameter δ. To be more precise, the lower (upper) bound is negatively (positively) monotonic in δ, as stated in the following proposition:

Proposition:

$$\hat{\varepsilon}_i^-(x,y,\delta'') \le \hat{\varepsilon}_i^-(x,y,\delta') \text{ and } \hat{\varepsilon}_i^+(x,y,\delta'') \ge \hat{\varepsilon}_i^+(x,y,\delta') \quad \forall\ \delta'' \ge \delta' \qquad (39)$$

Proof: See Appendix 3.

In Appendix 3 it is shown that $\Delta^+\hat{F}_i(\delta)$ is a piecewise linear increasing function of δ. This implies that $\hat{\varepsilon}_i^-(x,y,\delta)$ is a decreasing step function of δ. Hence, a natural definition of the endogenous delta for the lower bound estimate is given by $\delta_0^+ = \underset{\delta^+\ge 0}{arg\,max}\,\hat{\varepsilon}_i^-(x,y,\delta^+)$. In an analogous way, the endogenous delta for the upper bound estimate is defined by $\delta_0^- = \underset{\delta^-\ge 0}{arg\,min}\,\hat{\varepsilon}_i^+(x,y,\delta^-)$.

Using the endogenous deltas, the endogenous lower and upper bounds of the elasticity are defined as

$$\hat{\varepsilon}_{i*}^-(x,y) = \hat{\varepsilon}_i^-(x,y,\delta_0^+) \text{ and } \hat{\varepsilon}_{i*}^+(x,y) = \hat{\varepsilon}_i^+(x,y,\delta_0^-). \qquad (40), (41)$$

This two-step procedure of obtaining endogenous scale elasticity can be performed by a grid-search procedure, where the (36a) and (36b) are evaluated for successively smaller δ^+ and δ^-. However, this procedure can be somewhat tedious. Therefore, we propose an alternative approach where the lower and upper bounds for the elasticity are obtained directly as solutions (non-linear) optimization problems given by

$$\hat{\varepsilon}^{-}_{i*}(x_k, y_k) = \max_{\lambda^+,\delta^+,z} \left\{ \frac{\delta^+}{(\lambda^+ - 1)} : (1+\delta^+) y_k \le Yz, \lambda^+ \hat{x}^f_k \ge Xz, 1'_K z = 1, z \in R^K_+, \delta^+ \in R_+ \right\}, \quad (42)$$

and

$$\hat{\varepsilon}^{+}_{i*}(x_k, y_k) = \min_{\lambda^-,\delta^-,z} \left\{ \frac{\delta^-}{(1 - \lambda^-)} : (1-\delta^-) y_k \le Yz, \lambda^- \hat{x}^f_k \ge Xz, 1'_K z = 1, z \in R^K_+, \delta^- \in [0,1] \right\}. \quad (43)$$

A comment on the effects of an ad hoc choice of the approximation parameter δ

The ”quality” of the ad hoc approach can be established by comparing the optimal values of the endogenous deltas in (41) and (42) with the used $\delta = 0.01$. The following conclusions can be stated, based on the monotonicity property in (39), on using an ad hoc δ.

-The lower bound estimate:
If $\delta^+_0 \ge \delta = 0.01$, the ad hoc and the endogenous approach coincide, i.e., $\hat{\varepsilon}^-_i(x,y,\delta) = \hat{\varepsilon}^-_{i*}(x,y)$. If on the other hand $\delta^+_0 < \delta = 0.01$, the ad hoc and the endogenous approach differ in the sense that $\hat{\varepsilon}^-_i(x,y,\delta) < \hat{\varepsilon}^-_{i*}(x,y)$.

-The upper bound estimate:
If $\delta^-_0 \ge \delta = 0.01$, the ad hoc and the endogenous approach coincide, i.e., $\hat{\varepsilon}^+_i(x,y,\delta) = \hat{\varepsilon}^+_{i*}(x,y)$. If on the other hand $\delta^-_0 < \delta = 0.01$, the ad hoc and the endogenous approach differ in the sense that $\hat{\varepsilon}^+_i(x,y,\delta) > \hat{\varepsilon}^+_{i*}(x,y)$.

