Estimation Methodology of Economic Efficiency: Stochastic Frontier Analysis vs Data Envelopment Analysis

Taptuk Emre ERKOC  PhD Candidate in Economics, Centre for Economic and Management, Keele University, UK
t.erkoc@ippm.keele.ac.uk

ABSTRACT

This chapter investigates estimation methodology of economic efficiency on the basis of both mathematical programming and econometric techniques. Section II defines and illustrates the concept of efficiency regarding microeconomic framework. Section III examines Data Envelopment Analysis (DEA) which is the major linear programming model in efficiency analysis. Section IV deals with econometric techniques in particular Stochastic Frontier Analysis (SFA) that is the mostly used model among parametric techniques. Section V compares the merits and weaknesses of these aforementioned methodologies and concludes.

KEY WORDS  Economic Efficiency, Stochastic Frontier Analysis, Data Envelopment Analysis

JEL CODES  D01, D11, D12

I. Introduction

In microeconomic theory, the objective of firms is identified as producing maximum output using given inputs with minimum cost. In other words, it can be defined as utilising minimum amount of one input within a given output and other input levels. This microeconomic notion stipulates or presumes that firms –within the framework of free market rules- should allocate input and output efficiently with the aim of obtaining maximum profit and/or minimum cost. Up to now, productive efficiency of a firm has been calculated by means of measuring the distance to a particular frontier such as the revenue frontier, cost frontier and production frontier.

II. Definition of Efficiency

In market economies in which markets exercise power on the behaviours of firms and individuals, they are expected to achieve the theoretical maximum either in production and/or consumption. The failure of firms to produce at the “best-practicing” frontier which can be called as production inefficiency has been elaborated by researchers (Hicks: 1935, Debreu: 1951, Farrell: 1957, Leibenstein: 1966) on the basis of different approaches. Hicks (1935) argued that monopolistic firms don’t feel any market restraint on them to become fully
efficient as enjoying benefits of monopoly. In a similar vein, Debreu (1951) and Farrell (1957) proposed that lack of market power on managers in certain cases may cause inefficiencies among the firms.

The most controversial argument in explaining the inefficiencies of firms is Leibenstein’s X-inefficiency approach which contradicts with neo-classical microeconomics theory. To Leibenstein (1966), the failure of firms to produce on the efficient frontier is by and large motivated by following set of reasons including inadequate motivation, incomplete contracts, asymmetric information, agency problems and attendant monitoring difficulties which are lumped together and form X-inefficiency. Stigler (1976) objected to this approach and put forward that all sources of inefficiency according to Leibenstein can be shown as the evidence for incomplete production model in which whole set of relevant variables are failed to be incorporated (Fried et al: 2008, 9).

The pioneering work of Koopmans (1951) provided the earliest formal definition of technical efficiency as: “A producer is technically efficient if, and only if, it is impossible to produce more of any output without producing less of some other output or using more of some input.” Subsequently, Debreu (1951) and Farrell (1957) developed a slightly different definition of technical efficiency by ruling out the slack units: “one minus the maximum equiproportionate (radial) reduction in all inputs that is feasible with given technology and output” (Fried et al: 2008, 20).

To be able to examine those aforementioned means of measurement, it might be appropriate to introduce some certain notations and formulations:

\[ \text{Level Set: } L(y) = \{ x: (y, x) \text{ is producible } \} \]

(3.1)

the production function is derived from input isoquant function to produce \( y \)

\[ I(y) = \{ x: x \in L(y), \lambda x \in L(y) \text{ if } \lambda < 1 \} \]

(3.2)

and the efficient input subset is defined as:

\[ ES(y) = \{ x: x \in L(y), x' \in L(y), x' < x \} \]

(3.3)

eventually interrelation between these three subsets can be represented as:

\[ ES(y) \subseteq I(y) \subseteq L(y) \]

(3.4)

and depicted in Figure-1:

![Figure-1](image-url)
The Debreu-Farrell input-oriented technical efficiency can be formulated relying on the production function:

$$ TE_{DF}(y, x) = \min \{ \varnothing : \exists x \in L(y) \} $$

(3.5)

Shephard’s (1953) input distance function is another apparatus that has been used to figure out the technical efficiency of firms from a relatively different perspective. Shephard formulated the distance function (based on input measurements) as indicated below:

$$ D_S(y, x) = \max \{ \lambda: [1/\lambda] x \in L(y) \} $$

(3.6)

It is obvious that, Debreu-Farrell radial contradiction process of inputs is the inverse iteration of Shephard’s input distance function. Therefore, (3.5) and (3.6) can be related to each other as:

$$ TE_{DF}(y, x) = 1/D_S(y, x) $$

(3.7)

To Debreu-Farrell, the first and foremost requirement of being technically efficient is to be situated exactly on the isoquant curve I(y). However, Koopmans stipulates the “absence of coordinatewise improvements” which means “a simultaneous membership in both efficient subsets (Fried et al: 2008, 25).” For instance, while the point $X^a$ on Figure-1 is technically efficient according to the Debreu-Farrell definition, Koopmans spots this point -which is outside the efficient subset- as inefficient due to slack usage of $X_2$. As a consequence, it is convenient to state that “Debreu-Farrell technical efficiency is necessary, but not sufficient for Koopmans technical efficiency” (Kang: 1998, 63).

In efficiency analysis, two components have been put forward by Farrell (1957) as fundamentals of efficiency comprising of technical (TE) and allocative (AE). Whilst the former one arises when outputs fall short from ideal production given input level, the latter is the result of inappropriate input choices concerning certain input prices and output level. As indicated in Figure-2, producer utilises two inputs ($X_1$ and $X_2$) in order to produce a specific output. At the input bundle of $X^a$, this producer has the capability to decrease the amount of inputs all the points in “level set” back to isoquant curve until reaching to the point $\theta X^a$. That is to say, the input choices at $X^a$ can be radially contradicted with the “absence of...
coordinatewise improvements” up to the point $\theta X^A$. Therefore, relying on both Koopmans and Debreu-Farrell definitions, technical efficiency of this firm at the point $X^A$ is calculated as:

$$TE = \frac{\partial \theta X^A}{\partial X^A}$$

(3.8)

where $X^A$ denotes the observed input levels and $\theta X^A$ represents the combination of technically efficient amounts of inputs.

