Scand. J. of Economics 87 (4), 594–604, 1985

A Nonparametric Cost Approach to Scale Efficiency*

Rolf Färe and Shawna Grosskopf

Southern Illinois University, Carbondale, IL, USA

Abstract

Starting with Farrell & Fieldhouse (1962), a host of nonparametric models have been developed for calculation of scale efficiency. A common property of these models is that they are specified in terms of input/output data. Duality theory has taught us that technological characteristics may be analyzed from either the primal or the dual cost side. The purpose of this paper is to introduce a nonparametric dual method of calculating scale efficiency.

I. Introduction

Linear programming models have proved to be extremely useful in the measurement of productive efficiency. Early efforts were limited by the fairly strong restrictions imposed, implicitly and explicitly, on the reference technology. One commonly imposed restriction was that of linear homogeneity which forced the technology to exhibit constant returns to scale everywhere; see e.g. Farrell's original paper (1957). Efforts were made to circumvent the constant returns restriction by grouping data and adjusting for output level as in Farrell & Fieldhouse (1962). In the meantime, alternative linear production models were introduced which relaxed the linear homogeneity restriction allowing for increasing, constant or decreasing returns to scale. This work was pioneered by Afriat (1972) and extended by Diewert & Parkan (1983), Hanoch & Rothschild (1972) and Varian (1984). These models laid the groundwork for the recent work which focuses, in particular, on scale efficiency. Examples include Banker et al. (1984), Banker (1984), Färe et al. (1983) and Färe et al. (1985).

These studies have in common their method—nonparametric activity analysis—and their use of the production relationship or primal data as the reference set. In this framework, data on inputs and outputs are required to

* We would like to thank two anonymous referees for their comments.

calculate scale efficiency. From duality theory, we know that properties of the technology may be analyzed from the primal (input/output) side or deduced from the dual or cost side. In fact, many of the studies which calculate scale elasticities using stochastic models employ the cost framework, due in many cases to data availability. The purpose of this paper is to provide the option of employing cost rather than input data to those pursuing nonparametric methods. In particular, we introduce a nonparametric method of calculating scale efficiency using dual or cost data.

In terms of organization, we begin by summarizing the primal approach to nonparametric measurement of scale efficiency based on Färe et al. (1985). Next we introduce the cost approach. We conclude with a comparison of the two approaches.

II. The Primal Approach

In order to review the Färe et al. (1985) primal approach to calculating scale efficiency, suppose there are k establishments each producing a vector of outputs $u \in R^m_+$ using n inputs $x \in R^n_+$. We denote the matrix of observed outputs by U, where U is of order (k, m) and we denote the matrix of observed inputs by X, where X is of order (k, n). The strategy in calculating scale efficiency and determining its source is to construct a series of nonparametric frontier technologies. We then use linear programming techniques to calculate the efficiency of each observation relative to the different frontier technologies.

We start with the most restrictive frontier technology which exhibits constant returns to scale (CRS)[1] and strong disposability of inputs and outputs.[2] This technology corresponds to that originally used by Farrell (1957). The overall input[3] measure of technical efficiency of any observation (u, x) relative to this technology is calculated by solving

$$K_i(u, x) = \min \lambda \tag{1}$$

subject to

$$zU \geqslant u$$
$$zX \leqslant \lambda x$$
$$z \in R^k_+,$$

[1] A technology $T=\{(x, u)\colon u \text{ obtainable by } x \in R^n_+\}$ exhibits CRS if it is a cone.

[2] A technology T exhibits strong disposability of inputs if $(x, u) \in T$ and $y \geqslant x \Rightarrow (y, u) \in T$. T exhibits strong disposability of outputs if $(x, u) \in T$ and $v \leqslant u \Rightarrow (x, v) \in T$.

[3] These measures could also be derived in terms of outputs, where the corresponding measures $K_0(x, u)$ would yield the ratio of ray maximum to actual output, given observed input levels.

