Chapter 2

Empirical Methodology for Efficiency Estimation: An Introduction to Data Envelopment Analysis and Malmquist Analysis

This chapter sets the modeling stage with the specification of Data Envelopment Analysis (DEA) model and Malmquist TFP growth and their constituent components. The chapter is divided into two sections. Section one provides a brief introduction to efficiency measurement concepts developed by Farrell (1957) and others and section two outlines how these ideas may be empirically implemented by using Linear Programming method i.e. DEA method and Malmquist index.

2.1 Efficiency Measurement Concepts

This section provides a very brief introduction to modern efficiency measurement which begins with Farrell (1957) who drew upon the work of Debreu (1951) and Koopmans (1951) to define a simple measure of firm efficiency, which could account for multiple inputs. He proposed that the efficiency of a firm consists of two components: technical efficiency, which reflects the ability of a firm to obtain maximal output from a given set of inputs, and allocative efficiency (Price efficiency), which reflects the ability of a firm to use the inputs in optimal proportions, given their respective prices. These two measures are then combined to provide a measure of Overall efficiency.

2.1.1 Input-Orientated Measures

Farrell illustrated his ideas using a simple example involving firms which use two inputs \(x_1\) and \(x_2\) to produce a single output \(y\), under the assumption of constant returns to scale (The constant returns to scale assumption allows one to represent the technology using a unit isoquant. Furthermore, Farrell also discussed the extension of his method so as to accommodate more than two
inputs, multiple outputs, and non-constant returns to scale). Knowledge of the unit isoquant of the fully efficient firm (The production function of the fully efficient firm is not known in practice, and thus must be estimated from observations on a sample of firms in the industry concerned. In this study we use DEA to estimate this frontier), represented by SS' in chart 2.1, permits the measurement of technical efficiency. If a given firm uses quantities of inputs, defined by the point P, to produce a unit of output, the technical inefficiency of that firm could be represented by the distance QP, which is the amount by which all inputs could be proportionally reduced without a reduction in output. This is usually expressed in percentage terms by the ratio QP/OP, which represents the percentage by which all inputs could be reduced. The technical efficiency (TE) of a firm is most commonly measured by the ratio

$$TE_I = \frac{OQ}{OP}, \quad (1)$$

This is equal to one minus QP/OP (The subscript “I” is used on the TE measure to show that it is an input-orientated measure). It will take a value between zero and one, and hence provides an indicator of the degree of technical inefficiency of the firm. A value of one indicates the firm is fully technically efficient. For example, the point Q is technically efficient because it lies on the efficient isoquant.

Chart 2.1

Technical and Allocative Efficiencies

![Diagram showing technical and allocative efficiencies](chart2.1.png)
If the input price ratio, represented by the line AA’ in chart 2.1, is also known, allocative efficiency may also be calculated. The *allocative efficiency* (*AE*) of the firm operating at P is defined to be the ratio

\[ \frac{OR}{OQ}. \]  

(2)

Since the distance RQ represents the reduction in production costs that would occur if production were to occur at the allocatively (and technically) efficient point Q’, instead of at the technically efficient, but allocatively inefficient, point Q.

The *overall efficiency* (*OE*) is defined to be the ratio

\[ \frac{OR}{OP}. \]  

(3)

Where the distance RP can also be interpreted in terms of a cost reduction. Note that the product of technical and allocative efficiency provides the overall economic efficiency

\[ \frac{OQ}{OP} \times \frac{OR}{OQ} = \frac{OR}{OP} = OE. \]  

(4)

Note that all these measures are bounded by zero and one.

These efficiency measures assume the production function of the fully efficient firm is known. In practice this is not the case, and the efficient isoquant must be estimated from the sample data. Farrell suggested the use of either (a) a non-parametric piecewise-linear convex isoquant constructed such that no observed point should lie to the left or below it (refer to chart 2.2), or (b) a parametric function, such as the Cobb-Douglas form, fitted to the data, again such that no observed point should lie to the left or below it. Farrell provided an illustration of his methods using agricultural data for the 48 continental states of the US.
1.2 Output-Orientated Measures