To have an economically efficient production set, TE is not sufficient alone. The input combination should be selected appropriately on the basis of their prices. The best-practicing mixture of inputs concerning the prices is the intersection point of isoquant and isocost curves where technically feasible production units are produced at the lowest cost. According to the Figure-2, allocative efficiency at $X^A$ is:

$$AE = \frac{\alpha X^A}{\partial \theta X^A}$$

(3.9)

where $\theta X^A$ represents the combination of technically efficient amounts of inputs, $\alpha X^A$ refers to the mixture of inputs that has the lowest cost given this output and technology.

In order to convert production efficiency to cost efficiency (particularly for the multiple output cases), assume that producer faces input prices $w = \{w_1, w_2, ..., w_n\}$ and aims to minimise costs. For this case, cost frontier can be narrated as:

$$c \left( y, w \right) = \min_x \left\{ w^T x : D_x \left( y, x \right) \geq 1 \right\}$$

(3.10)

if the inputs are freely disposable and the level sets $L(y)$ are convex and closed, the cost frontier above is the dual function of input distance function proposed by Shephard (1953). Therefore:

$$D_x \left( y, x \right) = \min_w \left\{ w^T x : c \left( y, w \right) \geq 1 \right\}$$

(3.11)

cost efficiency can be calculated as the ratio of minimum cost to actual cost:

$$CE \left( y, x, w \right) = \frac{c \left( y, w \right)}{w^T x}$$

(3.12)

regarding to the points shown in Figure-2, cost efficiency at $X^A$ is:

$$CE = \frac{\partial \alpha X^A}{\partial X^A}$$

(3.13)

As being easily inferred from Figure-2, cost-efficiency has two components which are allocative and technical efficiency. Whereas $\frac{\partial \theta X^A}{\partial X^A}$ corresponds to the technical side of it, $\frac{\partial \alpha X^A}{\partial \theta X^A}$ is indicating the allocative component. The product of them gives the value of cost efficiency.

$$CE = \frac{\partial \theta X^A}{\partial X^A} \times \frac{\partial \alpha X^A}{\partial \theta X^A} = \frac{\partial \alpha X^A}{\partial X^A}$$

(3.14)

So as to measure the efficiency levels of firms, two separate methods have been developed by researchers under the rubric of mathematical programming approach and the econometric approach. Mathematical programming approach which is also known as Data Envelopment Analysis (DEA) was originated by Charnes, Cooper, and Rhodes (1978). In DEA, multiple outputs and inputs are reduced into a single output-input form in which efficiency measure is yielded after necessary calculations are completed with linear programming. Although DEA is frequently used in efficiency analysis its non-stochastic nature prevents researchers to attain comprehensive and sustainable results in many cases. Therefore,
econometric approach or stochastic frontier analysis became preferable owing to its ability to distinguish the impact of variation in technical efficiency from external stochastic error on the firm’s output. In the following sections, data envelopment and stochastic frontier analysis will be examined subsequently.

III. Data Envelopment Analysis (DEA)

One of the mainstream methods of efficiency analysis is DEA which doesn’t presume any functional form for production. It basically “involves the use of linear programming methods to construct a non-parametric piece-wise surface (or frontier) over the data” (Coelli et al, 2005:162). Therefore, efficiency of each decision making unit (DMU hereafter) which can be a bank, hospital, university and so forth is calculated regarding to the “best practising” producer. In other words, DEA is based upon a comparative analysis of observed producers to their counterparts (Greene, 2007). The comprehensive literature of this methodology can be reached in Charnes, Cooper and Rhodes (1978), Banker, Charnes and Cooper (1984), Dyson and Thanassoulis (1988), Seiford and Thrall (1990), Ali, Cook and Seiford (1991), Seiford and Tone (2000) and Thanassoulis (2001).

Data Envelopment Analysis was first coined by Charnes, Cooper and Rhodes (1978) which had an input-oriented model with constant return to scale (CRS). This method which is currently known as basic DEA was an extension of “Farrell’s measure to multiple - input multiple - output situations and operationalised it using mathematical programming” (Emrouznejad, 2000: 17). In subsequent researches, Färe, Grosskopf and Logan (1983) and Banker, Charnes and Cooper (1984), variable returns to scale (VRS) models were developed and introduced to the DEA literature. Furthermore, to capture the statistical error and separate it from efficiency term, two-sided deviation was brought in by Varian (1985) and besides “chance-constrained” efficiency analysis was integrated (Land et al., 1993) to the DEA models. And eventually, this efficiency estimation methodology is being used in the wide range of areas including management, operations research and economics.

Theoretical Framework

So as to illustrate basic DEA model mathematically, let’s assume that each decision-making units (DMUs) use \( m \) inputs for the production of \( n \) outputs in a given technology level. \( x_{ij} \) denotes the amount of input \( i \) \((i=1,2,\ldots,m)\) produced by \( j^{th} \) DMU \((j=1,2,\ldots,k)\) whereas \( y_{sj} \) represents the quantity of output \( s \) \((s=1,2,\ldots,n)\) produced by \( j^{th} \) DMU \((j=1,2,\ldots,k)\). The variables \( u_r \) \((r=1,2,\ldots,n)\) and \( w_i \) \((i=1,2,\ldots,m)\) are weights of each output and input respectively. The technical efficiency of \( DMU_0 \) can be written as:

\[
\text{Max} = \frac{\sum_{r=1}^{n} u_r y_{r0}}{\sum_{i=1}^{m} w_i x_{i0}}
\]  

(3.1)

Subject to:

\[
\sum_{s=1}^{n} u_r y_{sj} \leq 1 \quad \text{for } j=1,2,\ldots,k 
\]  

(3.2)

\[
\sum_{i=1}^{m} w_i x_{ij} \geq 0 \quad \text{for } (r=1,2,\ldots,n) \text{ and } (i=1,2,\ldots,m) 
\]  