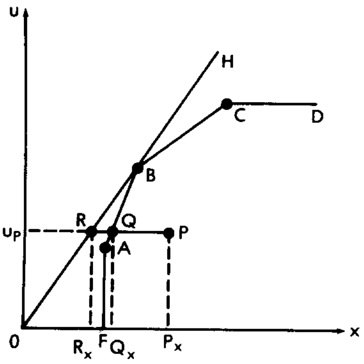

Fig. 1

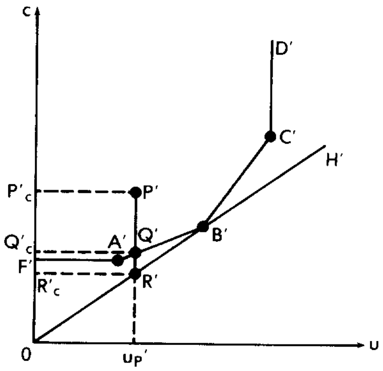

Fig.2

where $z=(z_1,z_2,...,z_k)$ is the intensity vector and serves to construct the smallest cone which covers the data in the sample.

In Figure 1, four observations are labelled A, B, C, D. The reference technology formed by the constraints in (1) is given by the ray OH and the x-axis. The measure $K_i(u,x)$, applied to observation P, is given by the ratio OR_x/OP_x, i.e., the ratio of minimum to actual input usage given output level u_p.

Next, we construct a less restrictive technology which satisfies nonincreasing returns to scale (NIRS)[4] and strong disposability of inputs and outputs. The measure of technical efficiency of any observation (u,x) relative to this technology is calculated by solving

$$W_i^*(u,x)=\min\lambda \tag{2}$$

subject to

$$zU\geqslant u$$

$$zX\leqslant \lambda x$$

$$z\in R_+^k$$

$$\sum_{i=1}^{k} z_i\leqslant 1.$$

Problem (2) differs from (1) only in the additional restriction placed on the intensity vector in (2). In the scalar output case, as shown in Afriat (1972), this restriction imposes concavity of the production function. In terms of the programming problem, the reference technology is constructed from the star closure of observed points, i.e., the original observed points and their radial contractions.

[4] A technology T exhibits NIRS if $(x,u)\in T\Rightarrow(\lambda x,\lambda u)\in T$, $0\leqslant\lambda\leqslant 1$.

In Figure 1 the reference technology given by the constraints in (2) is bounded by OBCD and the x-axis. Relative to this technology, observations B and C are efficient.

The least restrictive reference technology which satisfies variable returns to scale (VRS) and strong disposability of inputs and outputs is now used to calculate technical efficiency. This efficiency measure is obtained by solving

$$W_i(u,x) = \min \lambda \tag{3}$$

subject to

$$zU \geqslant u$$
$$zX \leqslant \lambda x$$
$$z \in R^k_+$$
$$\sum_{i=1}^{k} z_i = 1.$$

Expression (3) differs from (1) and (2) in terms of the restriction on the intensity vector. By restricting the intensity vector to sum to unity, the programming problem constructs a closed polytype. The inequality constraints allow for disposability. In the scalar output case, it is shown by Afriat (1972) that the $\Sigma z_i = 1$ restricts the production function to be quasi-concave, allowing for variable returns to scale (VRS).

In Figure 1, reference technology (3) is bounded by FABCD and the x-axis starting at F. In this technology, A, B and C are efficient, while applying the measure to observation P yields $W_i(u,x)=OQ_x/OP_x$. We can now derive the Färe et al. (1985) input measure of scale efficiency, which measures proportional input slack due to deviation from optimal scale, i.e., from CRS. The measure of scale efficiency is given by

$$S_i(u,x) = K_i(u,x)/W_i(u,x). \tag{4}$$

In terms of Figure 1, $S_i(u,x)$ for observation P is thus OR_x/OQ_x. From the earlier discussion it is clear that $1 \geqslant S_i(u,x) > 0$. $S_i(u,x)=1$, i.e., is called scale efficient, if and only if $(x \cdot W_i(u,x))$ belongs to the CRS technology given by the constraints in (1). If $S_i(u,x)<1$, then scale inefficiency is due to decreasing returns to scale (DRS)[5] if and only if $W_i^*(u,x)=W_i(u,x)$, i.e., $(x \cdot W_i(u,x))$ belongs to the NIRS technology given by the constraints in (2).

[5] A technology T exhibits DRS at $(x,u) \in T$ if (x,u) belongs to a technology that exhibits NIRS, but not to a technology that exhibits CRS.

