Frontier analysis in healthcare

C.A. Knox Lovell
Department of Economics, University of Georgia, Athens, GA 30602, USA
E-mail: knox@terry.uga.edu

Abstract: Frontier techniques can be of immense use in establishing the best practice benchmarks against which the performance of healthcare providers can be evaluated. If credible benchmarks are set by frontier techniques, then reimbursement schemes can be guided by the performance of individual providers. Although frontier techniques have gained acceptance in other regulatory arenas, their potential remains unfulfilled in healthcare. This paper speculates on the strengths and weaknesses of the application of frontier techniques in healthcare reimbursement schemes, although the main objective is to introduce the readers to the basics of the techniques. Two techniques are considered; one is regression-based and the other is based on linear programming techniques. Each technique has its own relative advantage depending on the confidence generated by the underlying data or the ability to model a complex service delivery technology.

Keywords: frontier analysis; best practice; cost containment.


Biographical notes: C.A. Knox Lovell is Professor of Economics in the Terry College of Business, University of Georgia. He was the Editor-in-Chief of Journal of Productivity Analysis, and is co-author of Stochastic Frontier Analysis and Production Frontiers, Cambridge University Press. Actively involved with issues related to the measurement of efficiency and productivity change, he publishes and lectures on the subject internationally.

1 Introduction
The essence of frontier analysis is to construct a best practice frontier against which the performance of individual producers or service providers can be evaluated. Using mediocrity as a benchmark makes little sense, therefore the role of frontier analysis in identifying best practice standards is important to those in management and to those involved in the performance evaluation exercise. Since the best practice frontier is not an ideal or hypothetical frontier, but a conservative data-driven frontier established by best practice in the sample, the goals are not out of reach. In some settings, performance is evaluated so that it can be monitored for continuous improvement. In other settings performance is evaluated as an integral component of the budget allocation or reimbursement process.
The use of frontier analysis is widespread in incentive-based regulation of utilities – in which reimbursement is guided by the cost efficiency of service provision. In this context, among the proclaimed virtues of frontier analysis are its ability to deal with uncertainty about the structure of complex technology; and with agency problems in the presence of asymmetric information, unclear social priorities and conflicts of interest among the agents involved. In such an environment, the use of frontier analysis to benchmark against best practice standards provides utilities with incentives to improve their cost efficiency. The regulator then sets a reimbursement scheme that enforces a sharing of potential cost reductions between utilities and their customers. The potential for applying frontier analysis to healthcare in such an environment is explored (Agrell and Bogetoft, 2003).

The setting in which hospitals are reimbursed is structurally similar to that of the setting of revenue caps in utilities regulation. The same objective applies, namely cost containment, and the same constraints apply as well. The technology is certainly complex, information asymmetries abound, and the priorities of funding agencies, hospital administrators and physicians are often in conflict. The question that naturally arises is why frontier techniques are so popular in utilities regulation, yet so under-utilised in hospital reimbursement schemes. Targets are frequently set for improvements in cost efficiency, but these targets do not appear to be determined, or even influenced, by frontier analysis. The UK National Health Service experience with target setting is analysed (Jacobs and Dawson, 2003).

A number of arguments have been proffered against the use of frontier techniques in the hospital cost evaluation and reimbursement exercise. Newhouse (1994) notes extreme difficulties in the measurement of healthcare services, particularly of service quality, and concludes that ‘…reimbursement should not be fully prospective’. Skinner (1994) and Street (2003) note technical difficulties with one version of frontier analysis. Dor (1994) is concerned with model misspecification and the resulting fragility of inferences concerning the nature of best practice itself. It should be noted, however, that the concerns expressed (Newhouse, 1994; Dor, 1994) are general concerns that are not specific either to frontier analysis or to healthcare.

Without questioning the relevance of these and other concerns, it remains the case that frontier analysis plays a prominent role in the rate setting procedure in utilities regulation, where measurement and model specification issues are important. Moreover, utilities rate setting and hospital reimbursement are structured similarly, with the same objectives and the same constraints. It seems that the rich potential of frontier analysis is largely unfulfilled in healthcare. With this in mind, the objective of this paper is to briefly survey two frontier techniques, with an eye toward their relative strengths and weaknesses in an incentive-based reimbursement setting.

In Section 2, a generic approach to frontier analysis is developed, showing what it is capable of offering. In Sections 3 and 4 the essentials of a pair of alternative frontier techniques are discussed. The first is regression-based, popular in economics, and goes by the name of Stochastic Frontier Analysis (SFA). The second is based on linear programming, popular in management science and operations research, and goes by the name of Data Envelopment Analysis (DEA). Both are widely used in academic circles, and both are gaining credibility outside academe. Numerous academic studies have used frontier techniques to investigate hospital performance; see (Rodriguez-Alvarez and Lovell, 2004) for a recent example. At the micro level, some studies have used these techniques to evaluate physician performance (Chilingerian, 1995; Chilingerian and
Sherman, 1990) for example. At the macro level, frontier techniques have been used (World Health Organization, 2000) to evaluate the performance of the health sectors of World Health Organization member countries, and the WHO has shown continuing interest in extending and applying these techniques. Throughout Sections 2–4, the best practice frontier as a minimum cost frontier is specified, because the objective of the funding agency is typically one of cost containment, without sacrifice of either quantity or quality of service provision. In Section 5 some cautiously optimistic conclusions are drawn concerning the value of frontier analysis in the hospital reimbursement exercise.