(3.3)

www/hrmar.com/journals
This mathematical representation can be clarified as finding the appropriate values for \( u \) and \( w \) that maximise efficiency level of the observed firm subject to all efficiency scores are less than or equal to 1. To avoid infinite solutions (Coelli et al., 2005:163) and obtain a linear programming model, Charnes-Cooper transformation can be used as following:

\[
\text{Max} = \sum_{r=1}^{n} \mu_r Y_{r0} \tag{3.4}
\]

subject to:
\[
\sum_{i=1}^{m} w_i X_{i0} = 1, \tag{3.5}
\]
\[
\sum_{r=1}^{n} \mu_r Y_{rj} - \sum_{i=1}^{m} w_i X_{ij} \leq 0, \tag{3.6}
\]
\[
\mu_r \text{ and } w_i \geq 0 \quad (r=1,2,\ldots,n) \text{ and } (i=1,2,\ldots,m) \tag{3.7}
\]

Via using duality property of linear programming, equivalent form of this envelopment system can be illustrated as:

\[
\text{Min } \Theta \tag{3.8}
\]

subject to:
\[
\sum_{j=1}^{k} \lambda_j X_{ij} \leq \Theta X_{i0} \quad (i=1,2,\ldots,m) \tag{3.9}
\]
\[
\sum_{j=1}^{k} \lambda_j Y_{rj} \geq Y_{r0} \quad (r=1,2,\ldots,n) \tag{3.10}
\]
\[
\lambda_j \geq 0 \quad \text{for } j=1,2,\ldots,k \tag{3.11}
\]

where \( \Theta \) is a scalar and \( \lambda \) is a \( k \times 1 \) vector of constants. The solution of this linear system will end up with finding \( \Theta \)s corresponding to the efficiency level of each DMU. Therefore \( \Theta \) should be less than or equal to 1 as well as the firm with \( \Theta=1 \) is technically efficient that means operating on the frontier concerning Farell’s (1957) proposition.

In the previous section where Farell’s (1957) and Koopman’s (1951) definitions of efficiency were discussed, the magnitude of “coordinatewise improvements” were highlighted inspiring from Koopman’s analysis. Therefore, there is a precise need to integrate slack variables into the linear programming model through which efficiency scores will be gauged concerning the slack usage of any input. The model becomes as follows:

\[
\text{Min } \Theta \cdot (\sum_{i=1}^{m} s_i^+ + \sum_{r=1}^{n} s_r^-) \tag{3.12}
\]

subject to:
\[
\sum_{j=1}^{k} \lambda_j X_{ij} + s_i^- = \Theta X_{i0} \quad (i=1,2,\ldots,m) \tag{3.13}
\]
\[
\sum_{j=1}^{k} \lambda_j Y_{rj} + s_r^+ = Y_{r0} \quad (r=1,2,\ldots,n) \tag{3.14}
\]
\[
s_r^+, \ s_i^-, \lambda_j \geq 0 \quad \text{for } j=1,2,\ldots,k \tag{3.15}
\]

\( s_r^+ \) and \( s_i^- \) are constrained to become non-negative and transformed inequalities into equations. \( s_r^+ \) means that \( Y_{r0} \leq \sum \lambda_j Y_{rj} \) must be satisfied by every single solution, whereas \( s_i^- \) denotes that \( \sum \lambda_j X_{ij} \leq X_{i0} \) must be sustained for each input used by DMU \( 0 \).
As a result of all these linear programming iterations, efficiency level of the observed DMU - $DMU_0$ in this case - is equal to 100% if and only if:

i. $= 1$

ii. $= 0$ for all $(i=1,2,\ldots,m)$ and $(r=1,2,\ldots,n)$

If we turn back again the debate between Farell and Koopmans, proposition (i) is a necessary condition to Farell for efficiency; however Koopmans states that full efficiency necessitates both (i) and (ii).

Figure-3 illustrates DEA in a very generic representation which allows discussing Farell and Koopmans’s efficiency approaches straightforwardly. To Farell, all the points on the isoquant curve can be named as efficient combinations of input-1 and input-2. However, Koopmans reveals the fact that points on the isoquant curve with slack usage of inputs (like A, F, I) can’t be shown as efficient combination of inputs.

![Figure-3](image)

**CRS vs. VRS Models**

The analysis up to this point was assuming that DMUs are operating at constant return to scale (CRS) as put forward by Charnes, Cooper and Rhodes (1978) where $t$ times increase in inputs will result in $t$ times increase in output (i.e. $t^*Y = t^*f(X)$). On the other hand, in many sectors due to “imperfect competition, government regulations and constraints on finance” firms can’t be run at optimal scale (Coelli et al., 2005:172). Therefore, scale efficiency which has an impact on technical efficiency of a firm arises in these circumstances.

So as to capture the magnitude of “scale effect”, Färe, Grosskopf and Logan (1983) and Banker, Charnes and Cooper (1984) developed a variable returns to scale (VRS) in which CRS assumption is relaxed. Figure-4 illustrates the divergence of VRS models from CRS ones in a quite generic way. For instance, the efficiency of point B is calculated as the ratio of $O_1/O_2$ regarding VRS frontier, whereas is equal to $O_1/O_3$ if CRS frontier is taken as the reference point. Eventually, it is apparent that VRS frontier takes the magnitude of scale efficiency into account while measuring the total efficiency.
Linear programming model of VRS is quite similar to the CRS as indicated in (3.8), (3.9), (3.10) and (3.11). Only difference is addition of a convexity constraint to the system:

$$\sum_{j=1}^{k} \lambda_j = 1, \text{ for } j=1,2,\ldots,k$$  \hspace{1cm} (3.16)

The mathematical relationship between VRS and CRS efficiency measurements can be illustrated as (Coelli et al., 2005:173):

$$TE_{CRS} = TE_{VRS} \times SE,$$  \hspace{1cm} (3.17)

which means that CRS technical efficiency of a firm can be decoupled into pure technical efficiency and scale efficiency (SE). Even though, an analytical association exists among CRS and VRS models, input and output efficiency scores are different in VRS unlike in CRS models (Emrouznejad, 2000: 25).