The above input-orientated technical efficiency measure addresses the question: "By how much can input quantities be proportionally reduced without changing the output quantities produced?" One could alternatively ask the question: "By how much can output quantities be proportionally expanded without altering the input quantities used?". This is an output-orientated measure as opposed to the input-oriented measure discussed above. The difference between the output- and input-orientated measures can be illustrated using a simple example involving one input and one output. This is depicted in chart 2.3(a) where we have decreasing returns to scale technology represented by $f(x)$, and an inefficient firm operating at the point P. The Farrell input-orientated measure of TE would be equal to the ratio $AB/AP$, while the output-orientated measure of TE would be $CP/CD$. The output- and input-orientated measures will only provide equivalent measures of technical efficiency when constant returns to scale exist, but will be unequal when increasing or decreasing returns to scale are present (Fare and Lovell 1978). The constant returns to scale case is depicted in chart 2.3(b) where we observe that $AB/AP=CP/CD$, for any inefficient point P we care to choose.
Input and Output-Orientated Technical Efficiency Measures and Returns to Scale

One can consider output-orientated measures further by considering the case where production involves two outputs \(y_1\) and \(y_2\) and a single input \(x_1\). Again, if we assume constant returns to scale, we can represent the technology by a unit production possibility curve in two dimensions. This example is depicted in chart 4 where the line ZZ' is the unit production possibility curve and the point A corresponds to an inefficient firm. Note that the inefficient point, A, lies below the curve in this case because ZZ' represents the upper bound of production possibilities.

Chart 2.4
The Farrell output-orientated efficiency measures would be defined as follows. In chart 2.4 the distance AB represents technical inefficiency. That is, the amount by which outputs could be increased without requiring extra inputs. Hence a measure of output-orientated technical efficiency is the ratio

\[ \text{TE}_o = \frac{OA}{OB}. \] (7)

If we have price information then we can draw the isorevenue line DD', and define the allocative efficiency to be

\[ \text{AE}_o = \frac{OB}{OC} \] (8)

which has a revenue increasing interpretation (similar to the cost reducing interpretation of allocative inefficiency in the input-orientated case). Furthermore, one can define overall economic efficiency as the product of these two measures

\[ \text{OE}_o = \frac{OA}{OC} = \frac{OA}{OB} \times \frac{OB}{OC} = \text{TE}_o \times \text{AE}_o. \] (9)

Again, zero and one bound all of these three measures.

### 2.2.1 DATA ENVELOPMENT ANALYSIS (DEA)

Data envelopment analysis (DEA) is the non-parametric mathematical programming approach to frontier estimation. The piecewise-linear convex hull approach to frontier estimation, proposed by Farrell (1957), was considered by only a handful of authors in the two decades following Farrell's paper. Authors such as Boles (1966)\(^3\) and Afriat (1972)\(^4\) suggested mathematical programming methods, which could achieve the task, but the method did not receive wide attention until the paper by Charnes, Cooper and Rhodes (1978)\(^5\) which coined the term Data Envelopment Analysis (DEA). There have since been a large number of papers, which have extended and applied the DEA methodology.
Charnes, Cooper and Rhodes (1978) proposed a model, which had an input orientation and assumed constant returns to scale (CRS) / (CRTS). Subsequent papers have considered alternative sets of assumptions, such as Banker, Charnes and Cooper (1984) who proposed a variable returns to scale (VRS) model. The following discussion of DEA begins with a description of the input-orientated CRS model, because this model was the first to be widely applied.

2.2.1.1 The Constant Returns to Scale Model (CRS)

The model begins by defining some notation. Assume there are data on K inputs and M outputs on each of N firms or Decision Making Units (DMUs) as they tend to be called in the DEA literature. For the i\textsuperscript{th} DMU these are represented by the vectors \( x_i \) and \( y_i \), respectively. The \( K \times N \) input matrix, \( X \), and the \( M \times N \) output matrix, \( Y \), represent the data of all \( N \) DMUs. The purpose of DEA is to construct a non-parametric envelopment frontier over the data points such that all observed points lie on or below the production frontier. For the simple example of an industry where one output is produced using two inputs, it can be visualized as a number of intersecting planes forming a tight fitting cover over a scatter of points in three-dimensional space. Given the CRS assumption, this can also be represented by a unit isoquant in input/input space (refer to Chart 2.2).

The best way to introduce DEA is via the ratio form. For each DMU we would like to obtain a measure of the ratio of all outputs over all inputs, such as \( u'y/v'x \), where \( u \) is an \( M \times 1 \) vector of output weights and \( v \) is a \( K \times 1 \) vector of input weights. To select optimal weights we specify the mathematical programming problem:
Max \(_{uv}(uy/vx_i)\),

\[
\begin{align*}
\text{st} & \quad uy_j/vx_j \leq 1, \quad j=1,2,...,N, \\
u, v & \geq 0.
\end{align*}
\]  

(10)

This involves finding values for \(u\) and \(v\), such that the efficiency measure of the \(i\)th DMU is maximized, subject to the constraint that all efficiency measures must be less than or equal to one. One problem with this particular ratio formulation is that it has an infinite number of solutions (That is, if \((u^*,v^*)\) is a solution, then \((\alpha u^*,\alpha v^*)\) is another solution, etc.). To avoid this one can impose the constraint \(v'x_i = 1\), which provides:

Max \(_{\mu v}(\mu'y_i)\),

\[
\begin{align*}
\text{st} & \quad v'x_i = 1, \\
\mu'y_j - v'x_j & \leq 0, \quad j=1,2,...,N, \\
\mu, v & \geq 0.
\end{align*}
\]  

(11)

Where the notation change from \(u\) and \(v\) to \(\mu\) and \(v\) reflects the transformation. This form is known as the multiplier form of the linear programming problem.