Otherwise scale inefficiency is due to increasing returns to scale (IRS).[6] Note that Q exhibits IRS.

We note that this measure is closely related to the (input) scale efficiency measure developed by Banker et al. (1984) and Banker (1984), although they use other methods to determine whether DRS, CRS or IRS prevail locally. Our measure is also conceptually similar to (some of) the scale measures discussed in Førsund & Hjalmarsson (1979). These measures have in common their use of the "smallest" CRS technology as the reference "scale efficient" set. This corresponds to the notion of the social efficiency of the competitive firm producing at the minimum point of the average cost curve in the long run.

III. The Cost Approach

In the primal approach discussed above, observations on inputs and outputs were used to calculate scale efficiency. For the cost approach, suppose once more that observations on outputs are given for each establishment, i.e., U is known. Moreover, assume that the total costs of producing outputs for each establishment are given. In order to maintain consistency with the primal approach, we require that each establishment face the same input price vector. Denote the vector of these costs by C, where C is of order $(k, 1)$. In order to calculate scale efficiency using U and C, we proceed as in Section II by constructing a series of nonparametric cost frontiers, and use linear programming techniques to calculate the cost efficiency of each observation relative to the different cost frontiers.

Let us start with the most restrictive cost frontier exhibiting CRS and strong disposability of outputs (and cost). The overall cost efficiency measure of some observation (u, c) relative to this frontier is calculated by solving

$$K_c(u, c) = \min \lambda \tag{5}$$

subject to

$$zU \geqslant u$$

$$zC \leqslant \lambda c$$

$$z \in R^k_+,$$

[6] A technology T exhibits Non-Decreasing Returns to Scale, NDRS, if $(x, u) \in T \Rightarrow (\lambda x, \lambda u) \in T$, $\lambda \geqslant 1$. T exhibits IRS at $(x, u) \in T$ if (x, u) belongs to a technology that exhibits NDRS, but not to a technology that exhibits CRS.

where again $z \in R^k_+$ is the intensity vector. In Figure 2, four observations are labelled A', B', C', D'.[7] The reference frontier formed by the constraints in (5) is given by the ray $O'H'$ and the C-axis. Only observation B' is efficient relative to this frontier. The cost measure $K_c(u, c)$ applied to observation P' is given by the ratio OR'_c/OP'_c.

Since C is of order $(k, 1)$ and z is of order $(1, k)$, zC is a scalar, so that $K_c(u, c)$ may also be calculated using a two step procedure. First calculate

$$\hat{K}_c(u) = \min zC \tag{6}$$

subject to

$$zU \geqslant u$$
$$z \in R^k_+.$$

From (5), $\lambda = zC/c$ where zC constructs the minimum cost frontier, which is the same as $\hat{K}_c(u)$. Thus we can now calculate: $K_c(u, c) = \hat{K}_c(u)/c$, where $\hat{K}_c(u)$ is minimum cost and c is observed cost of the same observation.

Next, we construct the less restrictive cost frontier satisfying NIRS and strong disposability of outputs (and cost). The measure of cost efficiency for some observation (u, c) relative to this frontier is obtained by solving

$$W^*_c(u, c) = \min \lambda \tag{7}$$

subject to

$$zU \geqslant u$$
$$zC \leqslant \lambda c$$
$$z \in R^k_+$$
$$\sum_{i=1}^{k} z_i \leqslant 1.$$

In Figure 2 the frontier obtained from (7) is bounded by $O'B'C'D'$ and the c-axis. Under this frontier, observations B' and C' are efficient.

The least restrictive reference frontier (satisfying VRS and strong disposability of outputs and cost) is next used in the calculation of cost efficiency.

$$W_c(u, c) = \min \lambda \tag{8}$$

[7] Note that if input price is assumed to be equal to one, Figure 2 is the mirror or dual image of Figure 1 with the axes interchanged.

subject to

$$zU \geqslant u$$
$$zC \leqslant \lambda c$$
$$z \in R^k_+$$
$$\sum_{i=1}^{k} z_i = 1.$$

In Figure 2, the restrictions imposed by (8) are given by $F'A'B'C'D'$ and the c-axis above F'. Here, observations A', B' and C' are all efficient. Applying $W_c(u, c)$ to observation P' implies that $W_c(u, c) = OQ_c'/OP_c'$.