2 Frontier analysis

Suppose we have data collected from a group of hospitals. Each hospital uses a vector \( x = (x_1, \ldots, x_N) \) of resources to provide a vector \( y = (y_1, \ldots, y_M) \) of services, the elements of which include both quantity and quality indicators. In the provision of its services each hospital incurs expense \( w^T x = \sum_{n=1}^{N} w_n x_n \), where \( w = (w_1, \ldots, w_N) \) is a vector of resource prices. The relationship between service provision and expense is illustrated in Figure 1 for a dozen hospitals, with scalar output representing the multidimensional service vector. The relationship is generally positive, although some hospitals provide more service at lower cost than some others do.

Figure 1  Relationship between service provision and expense

The objectives of frontier analysis are to uncover the nature of the relationship between service provision and expenditure, and to evaluate the performance of individual hospitals. Performance means the ability to minimise expenditure required to provide a service vector \( y \), in light of input price vector \( w \) and other exogenous variables represented by the vector \( z = (z_1, \ldots, z_K) \), whose elements characterise the operating environment.

Economic theory defines a minimum cost frontier as \( c(y, w, z) = \min_x \{ w^T x \mid (x, y) \in T_z \} \), where \( T_z \) is the set of technologically feasible combinations of resources and services in operating environment \( z \), and so
expression (1) states that actual expenditure is at least as great as minimum expenditure required to provide service vector \( y \) at resource prices \( w \) in operating environment \( z \). The minimum cost frontier \( c(y,w,z) \) expresses the desired nature of the relationship between service provision and minimum required expenditure. It provides the benchmark against which to evaluate the performance of individual hospitals. Their performance is evaluated in terms of their cost efficiency:

\[
CE(y,w,z,x) = \frac{c(y,w,z)}{w^Tx} \leq 1, \quad (2)
\]

with cost-efficient hospitals having \( CE(y,w,z,x) = 1 \) and the degree of cost inefficiency increasing as \( CE(y,w,z,x) \) declines.

In principle, reimbursement might be guided by \( CE(y,w,z,x) \). This performance criterion provides hospitals with an incentive to minimise the expenditure they incur in the provision of their service vector at the input price vector they face and in their operating environment.

The preceding analysis assumes that the minimum cost frontier \( c(y,w,z) \) is known. Unfortunately, in fact, it is not known. Therefore, considerable information asymmetries hinder the ability of funding agencies to calculate \( c(y,w,z) \), and hence \( CE(y,w,z,x) \). Two empirical approaches have been developed to uncover the structure of the minimum cost frontier, and to measure the cost efficiency of individual hospitals. One is regression-based, and is known as stochastic frontier analysis (SFA). The other uses linear programming techniques, and is known as data envelopment analysis (DEA). Both approaches construct best practice minimum cost frontiers, and so cost efficiency is measured relative to the most cost-efficient hospitals in the sample, rather than relative to some theoretical standard.

3 Stochastic frontier analysis

As its name suggests, SFA is based on the twin beliefs in the presence of both statistical noise and inefficiency. Consequently, a regression model incorporating these beliefs requires two error components, and so (1) is converted to an equality in the stochastic cost frontier regression model:

\[
w^Tx = c(y,w,z) \cdot \exp\{v+u\}, \quad (3)
\]

(3) states that actual expenditure \( w^Tx \) equals minimum required expenditure \( c(y,w,z) \) times the product of two error components. The first, \( \exp\{v\} \), captures the impact of statistical noise, reflecting random events beyond the control of the hospital, and the second, \( \exp\{u\} \), captures the magnitude of cost inefficiency. Since statistical noise can be unfavourable or favourable, \( v \geq 0 \), and \( \exp\{v\} \geq 1 \) raises or lowers expenditure. Since cost inefficiency can only raise expenditure, \( u \geq 0 \), and \( \exp\{u\} \geq 1 \) raises expenditure.

The deterministic kernel of a stochastic cost frontier is depicted in Figure 2, using the data from Figure 1. The second error component \( \exp\{u\} \geq 1 \) causes \( w^Tx \geq c(y,w,z) \cdot \exp\{v\} \) and leads to the expectation that cost inefficiency also causes \( w^Tx \geq c(y,w,z) \), and this is generally the case. However, the first error component \( \exp\{v\} \geq 1 \) allows for
the possibility that favourable statistical noise might dominate cost inefficiency, causing $w^T x \leq c(y,w,z)$, and this happens for two hospitals. The interpretation of the second case is that relatively low cost may be more a reflection of a favourable operating environment than of an ability to minimise cost in that environment.