Using the duality in linear programming, one can derive an equivalent envelopment form of this problem:

Min \(_{a,\lambda}\theta, 

\begin{align*}
\text{St.} & \quad -y_i + \theta \lambda \geq 0, \\
\theta x_i - X\lambda & \geq 0, \\
\lambda & \geq 0.
\end{align*}
\]  

(12)

Where \(\theta\) is a scalar and \(\lambda\) is a \(N \times 1\) vector of constants. This envelopment form involves fewer constraints than the multiplier form (\(K+M < N+1\)), and hence is generally the preferred form to solve. The value of \(\theta\)
obtained will be the efficiency score for the i-th DMU. It will satisfy $\theta \leq 1$, with a value of 1 indicating a point on the frontier and hence a technically efficient DMU, according to the Farrell (1957) definition. The linear programming problem must be solved N times, once for each DMU in the sample. A value of $\theta$ is then obtained for each DMU.

2.2.1.2 Slacks

The piecewise linear form of the non-parametric frontier in DEA can cause a few difficulties in efficiency measurement. The problem arises because of the sections of the piecewise linear frontier which run parallel to the axes (Chart 2.2) which do not occur in most parametric functions (Chart 2.1). To illustrate the problem, refer to Chart 2.5 where the DMU's using input combinations C and D are the two efficient DMU's which define the frontier, and DMU's A and B are inefficient.

The Farrell (1957) measure of technical efficiency gives the efficiency of DMU's A and B as $O_A'/O_A$ and $O_B'/O_B$, respectively. However, it is questionable as to whether the point $A'$ is an efficient point since one could reduce the amount of input $x_2$ used (by the amount $CA'$) and still produce the same output. This is known as input slack (input excess) in the literature. Once one considers a case involving more inputs and/or multiple outputs, the diagrams are no longer as simple, and the possibility of the related concept of output slack also occurs. Thus it could be argued that both the Farrell measure of technical efficiency ($\theta$) and any non-zero input or output slacks should be reported to provide an accurate indication of technical efficiency of a DMU in a DEA analysis. Note that for the $i^{th}$ DMU the output slacks will be equal to zero only if $Y_{i\lambda} - y_i=0$, while the input slacks will be equal to zero only if $\theta x_i - X_{i\lambda}= 0$ (for the given optimal values of $\theta$ and $\lambda$).
In Chart 2.5 the input slack associated with the point A' is CA' of input \( x_2 \). In cases when there are more inputs and outputs than considered in this simple example, the identification of the "nearest" efficient frontier point (such as C), and hence the subsequent calculation of slacks, is not a trivial task. Some authors (see Ali and Seiford 1993) have suggested the solution of a second-stage linear programming problem to move to an efficient frontier point by MAXIMISING the sum of slacks required to move from an inefficient frontier point (such as A' in Chart 2.5) to an efficient frontier point (such as point C). This second stage linear programming problem may be defined by:

\[
\begin{align*}
\text{Min}_{\lambda, os, is} & \quad (M_1' OS + K_1' IS), \\
\text{st} & \quad -y_i + Y_\lambda - S = 0, \\
& \quad \theta X_i - X_\lambda - IS = 0, \\
& \quad \lambda \geq 0, \ OS \geq 0, \ IS \geq 0, 
\end{align*}
\]  

(13)
Where \( OS \) is an \( M \times 1 \) vector of output slacks, \( IS \) is a \( K \times 1 \) vector of input slacks, and \( M_1 \) and \( K_1 \) are \( M \times 1 \) and \( K \times 1 \) vectors of ones, respectively. In this second-stage linear program, \( \vartheta \) is not a variable; its value is taken from the first-stage results. Furthermore, note that this second-stage linear program must also be solved for each of the \( N \) DMUs involved.