In short, the ad hoc approach can lead to a too wide interval for the elasticity. I.e.,

$$\hat{\varepsilon}^-_i(x,y,\delta) \le \hat{\varepsilon}^-_{i*}(x,y) \le \varepsilon(x,y) \le \hat{\varepsilon}^+_{i*}(x,y) \le \hat{\varepsilon}^+_i(x,y,\delta). \quad (44)$$

In section 6 a simple data example is provided that highlights different possibilities where an ad hoc approach in estimating scale elasticity will be erroneous and differ from the endogenous approach.

6. An illustrative example

To give a simple illustration of the different approaches to calculate (estimate) scale properties we use a simple data example with a single input and single output technology. The dataset is given in *Table 1* and illustrated in *Figure 2* below.

Table 1. The data example. A single input (x), single output (y) technology with observations for K = 10 DMUs.

DMU	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
x	1	1.5	2	2.5	3	4	5	5	5	5
y	1	3.5	6	7	8	9	10	9.95	8.04	5.97

This data is an extended version of the small dataset used in Banker and Thrall (1992). The first seven DMUs are the same as in Banker and Thrall. The remaining three DMUs are added to highlight the problem (see (44)) that may occur in the alternative ad hoc estimator of the scale elasticity described above in section 5. The common property of DMU 8-10 is that the frontier inputs are all located very close to kink points in the VRS envelope of the data.

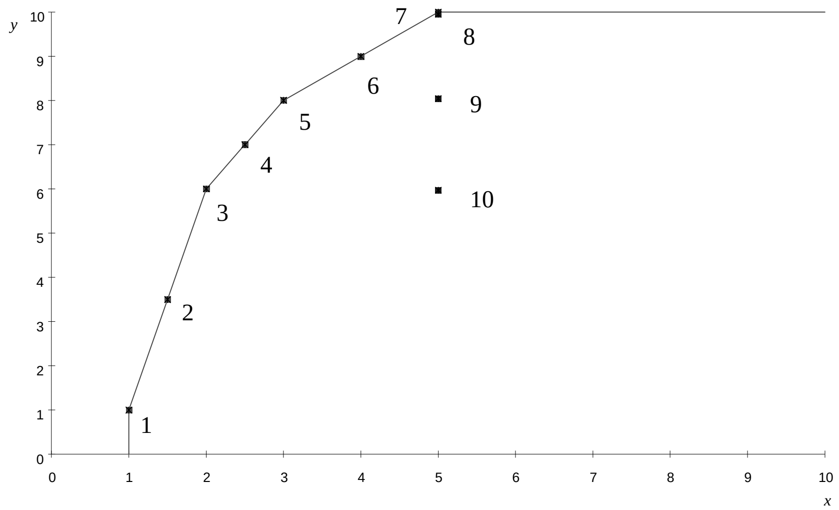

Figure 2. Illustration of the data example in *Table 1*.

In *Table 2*, the results[3] for the three qualitative methods are shown. In the first and second column the scale efficiency measures are shown. The third column is the sum of the intensity variables, followed by the values of the dual variable in column (4) - (6).

[3] All the results in this section are computed using LINGO.

Table 2. Qualitative scale measures. The scale efficiency approach, S_{i1} and S_{i2}, the sum of intensity variables approach, $\sum z$, and sign-of-u_0-approach, u_0^-, u_0, u_0^+.