**Input and Output Oriented Measurements**

As mentioned earlier, the chief objective of a firm in market economies is either minimizing input or output maximization. Both in CRS and VRS models, input and output oriented measurements can be conducted pertaining to the preference of researcher. As Figure-5 demonstrates clearly, output-oriented frontier represents all combinations of outputs that are attainable by the production unit. Whilst the efficient frontier in input-oriented model refers to the minimum usage of inputs to produce given output level, efficient frontier in output-oriented model denotes maximum amount of outputs given input level.
In the previous linear programming systems, input-oriented analysis was articulated relying on radially contraction of input vector without any change in output. In a similar vein, for output-oriented measurements, technical efficiency is calculated as a proportional increase in output level lacking any alteration in the amount of inputs. If this narrative is transliterated to mathematical lexicon as VRS:

\[
\begin{align*}
\text{Max } & \Phi \\
\text{subject to:} & \\
\sum_{j=1}^{k} \lambda_j x_{ij} & \leq x_{i0} & (i=1,2,\ldots,m) \\
\sum_{j=1}^{k} \lambda_j y_{rj} & \geq \Phi y_{r0} & (r=1,2,\ldots,n) \\
\sum_{j=1}^{k} \lambda_j & = 1, \text{ for } j=1,2,\ldots,k \\
\lambda_j & \geq 0 \text{ for } j=1,2,\ldots,k
\end{align*}
\]

In this case, firms confronting higher \( \Phi \)s will have lower technical efficiency scores and the firm with \( \Phi=1 \) can be identified as technically efficient in a given technological progress. Furthermore, due to the fact that input and output oriented DEA models are estimating same frontier, set of efficient firms will be the same, whilst there might be slight differences in the efficiency scores of inefficient firms.

**Extensions in DEA**

**Allocative Efficiency**

In DEA models, allocative efficiency of a DMU can be gauged alongside the technical efficiency scores by the means of cost minimisation or revenue/profit maximisation if price information about input set is available. Let’s take VRS cost minimization case with input-orientated model as an example to demonstrate the measurement of allocative efficiency via using same linear system of (3.12), (3.13), (3.14) and (3.15):
Min $\Theta_0 \cdot \varepsilon \left( \sum_{i=1}^{m} s^{-}_{i} + \sum_{r=1}^{n} s^{+}_{r} \right) \tag{3.12}$

subject to:

$\sum_{j=1}^{k} \lambda_j x_{ij} + s^{-}_{i} = \Theta x_{i0} \quad (i=1,2,\ldots,m) \tag{3.13}$

$\sum_{j=1}^{k} \lambda_j y_{rj} + s^{+}_{r} = y_{r0} \quad (r=1,2,\ldots,n) \tag{3.14}$

$s^{+}_{r}, s^{-}_{i}, \lambda_j \geq 0 \quad \text{for } j=1,2,\ldots,k \tag{3.15}$

$\sum_{j=1}^{k} \lambda_j = 1, \text{ for } j=1,2,\ldots,k \tag{3.16}$

where $p_i$ represents price data about input set and $X_{i1}^*$ is the cost minimising input quantities derived by linear programming. Eventually, cost efficiency (i.e. economic efficiency) of the firm is calculated as the minimum cost to observed cost:

$CE = \frac{\sum_{i=1}^{m} p_i X_{i0}^*}{\sum_{i=1}^{m} p_i X_{i0}} \tag{3.23}$

If the price information of output is available as well as “revenue maximisation is a more appropriate behavioural assumption” (Coelli et al., 2005:184), then programming model can be converted to revenue maximisation with VRS shown below:

Max $\sum_{i=1}^{n} q_i Y_{i0}^*$

subject to

$\sum_{j=1}^{k} \lambda_j x_{ij} \leq x_{i0}^* \quad (i=1,2,\ldots,m) \tag{3.24}$

$\sum_{j=1}^{k} \lambda_j y_{rj} \geq y_{r0} \quad (r=1,2,\ldots,n) \tag{3.25}$

$\lambda_j \geq 0 \quad \text{for } j=1,2,\ldots,k \tag{3.26}$

$\sum_{j=1}^{k} \lambda_j = 1, \text{ for } j=1,2,\ldots,k \tag{3.27}$

where $q_i$ refers to price information of corresponding output levels, $Y_{i0}^*$ is the revenue maximisation amounts of output attained at the end of solution iterations. Then, revenue efficiency of the observed DMU is computed as the ratio of observed revenue to maximum revenue:

$RE = \frac{\sum_{i=1}^{n} q_i Y_{i0}}{\sum_{i=1}^{n} q_i Y_{i0}^*} \tag{3.28}$

Lastly, if both price data of inputs and outputs is available, profit efficiency of the DMUs can be calculated via using DEA. Profit maximisation with VRS model is specified as:

Max $(\sum_{i=1}^{n} q_i Y_{i0}^* - \sum_{i=1}^{m} p_i X_{i0}^*) \tag{3.29}$

subject to

$\sum_{j=1}^{k} \lambda_j x_{ij} \leq x_{i0}^* \quad (i=1,2,\ldots,m) \tag{3.30}$

$\sum_{j=1}^{k} \lambda_j y_{rj} \geq y_{r0} \quad (r=1,2,\ldots,n) \tag{3.31}$
\[ \lambda_j \geq 0 \quad \text{for } j=1,2,\ldots,k \]  
\[ \sum_{j=1}^{k} \lambda_j = 1, \quad \text{for } j=1,2,\ldots,k \]  

Profit efficiency of the observed DMU can be computed as the ratio of observed profit to maximum attainable profit, however in this case efficiency values may be equal to 0 if observed profit is zero, and undefined if maximum profit is zero:

\[ PE = \frac{\sum_{i=1}^{n} q_i y_{i0} - \sum_{j=1}^{k} p_j x_{j0}}{\sum_{i=1}^{m} p_i x_{i0}^* - \sum_{j=1}^{k} p_j x_{j0}^*} \]

Heterogeneity

Another point worth examining in DEA models is that how this mathematical programming system deals with “environmental factors” that cause inefficiencies out of firm’s control (Coelli et al., 2005:190). In these cases, efficiency of the DMU could be influenced by external effects which aren’t controlled or directed by decision makers inside the firms and accordingly efficiency scores can be miscalculated due to these effects. Couple of methods with some weaknesses have been named to cope with this obstacle up to now.