There are two major problems associated with this second stage LP. The first and most obvious problem is that the sum of slacks is MAXIMISED rather than MINIMISED. Hence it will identify not the NEAREST efficient point but the FURTHEST efficient point. The second major problem associated with the above second-stage approach is that it is not invariant to units of measurement. The alteration of the units of measurement, say for a fertilizer input from kilograms to tones (while leaving other units of measurement unchanged), could result in the identification of different efficient boundary points and hence different slack and lambda measures.

However, that these two issues are not a problem in the simple example presented in Chart 2.5 because there is only one efficient point to choose from on the vertical facet. However, if slack occurs in 2 or more dimensions (which it often does) then the above-mentioned problems can come into play.

As a result of this problem, many studies simply solve the first-stage linear program (equation 12) for the values of the Farrell radial technical efficiency measures \( \vartheta \) for each DMU and ignore the slacks completely, or they report both the radial Farrell technical efficiency score \( \vartheta \) and the residual slacks, which may be calculated as \( OS = -y_j + Y\lambda, \) and \( IS = \vartheta x_j - X\lambda. \) However, this approach is not without problems either because these residual
slacks may not always provide all (Koopmans) slacks or hence may not always identify the nearest (Koopmans) efficient point for each DMU.

2.2.1.3 The Variable Returns to Scale Model (VRS) or Pure technical and Scale Efficiencies

The CRS assumption is only appropriate when all DMU's are operating at an optimal scale (i.e. one corresponding to the flat portion of the LRAC curve). Imperfect competition, constraints on finance, etc. may cause a DMU to be not operating at optimal scale. Banker, Charnes and Cooper (1984) suggested an extension of the CRS DEA model to account for variable returns to scale (VRS) situations. The use of the CRS specification when not all DMU's are operating at the optimal scale will result in measures of TE which are confounded by scale efficiencies (SE). The use of the VRS specification will permit the calculation of TE devoid of these SE effects.

The CRS linear programming problem can be easily modified to account for VRS by adding the convexity constraint: $N_1' \lambda = 1$ to (12) to provide:

$$\text{Min } \theta, \theta,$$

$$\text{st } -y + Y \lambda \geq 0,$$
$$\theta x - X \lambda \geq 0,$$
$$N_1 \lambda = 0$$
$$\lambda \geq 0$$

(14)

where $N_1$ is an $N \times 1$ vector of ones. This approach forms a convex hull of intersecting planes which envelope the data points more tightly than the CRS conical hull and thus provides technical efficiency scores which are greater than or equal to those obtained using the CRS model. The VRS specification has been the most commonly used specification in the 1990s.
2.1.4 Calculation of Scale Efficiencies

Many studies have decomposed the TE scores obtained from a CRS DEA into two components, one due to scale inefficiency and one due to "pure" technical inefficiency. This may be done by conducting both a CRS and a VRS DEA upon the same data. If there is a difference in the two TE scores for a particular DMU, then this indicates that the DMU has scale inefficiency, and that the scale inefficiency can be calculated from the difference between the VRS TE score and the CRS TE score.

Chart 2.6 attempts to illustrate this. In this chart we have a one-input one-output example and have drawn the CRS and VRS DEA frontiers. Under CRS the input-orientated technical inefficiency of the point P is the distance PP_C, while under VRS the technical inefficiency would only be PP_V. The difference between these two, PP_C - PP_V, is put down to scale inefficiency. One can also express all of this in ratio efficiency measures as:

\[
\text{TE}_{\text{CRS}} = \frac{AP_c}{AP} \\
\text{TE}_{\text{VRS}} = \frac{AP_v}{AP} \\
\text{SE}_i = \frac{AP_c}{AP_v}, \text{ where all of these measures will be bounded by zero and one. We also note that} \\
\text{TE}_{\text{CRS}} = \text{TE}_{\text{VRS}} \times \text{SE}_i
\]

because

\[
\frac{AP_c}{AP} = \left(\frac{AP_c}{AP_v}\right) \times \left(\frac{AP_v}{AP_v}\right).
\]

That is, the CRS technical efficiency measure is decomposed into "pure" technical efficiency and scale efficiency.
One shortcoming of this measure of scale efficiency is that the value does not indicate whether the DMU is operating in an area of increasing or the decreasing returns to scale. This may be determined by running an addition DEA problem with non-increasing returns to scale (NIRS) imposed. This can be done by altering the DEA model in equation 15 by substituting the \( N' A_r = 1 \) restriction with \( N' A_r < 1 \), to provide:

\[
\begin{align*}
\text{Min}_{\theta, \lambda, \theta}, \\
\text{St} \quad & -y_i + Y \lambda \geq 0, \\
& \theta x_i - X\lambda \geq 0, \\
& N1 \lambda = 0 \\
& \lambda \geq 0
\end{align*}
\]

(15)

The NIRS DEA frontier is also plotted in chart 2.6. The nature of the scale inefficiencies (i.e. due to increasing or decreasing returns to scale) for a particular DMU can be determined by seeing whether the NIRS TE score is equal to the VRS TE score. If they are unequal (as will be the case for the point 44...
P in Chart 2.6) then increasing returns to scale exist for that DMU. If they are equal (as is the case for point Q in chart 2.6) then decreasing returns to scale apply.