DMU	S_{i1}	S_{i2}	$\sum z$	u_0^-	u_0	u_0^+
	(1)	(2)	(3)	(4)	(5)	(6)
1	0.33	1.00	0.17	0.80	1.00	1.00
2	0.78	1.00	0.58	0.53	0.53	0.53
3	1.00	1.00	1.00	–0.50	0.40	0.40
4	0.93	0.93	1.17	–0.40	–0.40	–0.40
5	0.89	0.89	1.33	–1.67	–0.33	–0.33
6	0.75	0.75	1.50	–1.25	–1.25	–1.25
7	0.67	0.67	1.67	–∞	–1.00	–1.00
8	0.67	0.67	1.66	–1.01	–1.01	–1.01
9	0.88	0.88	1.34	–1.64	–1.64	–1.64
10	0.99	1.00	0.99	0.40	0.40	0.40

The data contains one observation that is technical efficient relative to the CRS technology, i.e., observation no. 3 is scale efficient. This is identified using either of the three methods. For observation 3 we thus have $S_{i1} = \sum z = 1$ and $u_0^+ \geq 0 \geq u_0^-$. Further, there are three DMUs operating a region of IRS (nos. 1, 2 and 10) and six DMUs operating in a region of DRS (nos. 4 - 9). Again, this information is obtained from all the three methods.

Table 3 contains the dual variables and the corresponding scale elasticity estimates.

Table 3. Scale elasticity estimates using the dual hyperplane approach

DMU	u_0^-	$\frac{1}{1-u_0^-}$	u_0	$\frac{1}{1-u_0}$	u_0^+	$\frac{1}{1-u_0^+}$
	(1)	(2)	(3)	(4)	(5)	(6)
1	0.800	5.000	1.000	+∞	1.000	+∞
2	0.533	2.143	0.533	2.143	0.533	2.143
3	–0.500	0.667	0.400	1.667	0.400	1.667
4	–0.400	0.714	–0.400	0.714	–0.400	0.714
5	–1.667	0.375	–0.3333	0.750	–0.333	0.750
6	–1.250	0.444	–1.250	0.444	–1.250	0.444
7	–∞	0.000	–1.000	0.500	–1.000	0.500
8	–1.010	0.498	–1.010	0.498	–1.010	0.498
9	–1.645	0.378	–1.645	0.378	–1.645	0.378
10	0.401	1.670	0.401	1.670	0.401	1.670

The results in *Table* 3 confirms the qualitative results in *Table 2*.

In *Table 4* results for the proposed scale elasticity estimators are given both for the endogenous and the ad hoc approach. Note that the ad hoc results are based on $\delta^- = \delta^+ = \delta = 0.01$.

Table 4. Scale elasticity estimates. Ad hoc lower, $\hat{\varepsilon}_i^-$, and upper, $\hat{\varepsilon}_i^+$, estimates. Endogenous lower, $\hat{\varepsilon}_{i*}^-$, and upper, $\hat{\varepsilon}_{i*}^+$, and the corresponding endogenous delta, δ_0^+, δ_0^-.

DMU	δ_0^+	$\hat{\varepsilon}_{i*}^-$	$\hat{\varepsilon}_i^-$	δ_0^-	$\hat{\varepsilon}_{i*}^+$	$\hat{\varepsilon}_i^+$
	(1)	(2)	(3)	(4)	(5)	(6)
1	5.000	5.000	5.000	.	.	.
2	0.714	2.143	2.143	0.714	2.143	2.143
3	0.333	0.667	0.667	0.136	1.667	1.667
4	0.124	0.714	0.714	0.143	0.714	0.714
5	0.250	0.375	0.375	0.250	0.750	0.750
6	0.111	0.444	0.444	0.107	0.444	0.444
7	0.000	0.000	0.000	0.197	0.500	0.500
8	*0.005*	0.498	**0.000**	0.196	0.498	0.498
9	0.244	0.378	0.378	*0.004*	0.378	***0.505***
10	*0.005*	1.670	**0.956**	0.832	1.670	1.670

The estimates of the upper and lower bounds in the hyperplane approach coincide with the results from the endogenous delta approach for all DMUs.

From the results in *Table 4* it is seen that the ad hoc and the endogenous approach differs in the calculations for DMUs no. 8 and 10 for the lower bounds, where $\delta_0^+ < 0.01$ and hence $\hat{\varepsilon}_i^-\left(x,y \middle| \delta^+ = 0.01\right) < \hat{\varepsilon}_{i*}^-$, and for DMU no. 9 for the upper bound, where $\delta_0^- < 0.01$ and hence $\hat{\varepsilon}_i^+\left(x,y \middle| \delta'^- = 0.01\right) > \hat{\varepsilon}_{i*}^-$. This observation confirms the potential problem with the ad hoc approach stated in (44).