First method (Banker and Morey, 1986) proposes comparing the efficiency of firm A with the firms in the sample which have the value for environmental factor (which can be ordered from worst to best) less than or equal to firm A. The second method developed by Charnes, Cooper and Rhodes (1981) puts forward that 1) dividing the sample into two sub-samples and solution by DEA 2) project all observed points on the frontier 3) solving a single DEA system and check whether there is any difference in the mean efficiencies of two sub-samples. Third one suggests the inclusion of environmental variables directly into the linear system by integrating the expression following where Z indicates environmental variable:

\[ \sum_{j=1}^{k} \lambda_j Z_j \leq Z_1 \]  

The last one encourages researchers to conduct two-stage method in which solving the system by DEA in a traditional way forms the former leg and regressing efficiency scores onto the environmental factors forms the latter one. Although all these four methods have deficient aspects, they give a decent insight to separate the effects of external factors from efficiency scores within the DEA method.

Additional Methods

Even though flexibility in DEA is praised frequently in the literature, its structure may cause troubles if assigned weights to the input/output sets show unrealistic properties (Coelli et al., 2005:199). For this reason, researchers can construct more realistic models to “improve the discrimination of models” through weight restrictions (WR) on output and/or input bundles (Podinovski, 2002). Lower and upper bounds are specified for weights of input and output sets and then incorporated to the linear programming system:

\[ \mu_L \leq \mu \leq \mu_U \quad \text{(restrictions on output)}, \quad w_L \leq w \leq w_U \quad \text{(restrictions on input)} \]  

The main problem in WR models is the possibility of ending up with an inappropriate boundary which is solely contingent upon researchers’ own value judgments.
Another concept in additional methods is super efficiency. This method relaxes the linear programming system not to use observed DMU as its own peer which results in efficiency scores with greater than 1. For instance, regarding to this chapter’s representation, $X_{11}$ and $Y_{11}$ are dropped from the left hand-side of the input (3.9) and output (3.10) inequalities correspondingly to eliminate peer effect of $DMU_1$. Thanks to this omission, the analyst is able to compare efficient firms (with $\theta=1$) operating just at the frontier with its efficient counterparts, however no changes could be observed for inefficient firms.

Last additional method in DEA is bootstrapping which provides statistical properties to DEA estimations (Coelli et al., 2005:202). This method fundamentally produces a number of random samples which have same sample sizes from initial data set (Sena, 2003). The chief advantage of this “re-sampling technique” is to allow constructing confidence intervals and thus conducting hypothesis testing on estimated efficiency scores. As Coelli et al. (2005) articulates clearly, this re-sampling shouldn’t be confused with random noise motivated from measurement or specification error in stochastic analysis.

Previous Empirical Studies

Since the first paper was published by Charnes, Cooper and Rhodes (1978), DEA has been applied to a wide range of sectors compromising health care, education and banking. Particularly, due to their significance in public services provision, hospitals and higher education institutions (HEIs) have extremely attracted attentions of researchers to conduct efficiency analysis by the means of DEA. In the following sentences, only milestone papers in this research field will be exposed and discussed, for a detailed review of literature bibliography of this paper can be glanced at.

One of the first papers in this area is Sherman’s (1984) research in which efficiency scores of seven US teaching hospitals were calculated. And subsequently, physician efficiency in hospital services (Chilingerian and Sherman, 1990; Chilingerian, 1994), efficiency of nursing homes (Chattopadhyah and Ray, 1996) and health maintenance organizations (Siddharthan et al., 2000) were examined hinging upon the analytical premises of preceding theoretical works. In addition to these studies, Chang (1998) attempted to figure out the determinants of hospital efficiency via taking central-government owned hospitals in Taiwan between 1990 and 1994 as an example. The results revealed that proportion of retired veteran patients and scope of services have significantly negative impact on efficiency, even as occupancy is motivating efficiency in a positive way (Chang, 1998).

Another area of interest in DEA is HEIs on which a plethora of academic papers have been put forward. The pioneering empirical study in university sector is published by Coelli (1996) on 36 Australian universities through input orientated VRS DEA model for the calendar year 1994. Throughout the following years, comparative efficiency of UK HEIs (Athanassopoulos & Shale, 1997) efficiency of Finnish universities by efficient facets model (Raty, 2002) efficiency levels of Australian universities (Abbott & Doucouliagos, 2003) and UK universities (Johnes, 2005) were investigated using primarily VRS-based models.
In the banking sector analysis, Sherman and Gold (1985) were the first to adopt DEA to calculate efficiency scores of banks. Bhattacharyya et al. (1997) use VRS DEA to examine Indian commercial banks between 1986 and 1991 within the framework of “grand frontier” approach in which data for banks through all years were pooled. A similar research to the previous work was initiated by Sathye (2003) relying on 1997-1998 dataset consisted of 94 banks including public sector, private and foreign ones. Recently, researchers (Anouze and Emrouznejad, 2006; Mostafa, 2007; Johnes et al., 2009) began to focus on the banks in Gulf region that have a different idea in banking methodology motivated by Islamic finance.

IV. Stochastic Frontier Analysis (SFA)

As Greene (1997) figured out, in general, frontier production function can be described as “an extension of the familiar regression model based on the microeconomic premise that a production function represents some sort of ideal, the maximum output attainable given a set of inputs.” In recent researches, to measure the efficiency level of a firm/organization, distance between estimated production frontier and observed one is computed. Prior to current analysis, different approaches have been developed for efficiency measurement in an econometric way by researchers (Farell: 1957, Aigner and Chu: 1968).