The DEA model can be categorized according to the type of data available (Cross sections or panel) and according to the type of variables (quantities only, or quantities and price). With quantities and prices OE can be calculated and decomposed into TE and AE. In our model as panel data is used, so the Malmquist efficiency Index is used. The Malmquist index is defined using distance function.

### 2.2.2 Malmquist Efficiency Index

A production process, which employs input vector $X^t$ to produce output vector $Y^t$ at time $t$, can be defined by using an output set (Shao and Shu 2002)\(^7\)

$$P^t(X^t) = \{Y^t: X^t \text{ can produce } Y^t\}$$  \hspace{1cm} (1)

The output set $P^t(X^t)$ is assumed to be closed, bounded and convex, and it satisfies the strong disposability of inputs and outputs (Coelli et al 1998)\(^8\). The output distance function is defined on the output set as:

$$D^t(X^t, Y^t) = \min_{\varphi \in \varphi} \{\varphi(Y^t/\varphi) \in P^t(x^t)\}$$  \hspace{1cm} (2)

$$= (\max_{\varphi \in \varphi} \{\varphi Y^t \in P^t(X^t)\})^{-1}$$  \hspace{1cm} (3)

The formulation in equation (2) indicates that the distance function $D^t$ is the inverse of the measure of Farrell’s output-oriented technical efficiency (Farrell 1957) and it facilitates the computation of an output distance function through the methods of efficiency measurement like DEA. The value of the output distance function ranges from 0 to 1, with a highest score indicating a location closer to the boundary of the output set (i.e., the production frontier).
In equation (4), the ratio outside the brackets is equal to the change of technical efficiency between time $t$ and $t+1$. In other words, it represents the change in the relative distance of the observed production from the maximum potential production. The component inside the brackets of equation (3) is the geometric mean of the two productivity indexes and represents the shift in production technologies (technological efficiency change) between time $t$ and $t+1$.

That is, technical efficiency change

$$
TEC = \frac{D^t(x^t, y^t)}{D^t(x^t, y^t)}
$$

Technological efficiency change

$$
TCH = \left( \frac{D^t(x^t, y^t)}{D^t(x^t, y^t)} \times \frac{D^{t+1}(x^{t+1}, y^{t+1})}{D^{t+1}(x^{t+1}, y^{t+1})} \right)^{1/2}
$$

Technical efficiency change (TEC) in equation (5) can be further decomposed as the product of two components—pure technical efficiency change and scale change—as follows (Fare et al 1994):
The ratio outside the brackets in equation (7) represents the pure change of technical efficiency, subject to a distance function \( D_r \) with variable returns to scale, between time \( t \) and \( t+1 \) and is denoted by PEC hereafter. In other words,

\[
P_{EC} = D_t'(x^t, y^t) \frac{D_{t+1}(x^{t+1}, y^{t+1})}{D_t(x^t, y^t)} \quad (8)
\]

The components inside the brackets of equation (7) represent the effects of economics of scale on productivity and are expressed as SCH. It is noted that SCH can be readily derived by dividing TEC of equation (5) by PEC of equation (8) and would not involve its own computation of additional output distance functions. That is, scale change.

\[
SCH = \frac{TEC}{PEC} \quad (9)
\]

After incorporating equation (7)-(9) into equation (3), we obtain the complete decomposition of the Malmquist Efficiency Index.

\[
M'(x^t, y^t, x^{t+1}, y^{t+1}) = \text{Technical Efficiency Change} \times \text{Technological Change} \times \text{(Catching up Effect)} \times \text{(Frontier Effect)}
\]

\[
= (PEC) \times (SCH) \times (TCH)
\]

47
Another merit of defining the MPI using the output distance function $D^j$ is that Malmquist Efficiency Index and its corresponding components (TCH, TEC, PEC and SCH) are all calculated in an index from and have a threshold value of one. In other words, if a derived value is equal to one, it indicates that a bank’s performance remains unchanged in that performance measure. A value greater than one represents an improvement, and a value less than one indicates a decline. The product of the index components of TCH, PEC and SCH then amounts to the final Malmquist Efficiency Index.
References