Both the ad hoc and the endogenous approach correctly identifies the different DRS/IRS-region for almost all cases. The only exception is the ad hoc approach where for DMU no. 10 the lower bound erroneously indicates DRS contrary to the actual IRS, relevant under the input-based approach used here. However, the fact that the upper bound is larger than one implies that a conclusion of IRS for DMU no. 10 cannot be rejected.

7. Summary and conclusions

This paper presents several approaches to define and calculate estimates of returns to scale properties for multiple-input multiple-output technologies using DEA-models. Three approaches to estimate returns qualitatively are presented: the scale efficiency-method, the sum of intensity variables-method and the dual variable sign-method. The equivalence of these methods has been established earlier. Two approaches to estimate returns to scale quantitatively are presented. The first method is the already established ”hyperplane-approach” presented in a sequence of papers by Banker and colleagues. The second method, which is the major contribution of this paper, is based on a direct approximation of the differentials of the efficiency measures in the scale elasticity definition. The connection of this method to the hyperplane approach is established.

The new approach is introduced in two steps: First a ”naive” version is considered where an approximation parameter must be specified a priori. We show that this scale elasticity estimator is a monotonic function of the approximation parameter. The approach is further developed by defining the estimator as the solution to an optimization problem, free of ad hoc specification of the approximation parameter. It is shown that the *ad hoc* approach can lead to an over- or under-estimated scale elasticity if the observation is located too ”close” to a kink point on the DEA-frontier.

The different methods are exemplified using a small dataset. As expected, the qualitative methods give the same results. For the scale elasticity, the results for the hyperplane approach and the proposed endogenous approach coincide in all cases. The example illustrates the potential problem with the ad hoc approach in the sense that the lower bound can be under estimated and the upper bound can be over estimated.

Appendix 1

The output-based efficiency and scale efficiency measures are based on the inverse of the output-based distance function

$$F_o(y,x) = max\{\lambda\colon \lambda y \in P(x)\} \qquad \text{(A1.1)}$$

The estimated efficiency under CRS for DMU *k*, is the solution to the LP problem

$$\hat{F}_o(y_k, x_k | CRS) = \max_{\theta,z}\{\theta\colon \theta y_k \leq Yz, x_k \geq Xz, z \in R_+^K\} \qquad \text{(A1.2)}$$

The same efficiency measures, but with the assumption of NIRS or VRS are obtained by varying the restrictions on the intensity variables, as in (13) and (14). The output-based efficiency measures are bounded by

$$\hat{F}_o(y_k, x_k | CRS) \geq \hat{F}_o(y_k, x_k | NIRS) \geq \hat{F}_o(y_k, x_k | VRS) \geq 1\ . \qquad \text{(A1.3)}$$

with a value greater than one indicating technical inefficiency relative to the isoquant of the output set (see Färe *et al.* (1994)). The output-based scale efficiency measure is defined as

$$S_{o1}(y,x) = \frac{F_o(y,x| \ CRS)}{F_o(y,x| \ VRS)}\ . \qquad \text{(A1.4)}$$

Since $F_o(y,x|CRS) \geq F_o(y,x|VRS)$, S_{o1} satisfies $S_{o1} \geq 1$. S_{o1} = 1 indicates scale efficiency. Given S_{o1} > 1, the input-output mix is not scale efficient and the ratio

$$S_{o2}(y,x) = \frac{F_o(y,x| \ CRS)}{F_o(y,x| \ NIRS)} \qquad \text{(A1.5)}$$

indicates increasing returns to scale when $S_{o2}(y,x|S_{o1} > 1) = 1$, and decreasing returns to scale when $S_{o2}(y,x|S_{o1} > 1) > 1$.