The initial framework on parametric frontier analysis commenced with Farell’s (1957) cross-sectional model where goal programming techniques were used to estimate production function. Parametric frontier is specified as:

\[ Y_i = f(X_i, \beta) \cdot TE_i \]  

(3.36)

where \( i = (1, 2, 3, \ldots, I) \) represents the corresponding produces, \( Y \) is the level of output, \( X \) refers to a vector of \( N \) inputs, \( f(\cdot) \) is the production frontier depending on inputs and technology parameters \( (\beta) \) to be estimated. The last term \( TE_i \) is the technical efficiency of the \( i^{th} \) firm calculated as the ratio of observed output over maximum feasible output:

\[ TE_i = \frac{Y_i}{f(X_i, \beta)} \]  

(3.37)

Aigner and Chu (1968) reformulated frontier function above with log-linear Cobb-Douglas production function which was still reflecting the behaviours of deterministic frontiers:

\[ \ln Y_i = \beta_0 + \beta_n \ln X_n - u_i \]  

(3.38)

Even though frontier functions became parameterised with these extensions, technology parameters aren’t estimated in any statistical sense, rather they are calculated via using mathematical programming techniques. Therefore, to be able to “capture the effects of exogenous shocks beyond the control of analysed units”, alternative econometric approaches were put forward during the subsequent researches in this particular area of research (Murillo and Zamorano: 2004).

In two independent papers by Aigner et al. (1977) and Meeusen and van den Broeck (1977) stochastic frontier function for Cobb-Douglas case was specified as following:

\[ \ln Y_i = \beta_0 + \sum_{n=1}^{N} \beta_n \ln X_{ni} + \epsilon_i \]  

(3.39)
Where \( \ln y_i \) represents the logarithm of observed output, \( X_{ni} \) is the vector of given inputs and \( \beta_n \) is a vector of unknown parameters. Accordingly \( \epsilon_i \) is specified as:

\[
\epsilon_i = v_i - u_i, \quad u_i \geq 0
\]  

(3.40)

First error component \( v_i \) is independently and identically distributed as \( v_i \sim N(0, \sigma_v^2) \) and captures the effects of statistical noise such as random effects of measurement error and external shocks out of firm’s control, while \( u_i \) is independently and identically half-normal distributed \( u_i \sim N^+ (0, \sigma_u^2) \) and intended to capture technical inefficiency which can be measured as the deficiency in output away from the maximum possible output given by the stochastic production frontier:

\[
\ln y_i = f(x_i, \beta) + v_i
\]  

(3.41)

The property that \( u_i \geq 0 \) ensures all the observed outputs should lie below or on the stochastic frontier. Any deviation from the aforementioned frontier will be treated as the result of factors controlled by firm that named as technical and economic inefficiency (Aigner et al., 1977). Eventually, technical efficiency (economic efficiency will be discussed in the next section) of the \( i^{th} \) firm can be depicted as:

\[
TE_i = \frac{x_i}{\exp (f(x_i, \beta) + v_i)} = \exp \left( \frac{f(x_i, \beta) + v_i - u_i}{\exp(f(x_i, \beta) + v_i)} \right) = \exp (-u_i)
\]  

(3.42)

First and foremost motivation behind efficiency analysis is to estimate maximum feasible frontier and accordingly measure the efficiency scores of each and every DMU relative to that frontier. This estimation process was initially originated with different versions of Ordinary Least Squares (OLS) as shown in Figure-6 through which average-practising frontier was estimated and shifted up by (i) the maximum amount of residuals (Corrected OLS) coined by Gabrielsen (1975) or (ii) the mean of the residuals (Modified OLS) used by Richmond (1974). Apparently, the main drawback of these estimation procedures is taking the efficiency performance of the average producer as a benchmark (instead of best-practising one) and calculating other observed units’ efficiency concerning that point.

In lieu of using OLS, Greene (1980a) preferred Maximum Likelihood Estimation (MLE) to estimate technology parameters as well as residuals which were eventually decomposed into
statistical noise and inefficiency term by Jondrow et al (1982). The parameters of this regression ($\beta_s$) are being estimated by log-likelihood function in which $\sigma^2 = \sigma^2_\beta + \sigma^2$ and $\lambda^2 = \sigma^2/\sigma^2 \geq 0$. If $\lambda = 0$, then that means all deviations from the stochastic frontier is linked to noise term and there isn’t any sort of technical inefficiency. The log-likelihood function is written as:

$$\ln L(y \mid \beta, \sigma, \lambda) = -\frac{1}{2} \ln \left( \frac{\pi \sigma^2}{2} \right) + \sum_{i=1}^{I} \ln \Phi \left( \frac{-\mu_i}{\sigma} \right) - \frac{1}{2\sigma^2} \sum_{i=1}^{I} (e_i^2)$$

(3.43)

Where $y$ and $e_i$ are corresponding same representations in (3.22), $\Phi(x)$ is the cumulative density function (cdf) of the standard normal distribution. To estimate $\beta_s$, “iterative optimisation procedure” is undertaken until the values maximising the function is able to be obtained (Coelli et al., 2005:246).

**Estimating Inefficiency Term**

It is obvious that the very goal of efficiency estimation procedures is not solely about figuring out technology parameters but to gauge efficiency performances of each individual unit. So as to estimate them, residuals ($e_i$) obtained from MLE must be decomposed into their components. As indicated previously, Jondrow et al. (1982) developed a method (known as JLMS) which is an indirect estimation of inefficiency term ($u_i$) dependent on $\epsilon_i = v_i - u_i$:

(3.44)

$\phi$ and $\Phi$ denote standard normal density and cumulative density function, respectively. The JLMS technique includes two different distribution assumptions for $u_i$ consisting of normal-truncated normal and normal-exponential in separate analysis. The former terms in the distribution types correspond to the distribution of $v_i$s, whilst latter terms are indicating the distribution of $u_i$s. In the case of normal-truncated normal model where the conditional distribution of $u_i$ on $\epsilon_i$ is $N(\mu, \sigma^2)$ truncated at zero:

$$E(u_i \mid \epsilon_i) = \mu + \sigma \left[ \frac{\phi(-\mu_i/\sigma)}{1-\Phi(-\mu_i/\sigma)} \right] = -\lambda \epsilon_i$$