The cost measure of scale efficiency may now be defined analogous to the primal scale measure in (4) as

$$S_c(u, c) = K_c(u, c)/W_c(u, c). \tag{9}$$

In terms of Figure 2, the dual scale efficiency of observation P' is thus OR_c'/OQ_c. Clearly, $1 \geqslant S_c(u, c) > 0$, $S_c(u, c) = 1$, i.e., an observation is called scale efficient if and only if $(c \cdot W_c(u, c))$ belongs to the CRS frontier given by the constraints in (5). In an approach analogous to Section II, if $S_c(u, c) < 1$, we can determine the source of inefficiency. If $S_c(u, c) < 1$ and $W_c^*(u, c) = W_c(u, c)$, scale inefficiency is due to DRS, otherwise it is due to IRS. We observe that scale inefficiency for Q' is due to IRS.

IV. Comparison of the Approaches

In order to show conditions under which the two measures $S_i(u, x)$ and $S_c(u, c)$ of scale efficiency are equivalent, we first introduce a third intermediate method. Thus assume that in addition to X and U, input prices are known for each establishment. Let $p \in R^n_+$ denote input prices and compute

$$C(u, p) = \min px \tag{10}$$

subject to

$$zU \geqslant u$$
$$zX \leqslant x$$
$$z \in R^k_+,$$

where x is the choice variable, and $C(u, p)$ is the minimal cost relative to the reference technology in (1). Since both p and x are given for each establishment, the observed total cost c is also known. Following Farrell (1957), the overall measure of efficiency is

$$C(u,p)/c, \tag{11}$$

i.e., minimum potential cost divided by observed cost. The overall measure of efficiency can be decomposed into overall technical (*TE*) and allocative efficiency, *AE*. Thus,

$$C(u,p)/c=AE\cdot TE.^{8} \tag{12}$$

The technical efficiency referred to in (12) is equivalent to $K_i(u,x)$. Thus we can rewrite the Farrell decomposition as

$$C(u,p)/c=AE\cdot K_i(u,x). \tag{13}$$

All of the measures in (13) are calculated relative to the same reference technology, namely a technology satisfying CRS and strong disposability as in (1) and (10). We can derive a similar decomposition for measures calculated relative to a technology which satisfies VRS and strong disposability as in (3) and (14). For simplicity, we denote measures calculated relative to this alternate VRS technology by a V superscript. For example, we can calculate the cost minimum relative to the VRS technology as

$$C^V(u,p)=\min px \tag{14}$$

subject to

$$zU\geqslant u$$

$$zX\leqslant x$$

$$z\in R^k_+$$

$$\Sigma z_i=1,$$

where again the restriction $\Sigma z_i=1$ allows for VRS rather than CRS. We can now write a Farrell decomposition for $C^V(u,p)/c$ as

$$C^V(u,p)/c=AE^V\cdot W_i(u,x), \tag{15}$$

where $W_i(u,x)$ is technical efficiency (*TE*) calculated relative to the VRS reference technology; see (3).

Dividing (13) by (15) we obtain

$$[K_i(u,x)/W_i(u,x)]\cdot[AE/AE^V]=C(u,p)/C^V(u,p). \tag{16}$$

[8] *AE* is defined by the identity (12), i.e., $AE=C(u,p)/(c\cdot TE)$. Whereas *TE* is independent of input prices, *AE* does depend on input prices.

If $AE = AE^V$, then this simplifies to

$$S_i(u,x) \equiv K_i(u,x)/W_i(u,x) = C(u,p)/C^V(\mathrm{u},\mathrm{p}). \tag{17}$$

Expression (17) shows that if allocative efficiencies for the two reference technologies are equal, scale efficiency may be calculated as the quotient of the two minimum costs $C(u,p)$ and $C^V(u,p)$. However, this method requires information on inputs, outputs and input prices.

On the other hand, our dual scale efficiency measure $S_c(u,c)$ required data on outputs and costs—information on inputs and input prices were not required. In fact, our dual scale measure can be shown to be equivalent to the minimum cost measures discussed above. This relationship then allows us to derive the relationship between $S_i(u,x)$ and $S_c(u,c)$.