Appendix 2

In an output-based setting, scale elasticity can be defined in terms of Farrell-type efficiency measure as (see Färe *et al.* (1988))

$$\varepsilon_o(y,x)=\sum_{n=1}^{N}\frac{\partial F_o(y,x)}{\partial x_n}\frac{x_n}{F_o(y,x)}=\frac{\nabla_x F_o(y,x)\cdot x}{F_o(y,x)} \tag{A2.1}$$

The scale elasticity can be approximated in an output-based setting analogous to the procedure in section 5. Hence, considering the outputs as fixed and considering equal proportional changes in each input dimension $dx_n=\delta x_n$, $n=1,...,N$, i.e., $dx=\delta x$ for δ a positive scalar, the total differential of the output efficiency measure can be expressed as

$$dF_o(y,x)=\delta\sum_{n=1}^{N}\frac{\partial F_o}{\partial x_n}x_n=\delta\nabla_x F_o(y,x)\cdot x\,. \tag{A2.2}$$

Using this, the scale elasticity can be expressed in terms of the differential dF_o, as

$$\varepsilon_o(y,x)=\frac{1}{F_o(y,x)}\nabla_x F_o(y,x)\cdot x=\frac{dF_o(y,x)}{\delta F_o(y,x)}\,. \tag{A2.3}$$

The differential dF_o can be approximated as $dF_o(y,x)\approx\Delta\hat{F}_o(y,x)=\hat{F}_o(y,(1+\delta)x)-\hat{F}_o(y,x)$, where $\Delta\hat{F}_o(y,x)$ is the DEA-based difference of the efficiency measure. As in the input-based setting we need to consider two DEA-approximations of the (output-based) scale elasticity measure depending on whether increases or decreases in x are considered. That leaves us with the approximations:

$$d^+F_o(y,x)\approx\Delta^+\hat{F}_o(y,x,\delta^+)=\hat{F}_o(y,(1+\delta^+)x)-\hat{F}_o(y,x)\,, \tag{A2.4}$$

$$d^-F_o(y,x)\approx\Delta^-\hat{F}_o(y,x,\delta^-)=\hat{F}_o(y,x)-\hat{F}_o(y,(1-\delta^-)x)\,. \tag{A2.5}$$

Given these two differential-approximations, the scale elasticity estimators are given by:

$$\hat{\varepsilon}_o^-(y,x,\delta^+)=\frac{\Delta^+\hat{F}_o(y,x)}{\delta^+\hat{F}_o(y,x)}, \tag{A2.6}$$

and

$$\hat{\varepsilon}_o^+(y,x,\delta^-)=\frac{\Delta^-\hat{F}_o(y,x)}{\delta^-\hat{F}_o(y,x)}. \tag{A2.7}$$

Appendix 3

Proof of the stated monotonicity property of the scale elasticity estimator.

The proof that follows establishes the negative monotonicity of the estimator of the scale elasticity lower bound.

Proposition:

$$\hat{\varepsilon}_i^-(x,y,\delta'') \le \hat{\varepsilon}_i^-(x,y,\delta') \;\; \forall\, \delta'' \ge \delta' \qquad \text{(A3.1)}$$

This proposition is equivalent with the following property of the efficiency measure estimate

$$\frac{\Delta^+ \hat{F}_i(\delta'')}{\delta''} \ge \frac{\Delta^+ \hat{F}_i(\delta')}{\delta'} \;\; \forall\, \delta'' \ge \delta' \;. \qquad \text{(A3.2)}$$

I.e., the property is equivalent with a non decreasing ”average” efficiency difference function.

Proof:

The proof is based on the concept of supporting hyperplanes. Generally, the hyperplanes are given by $HP = \{(x,y): u'y - v'x + u_0 = 0\}$.