(3.45)

The normal-exponential model in which the conditional distribution of $u_i$ on $\epsilon_i$ is assumed as exponential with density function $f(u) = \exp(-u/\sigma_u)/\sigma_u$ and have a conditional distribution $N(-\sigma_u A, \sigma^2_u)$:

$$E(u_i \mid \epsilon_i) = \sigma_u \left[ \frac{\phi(A)}{1-\Phi(A)} - A \right] = \frac{\epsilon_i}{\sigma_u} + \frac{\sigma_u}{\sigma_u}$$

(3.46)

Thanks to JLMS technique, error term was begun to be separated into its components that are statistical noise and inefficiency term which is the main notion under examination for this particular research field.

In the estimation of inefficiency term, the major concern of researchers is to decide on the appropriate distribution function of it. Up till now, Aigner et al. (1977) proposed half-normal, Stevenson (1980) used truncated normal, Greene (1980) preferred to use gamma, and finally Beckers and Hammond (1997) extended exponential distribution function for inefficiency component of error term. Although, to opt for the best-fitted distribution is overwhelmingly difficult, prior theoretical insights of researchers do shape this decision making process. Coelli et al. (2005) underlines the notion of parsimony which is in favour of choosing the less complicated one ceteris paribus. Therefore, half-normal and exponential distributions are the
best candidates which have simpler structures than other aforesaid options (Coelli et al., 2005: 252).

**Stochastic Cost Frontier Approach**

After examining the concept of stochastic frontier analysis, in this section stochastic cost frontier approach (SCFA) which is expected to be used during this PhD research will be scrutinized specifically. SCFA paves the way for researchers who are dealing with economic efficiency analysis in which both technical and allocative efficiency of the given firms can be worked out.

SCFA basically defines minimum cost in a given output level and input prices relying on existing technology of production (Farsi and Flippini: 2003). In this way of measurement, efficiency level of a particular institution or a firm is gauged with respect to the inefficient usage of inputs within a given cost function. The key difference between stochastic and deterministic models is that stochastic analysis comprises error term (Karim and Jhantasana: 2005), therefore it can separate the inefficiency effect from statistical noise. That is to say, deterministic models aren’t capable of differentiating the influence of irrelevant factors or unexpected shocks on output level.

The cost function of a firm represents the minimum amount of expenditure for a production of a given output, therefore if the producer is operating inefficiently its production costs must be greater than theoretical minimum. Then, it is quite obvious that frontier cost function can be assigned as an alternative to frontier cost production (Greene: 1997). In a similar vein, frontier production function illustrated above can be converted to frontier cost function which will be articulated below via changing the sign of inefficiency error component consisting of both technical and allocative inefficiency (Kumbkahar and Lovell: 2000). Decomposition of the inefficiency term into the technical and allocative ones is examined by Schmidt and Lovell (1977) for Cobb Douglas functions and Kopp and Divert (1982) for general translog cases.

Unlike in the estimation of technical efficiency relying on output-oriented approaches, SCFA prioritise input-oriented approaches to estimate efficiency on the cost frontier (Zhao: 2006). Furthermore, Zhao (2006) puts forward that estimating cost efficiency differs from technical efficiency estimations in the sense of ‘data requirements, number of outputs, quasi-fixity of some inputs and decomposition of efficiency itself’. Eventually, the function is specified as:

\[ \ln(C_i) = \ln C(p_i, q_i; \theta) + v_i + u_i \]  \hspace{1cm} (3.47)

Where \( C_i \) is the observed cost, \( p_i \) is a vector of input prices, \( q_i \) is a vector of output prices, \( \theta \) is a vector of technology parameters to be estimated, \( u_i \) is a non-negative stochastic error capturing the effects of inefficiency and \( v_i \) is a symmetric error component has the same distribution in (3.23) and reflecting the statistical noise. Cost efficiency can be illustrated as:

\[ CE_i = \frac{C(p_i, q_i; \theta) \exp(u_i)}{q_i} \]  \hspace{1cm} (3.48)
Where $CE_i$ reflects the ratio of the minimum possible cost, given inefficiency $u_i$, to actual total cost. If $C_i = C(p_i, q_i; \theta) \exp(u_i)$, then $CE_i=1$ and we can say that firm $i$ is fully efficient. Otherwise actual cost for firm $i$ exceeds the minimum cost so that $0 < CE_i \leq 1$.

**Extensions in SFA**

Recent studies in SFA extended the volume of analysis via integrating observed/unobserved heterogeneity, panel data models and Bayesian inferences to the literature. Each of these extensions will be identified and examined in the following sections separately.

**Heterogeneity**

Heterogeneity among the organizations is often classified as observed and unobserved heterogeneity. The former conception refers to the cases where variations can be reflected in measured variables, whereas the latter term which is usually assumed time-invariant comes into the function as effects (Greene, 2007). Major concern for researchers to figure out heterogeneity is the likelihood of treating this given variation as inefficiency which is actually not. To account for observable heterogeneity in efficiency analysis, variable $z$ is identified and incorporated to the non-stochastic part of frontier function which has similar distribution properties for $u_i$ and $u_k$:

$$
\ln y_i = f(x_i, \beta) + Z_i + v_i - u_i
$$

(3.49)

Couple of models have been developed in order to explore the relationship between exogenous (environmental) factors ($Z$s) and inefficiencies as well as separating them from each other. The earliest paper conducted this analysis is the study of Pitt and Lee (1981) whom used two-stage approach. In the initial stage, they estimated conventional frontier function without taking any environmental variables into consideration, and secondly, the projected efficiencies are regressed onto $Z$s. The chief problem arises here is that exclusion of environmental variables in the first stage leads to biased estimators both for parameters of non-stochastic part of function and inefficiency terms as indicated by Caudill et al.(1995) and Wang and Schmidt (2002). To achieve the same target from a different technique, Kumbhakar et al. (1991) allowed $Z$ terms to influence $u_i$ directly by assuming the distribution of it as $N^+(Z_i, \sigma_u^2)$.