In order to establish the relationship between the minimum cost measures and $S_c(u,c)$, we begin by returning to the simple cost minimization problem used to calculate $C(u,p)$ in (10). Again assuming that all establishments face the same given input prices, we can write:

$$C(u,p) = \min px \tag{18}$$

subject to

$$zU \geqslant u$$

$$zX \leqslant x$$

$$\mathrm{z} \in R^k_+.$$

Since, by assumption, each firm faces the same input prices, the constraints $zX \leqslant x$ can be multiplied by p. Since we are minimizing px, it can be summed across the different inputs for each establishment, thus (18) becomes

$$\min px \tag{19}$$

subject to

$$zU \geqslant u$$

$$zC \leqslant px$$

$$\mathrm{z} \in R^k_+,$$

where C is the vector of observed costs. Equivalently, we can calculate minimum costs directly

$$\min zC \tag{20}$$

subject to

$zU \geqslant u$

$z \in R^k_+$,

which is precisely how we calculated $\hat{K}_c(u)$ in (6), thus $\hat{K}_c(u)=C(u,p)$. Recalling that $K_c(u,c)=\hat{K}_c(u)/c$, it follows directly that $K_c(u,c)=C(u,p)/c$. Following this method it also follows that $W_c(u,c)=C^V(u,p)/c$, which implies that

$$S_c(u,c) \equiv K_c(u,c)/W_c(u,c) = C(u,p)/C^V(u,p). \tag{21}$$

Substituting into (16) we now have

$$S_i(u,x) \cdot [AE/AE^V] = S_c(u,c). \tag{22}$$

Thus, if $AE=AE^V$, then $S_i(u,x)=S_c(u,c)$. Or, more generally, $AE=AE^V \Leftrightarrow S_i(u,x)=S_c(u,c)$, given that each firm faces the same input prices.

In summary, we have derived an alternate programming method of calculating scale efficiency which uses cost or dual information rather than production or primal data. Using the Farrell decomposition of overall efficiency, we have shown that the two approaches yield the same efficiency rating if $AE=AE^V$, whenever all firms face the same input prices.

This alternate dual method affords the researcher a wider range of choice in terms of data requirements when calculating scale efficiency. Since $K_c(u,c)/c=C(u,p)/c$, this approach can also be used to calculate overall efficiency when information on inputs and their prices is not known, as long as all establishments face the same input prices.

References

Afriat, S.: Efficiency estimation of production functions. *International Economics Review 13,* 568–598, 1972.

Banker, R. D.: Estimating most productive scale size using data envelopment analysis. *European Journal of Operational Research 17,* 35–44, 1984.

Banker, R. D., Charnes, A. & Cooper, W. W.: Some models for estimating technical and returns to scale inefficiencies. *Management Science 30,* 1078–1092, 1984.

Diewert, E. & Parkan, C.: Linear programming tests of regulatory conditions for production functions. *Quantitative studies on production and prices* (ed. by W. Eichhorn, R. Henn, K. Neumann and R. W. Shephard), pp. 131–158, Physica Verlag, Würzburg, West Germany, 1983.

Färe, R., Grosskopf, S. & Lovell, C. A. K.: The structure of technical efficiency. *The Scandinavian Journal of Economics 85,* 181–190, 1983.

Färe, R., Grosskopf, S. & Lovell, C. A. K.: *The measurement of efficiency of production.* Kluwer-Nijhoff, Boston, 1985.

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

Farrell, M. J. & Fieldhouse, M.: Estimating efficient production functions under increasing returns to scale. *Journal of the Royal Statistical Society 125,* Series A, part 2, 252–267, 1962.

Førsund, F. & Hjalmarsson, L.: Generalized Farrell measures of efficiency: An application to milk processing in Swedish dairy plants. *Economic Journal 89*, 274–315, 1979.

Hanoch, G. & Rothschild, M.: Testing the assumptions of production theory: A nonparametric approach. *Journal of Political Economy 80*, 256–275, 1972.

Varian, H.: The nonparametric approach to production analysis. *Econometrica 52*, 579–599, 1984.

First version submitted February 1985;
final version received June 1985.