Consider a radially input efficient observation $\hat{x}^f \in Isoq\hat{L}(y)$. $\hat{x}^f$ belongs to a segment of the isoquant that consists of one of the supporting hyperplanes. Call this hyperplane *HP₀*. I.e., we have $(\hat{x}^f, y) \in HP_0$. Now, increase the output to $(1+\delta)y$. Denote the estimate of the input efficiency measure obtained as a rescaling of the input $\hat{x}^f$ relative to HP_0 by $\hat{F}_i^{(0)}(\hat{x}^f, (1+\delta)y) = \hat{F}_i^{(0)}(\delta)$. This efficiency estimate is implicitly obtained by the condition $(\hat{F}_i^{(0)}(\delta)\hat{x}^f, (1+\delta)y) \in HP_0$. I.e., $\hat{F}_i^{(0)}(\delta)$ is given by the solution to the equation $u'(1+\delta)y - v'\hat{F}_i^{(0)}(\delta)\hat{x}^f + u_0 = 0$. Imposing the restriction $v'\hat{x}^f = 1$ from the dual LP-formulation in (22) implies that $\hat{F}_i^{(0)}(\delta)$ is given by $\hat{F}_i^{(0)}(\delta) = u'(1+\delta)y + u_0$.

An important point to note here is that the increase in *y* may be ”too large” in the sense that HP_0 is no longer the supporting hyperplane for the isoquant $Isoq\hat{L}((1+\delta)y)$of the DEA estimate of the technology. In other words, δ may be of such a size that a ”kink” point in the enveloping hyperplanes has been ”jumped” over and the point $\hat{F}_i^{(0)}(\delta)\hat{x}^f \notin Isoq\hat{L}((1+\delta)y)$. This means that the efficiency estimate is based on a rescaling of $\hat{x}^f$ to another hyperplane that intersects HP_0. Call this alternative supporting hyperplane HP_1. The efficiency estimate obtained from a DEA-analysis for the point $(\hat{x}^f, (1+\delta)y)$is thus given by $\hat{F}_i(\hat{x}^f, (1+\delta)y) = \hat{F}_i^{(1)}(\delta)$.

Since $(\hat{F}_i^{(1)}(\delta)\hat{x}^f, (1+\delta)y) \in HP_1$, we have the following inequality

$u'(1+\delta)y - v'\hat{F}_i^{(1)}(\delta)\hat{x}^f + u_0 \leq 0$. This can be rearranged as $\hat{F}_i^{(1)}(\delta) \geq u'(1+\delta)y + u_0 = \hat{F}_i^{(0)}(\delta)$.

This result can be stated in terms of the difference $\Delta^+\hat{F}_i$ as $\Delta^+\hat{F}_i^{(1)}(\delta) \geq \Delta^+\hat{F}_i^{(0)}(\delta)$. Furthermore, note that the difference function, given that the rescaling are based on a unique hyperplane, is a linear function of δ. I.e., $\Delta^+\hat{F}_i(\delta) = \delta u'y = \beta\delta$. These results imply that the difference function is piecewise linear and positive monotonic in δ. *Figure A3.1* below illustrates a ”typical” $\Delta^+\hat{F}_i$. Note that $\Delta^+\hat{F}_i(0) = 0$.

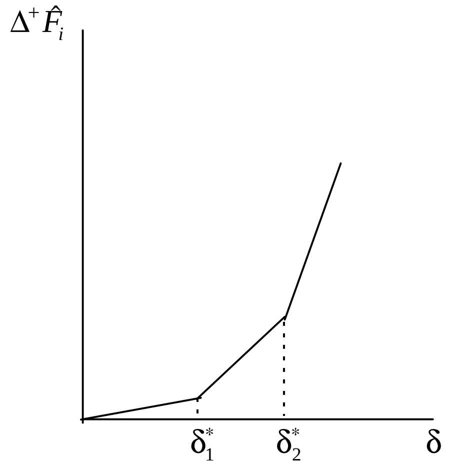

Figure A3.1. Illustration of the function $\Delta^+\hat{F}_i(\delta)$, where two kink points in the enveloping hyperplane surface is encountered for $\delta = \delta_1^*$ and $\delta = \delta_2^*$.