Orea and Kumbhakar (2003) came up with another method (one-stage approach) to deal with heterogeneity named as “latent class stochastic frontier model” (LCSFM) which is the combination of latent class structure and SFA. This method -applied to Spanish banking sector- basically segregates the whole dataset to the number of classes which is usually determined by Akaike Information Criterion (AIC) or Schwarz Criterion (BIC) and estimates a unique frontier for each class in the sample. Consequently, predicting biased estimators in “one sample case” due to heterogeneity is able to be avoided owing to this class segregation methodology.
Panel Data (Fixed Effects and Random Effects Models)

The stochastic frontier function in previous studies was lack of “time effect” which is indispensable for panel data models. Preferring panel data in lieu of cross-section offers a number of advantages due to its data enriched structure. Coelli et al. (2005:275) enumerates three of these as following:

- Some of the distributional assumptions to differentiate statistical noise and inefficiency terms is relaxed
- To obtain more consistent estimators of inefficiencies
- Examining the change in inefficiencies over time (which might be a good indication of technological progress)

To incorporate time effect into the stochastic frontier in (3.22), “t” will be added as a subscript alongside with i:

\[
\ln y_{it} = f(x_{it}, \beta) + \epsilon_{it}, \quad \epsilon_{it} = v_{it} - u_{it}
\]

(3.50)

The main matter of discussion for this modelling in panel data analysis is whether to treat inefficiency term as time variant or invariant. Time-invariant inefficiency model imposes \( u_{it} = u_i \) and is estimated either by fixed effects or random effects approach. On the other hand, time-variant inefficiency supposes that firms learn from their experiences to enhance efficiency levels incrementally that can be formulated as \( u_{it} = f(t) u_i \) (Coelli et al., 2005:278). Two diverse functional forms for capturing time effect have been generated: first one is by Kumbhakar (1990) \( f(t) = [1 + \exp (\alpha t + \beta t^2)]^{-1} \) and the second belongs to Battese and Coelli (1992) \( f(t) = \exp[(t - T)] \) where \( f, \alpha \) and \( \beta \) are unknown parameters to be estimated.

Last point that should be touched upon in this section is heterogeneity in panel data estimation of efficiency terms which was elucidated in Greene’s (2005) seminal work. In this paper, Greene (2005) discusses pros and cons of fixed effects and random effects models as well as proposes his own methodology called as “true fixed effects” and “true random effects” models. The fixed effects model illustrated below treats \( \alpha_i \) as firm-specific inefficiency, thus any heterogeneity among firms is omitted:

\[
y_{it} = \alpha_i + \beta x_{it} + v_{it}
\]

(3.51)

To overcome this problem, true fixed effects is brought into:

\[
y_{it} = \alpha_i + \beta x_{it} + v_{it} + u_{it}
\]

(3.52)

In random effects model where firm-specific inefficiency is assumed as constant over time, the frontier function is narrated as:

\[
y_{it} = \alpha_i + \beta x_{it} + v_{it} + u_{it}
\]

(3.53)

Due to the shortcomings of this model indicated by Greene (2005), he specified a comprehensive frontier production function to separate firm-specific effects denoted by \( w_i \) from inefficiency terms and called it true random effects:

\[
y_{it} = \alpha_i + \beta x_{it} + w_i + v_{it} + u_{it}
\]

(3.54)

Some may think that there are three disturbance terms in the regression \(\{w_i, v_{it}, u_{it}\}\) but indeed not because the real model has two disturbances:

\[
y_{it} = \alpha_i + \beta x_{it} + w_i + v_{it}
\]

(3.55)
Comparison of DEA and SFA

Whereas the superiority of SFA over to the DEA was revealed as a) including statistical noise into the frontier b) allowing statistical tests on the estimates, DEA is seen advantageous at times due to the fact that it doesn’t require any specific functional form for production function and distributional form for inefficiency terms. For that reason, trade-off between misspecification bias (in SFA) and measurement error (in DEA) determines the preference of researchers conducting efficiency analysis. To alleviate the repercussions motivated from this trade-off, statistical properties are trying to be integrated to the deterministic approaches, even as recent applications using diverse collection of functional forms prevents stochastic methods to be over-parameterised (Fried et al., 2007).

The first paper comparing SFA and DEA relying on a sample data was Gong and Sickles (2002). The result put forward by them was claiming that “Our results indicate that for simple underlying technologies the relative performance of the stochastic frontier models vis-a-vis DEA relies on the choice of functional forms”. In addition to that, severity of misspecification error accompanied by “degree of correlatedness of inefficiency with regressors” makes DEA more appealing (Gong and Sickles, 2002). In this particular case, the loser of the trade-off between misspecification and measurement errors is named as the former one. In another two separate papers, Bauer et al. (1998) and Cummins and Zi (1998) using dataset of US banks and life insurance companies, they explored a weak but positive rank correlation between point estimators of econometric and mathematical programming techniques (Fried et al., 2007).

The other point that gives idea about the robustness and appropriateness of these two methodologies is the value of “λ” corresponding to $\frac{\sigma_u}{\sigma_v}$. If λ gets closer to $+\infty$, this means all variation from frontier is being motivated from inefficiency which is the chief argument of deterministic frontiers. In a similar vein, in the cases where λ is close to 0, stochastic analysis is worth opting for (Greene, 2007).

As a result of all these aforementioned arguments, it’s extremely obvious that choosing one method to another will always have a certain amount of opportunity cost. Therefore it’d be better to finish this chapter with Sena’s (2003) arguments: “It is really impossible to suggest one approach to the other, as they both have positive and negative features; in a sense, they could be used jointly as they provide complementary information. At any rate, it is clear that the frontier approach offers an interesting set of tools to measure efficiency and total factor productivity (TFP) and so contribute to decision-making within both private and public organizations”
References