Note that the piecewise linear $\Delta^+\hat{F}_i(\delta)$ in a general situation with S kink points can be written as:

$$\Delta^+\hat{F}_i(\delta) = \begin{cases} \beta_1\delta, & 0 \leq \delta \leq \delta_1^*, \\ \Delta^+\hat{F}_i(\delta_1^*) + \beta_2(\delta - \delta_1^*), & \delta_1^* \leq \delta \leq \delta_2^*, \\ \vdots & \\ \Delta^+\hat{F}_i(\delta_{S-1}^*) + \beta_S(\delta - \delta_{S-1}^*), & \delta_{S-1}^* \leq \delta \leq \delta_S^*, \\ \infty, & \delta > \delta_S^*. \end{cases} \tag{A3.3}$$

Where the slope coefficients $\beta_s, s = 1,\ldots S$, satisfy the following nested inequalities $\beta_S > \beta_{S-1} > \ldots > \beta_2 > \beta_1$. Note that the last line in (A3.3) represents the limiting case for a horizontal segment of the enveloping surface.

The varying intercept terms $\Delta^+\hat{F}_i(\delta_s^*), s = 1,\ldots,S$, in (A3.3) can be written in recursive form as

$$\Delta^+ \hat{F}_i\left(\delta_s^*\right) = \sum_{j=1}^{s} \beta_j \left(\delta_j^* - \delta_{j-1}^*\right), \tag{A3.4}$$

where $\delta_0^* = 0$.

From (A3.3) it follows that the inequality in (A3.2) holds with equality if $0 \le \delta' < \delta'' \le \delta_1^*$, since

$$\frac{\Delta^+ \hat{F}_i(\delta'')}{\delta''} - \frac{\Delta^+ \hat{F}_i(\delta')}{\delta'} = \beta_1 - \beta_1 = 0\ . \tag{A3.5}$$

That is, if both δ’ and δ’’ belong to the same (first) kink segment, (A3.2) holds with equality and the estimated lower bound of the scale elasticity satisfies $\hat{\varepsilon}_i^-\left(x,y,\delta''\right) = \hat{\varepsilon}_i^-\left(x,y,\delta'\right)$. If, on the other hand $\delta_{s-1}^* \le \delta' < \delta'' \le \delta_s^*, s = 2,\dots,S$,

$$\begin{aligned}
\frac{\Delta^+ \hat{F}_i(\delta'')}{\delta''} - \frac{\Delta^+ \hat{F}_i(\delta')}{\delta'} &= \frac{(\delta''-\delta')}{\delta'\delta''}\left(\beta_s \delta_{s-1}^* - \Delta^+ \hat{F}_i\left(\delta_{s-1}^*\right)\right) \\
&= \frac{(\delta''-\delta')}{\delta'\delta''}\left(\beta_s \sum_{j=1}^{s-1}\left(\delta_j^* - \delta_{j-1}^*\right) - \sum_{j=1}^{s-1}\beta_j\left(\delta_j^* - \delta_{j-1}^*\right)\right) \\
&= \frac{(\delta''-\delta')}{\delta'\delta''}\left(\sum_{j=1}^{s-1}\left(\beta_s - \beta_j\right)\left(\delta_j^* - \delta_{j-1}^*\right)\right) > 0.
\end{aligned} \tag{A3.6}$$

So, even if δ’ and δ’’ belong to the same kink segment, s = 2, ..., S, but not the first one, (A3.2) will hold with strict inequality and $\hat{\varepsilon}_i^-\left(x,y,\delta''\right) < \hat{\varepsilon}_i^-\left(x,y,\delta'\right)$.

Finally, if $\delta_{s'}^* \le \delta' < \delta_{s-1}^* < \delta'' \le \delta_s^*, s = 2,\dots,S,\ s' = 2,\dots,s-1$, we have

$$\begin{aligned}
\frac{\Delta^+ \hat{F}_i(\delta'')}{\delta''} - \frac{\Delta^+ \hat{F}_i(\delta')}{\delta'} &= \frac{(\delta''-\delta')}{\delta'\delta''}\left(\beta_s \delta_{s-1}^* - \Delta^+ \hat{F}_i\left(\delta_{s'}^*\right)\right) \\
&= \frac{(\delta''-\delta')}{\delta'\delta''}\left(\beta_s \sum_{j=1}^{s-1}\left(\delta_j^* - \delta_{j-1}^*\right) - \sum_{j=1}^{s'}\beta_j\left(\delta_j^* - \delta_{j-1}^*\right)\right) \\
&= \frac{(\delta''-\delta')}{\delta'\delta''}\left(\sum_{j=s'+1}^{s-1}\beta_s\left(\delta_j^* - \delta_{j-1}^*\right) + \sum_{j=1}^{s'}\left(\beta_s - \beta_j\right)\left(\delta_j^* - \delta_{j-1}^*\right)\right) > 0.
\end{aligned} \tag{A3.7}$$

That is, if δ’ and δ’’ belong to different kink segments, the inequality in (A3.2) holds with strict inequality. Hence, in this case we have $\hat{\varepsilon}_i^-\left(x,y,\delta''\right) < \hat{\varepsilon}_i^-\left(x,y,\delta'\right)$.

This concludes the proof of the stated property. The positive monotonicity of the upper bound estimator can be established in an analogous way and is left to the reader.

References

Banker, R. D., (1984), ”Estimating Most Productive Scale Size Using Data Envelopment Analysis”, *European Journal of Operational Research*, Vol. 17, 35-44.

Banker, R. D., Chang, H. and Cooper, W. W., (1994), ”Equivalence of Alternative Methods for Returns-to-Scale Estimation in Data Envelopment Analysis, Unpublished working paper.

Banker, R. D., Charnes, A. and Cooper, W. W., (1984), ”Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis”, *Management Science*, Vol. 30, No. 9, 1078-1092.

Banker, R.D. and Thrall, R.M., (1992), ”Estimation of Returns to Scale using Data Envelopment Analysis”, *European Journal of Operational Research*, Vol. 62, 74-84.

Chang, K-P. and Guh, Y-Y., (1991), ”Linear Production Functions and the Data Envelopment Analysis”, *European Journal of Operational Research*, Vol. 52, 215-223.

Charnes, A., Cooper, W. W. and Rhodes, E., (1978), ”Measuring the Efficiency of Decision Making Units”, *European Journal of Operational Research*, Vol. 2, No. 6, 429-444.

Farrell, J., (1957), ”The Measurement of Productive Efficiency”, *Journal of the Royal Statistical Society*", Series A (General), Vol. 120, Part III, 253-281.

Frisch, R., (1965), ”Theory of Production”, D. Riedel Publishing Company, Dordrecht-Holland.

Färe, R., (1988), "Fundamentals of Production Theory", *Lecture Notes in Economics and Mathematical Systems*, Vol. 311, Springer-Verlag, Berlin Heidelberg.

Färe, R., Grosskopf, S. and Lovell, C. A. K., (1986), ”Scale Economies and Duality”, *Journal of Economics*, Vol. 46, No. 2, 175-182.

Färe, R., Grosskopf, S. and Lovell, C. A. K., (1988), ”Scale Elasticity and Scale Efficiency”, *Journal of Institutional and Theoretical Economics*, Vol. 144, 721-729.

Färe, R., Grosskopf, S., (1994), ”Estimation of Returns to Scale using Data Envelopment Analysis: A Comment”, *European Journal of Operational Research*, Vol. 79, 379-382.

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

Forsund, F. R., (1995), ”A Note on the Calculation of the Scale Elasticity in DEA Models”, Unpublished working paper, Department of Economics, University of Oslo.

Forsund, F. R. and Hjalmarsson, L., (1979), ”Generalised Farrell Measures of Efficiency: An Application to Milk Processing in Swedish Diary Plants”, *The Economic Journal*, Vol. 89, 294-315.

Panzar, J. C. and Willig, R. D., (1977), ”Economies of Scale in Multi-Output Production”. *Quarterly Journal of Economics*, Vol. 91, 481-493.

Shephard, R., (1953), "Cost and Production Functions", Princeton University Press, Princeton, New Jersey.

Zhu, J. and Shen, Z-H. (1995), ”A Discussion of Testing DMUs Returns to Scale”, *European Journal of Operational Research*, Vol. 81, 590-596.