**Figure 2** Deterministic kernel of a stochastic cost frontier

In this stochastic framework the cost efficiency of each hospital as defined in (2) is estimated by:

$$CE(y,w,z,x) = \frac{c(y,w,z) \cdot \exp\{v\}}{w^T x} = \exp\{-u\} \leq 1,$$

which makes it clear that stochastic frontier analysis incorporates the impacts of statistical noise in the estimation of cost efficiency. The trick is to disentangle the separate impacts of noise, reflecting random events not under the control of hospital management, and cost inefficiency, reflecting suboptimal resource allocation decisions on the part of management, in the composed error term $\exp\{v+u\}$.

Several estimation and decomposition strategies are available, particularly if one has panel data; details for both cross-sectional and panel data environments are provided (Kumbhakar and Lovell, 2000). Here, the application of maximum likelihood techniques to cross-sectional data is considered. The first step is to endow $c(y,w,z)$ with a functional form, preferably a flexible functional form capable of capturing the nature of economies of scale and diversification, as well as the nature of input substitutability and complementarity. The parametric cost frontier is written as $c(y,w,z;\beta)$, where $\beta$ is a parameter vector attached to $(y,w,z)$ that characterises the structure of the cost frontier. The second step is to make distributional assumptions on the two error components; this is the key to disentangling their separate impacts. A tractable and popular set of assumptions is:

- $v \sim iid N(0,\sigma_v^2)$
- $u \sim iid N(0,\sigma_u^2)$
- $v$ and $u$ are distributed independently of each other, and of the regressors
which state that statistical noise \( v \geq 0 \) is normally distributed around zero with constant variance \( \sigma_v^2 \), that cost inefficiency \( u \geq 0 \) is distributed as the non-negative half of a normal distribution having zero mean and constant variance \( \sigma_u^2 \), and finally that \( v, u \) and the regressors are independently distributed. The half-normal assumption on \( u \) is simply a parsimonious parameterisation of the basic idea that larger values of cost inefficiency are less likely than smaller values of cost inefficiency. These assumptions introduce two additional parameters, \( \sigma_v^2 \) and \( \sigma_u^2 \), to be estimated along with the elements of the parameter vector \( \beta \).

With these distributional assumptions, the mean and variance of the composed error term are:

\[
E(v+u) = \left( \frac{2}{\pi} \right) \sigma_u, \quad V(v+u) = \left( \frac{\pi - 2}{\pi} \right) \sigma_u^2 + \sigma_v^2. \tag{5}
\]

The expected value of the error term is positive unless there is no variation in cost inefficiency, in which case maximum likelihood reduces to ordinary least squares in which \( E(v+u) = 0 \) and \( V(v+u) = \sigma_v^2 \). Indeed OLS provides a simple test for the presence of inefficiency in the data. If \( u = 0 \), the error term is symmetric, and the data do not support a cost inefficiency story. However if \( u > 0 \), then \( v+u \) is positively skewed, and there is evidence of cost inefficiency in the data. This suggests that a test for the presence of cost inefficiency can be based directly on the OLS residuals. An appropriate test statistic is:

\[
(b_1)^{1/2} = m_3 / (m_2)^{3/2}, \tag{6}
\]

where \( m_2 \) and \( m_3 \) are the second and third sample moments of the OLS residuals. Since \( v \) is symmetrically distributed, \( m_3 \) is simply the third sample moment of \( u \). Thus \( m_3 > 0 \) implies that the OLS residuals are positively skewed, and suggests the presence of cost inefficiency, while \( m_3 < 0 \) implies that the OLS residuals are negatively skewed, which makes no sense in this context. Thus, the negative skewness in the OLS residuals provides an indication that the model is mis-specified, and cannot be used to estimate cost efficiency.

Recalling (4), a point estimate of the cost inefficiency of an individual hospital is provided by:

\[
CE_i = E(exp\{-u_i|\epsilon_i\}) = \left[ \frac{1 - \Phi(\sigma_* - \mu_i/\sigma_*)}{1 - \Phi(-\mu_i/\sigma_*)} \right] \cdot \exp\left\{ -\mu_i + \frac{1}{2} \sigma_*^2 \right\}, \tag{7}
\]

where \( \epsilon_i = v_i + u_i \), \( \sigma_* = \sigma_u \sigma_v / \sigma_v \), \( \sigma_v = (\sigma_u^2 + \sigma_v^2)^{1/2} \), \( \mu_i = \epsilon_i \sigma^v / \sigma_v^2 \), \( \Phi(\cdot) \) is the standard normal cumulative distribution function, and the subscript \( i = 1, \ldots, I \) indexes individual hospitals. \( CE_i \), as defined is a conditional expectation, the expected value of \( \exp\{-u_i\} \) given the error term \( \epsilon_i = v_i + u_i \) and given the distributional assumptions made on \( v \) and \( u \). It is these distributional assumptions that enable one to decompose the composed error term.

Although, to report confidence intervals around point estimates of \( \beta \) is the standard procedure, it is unusual to report confidence intervals around point estimates of cost efficiency as defined in (7). Nonetheless it is possible, and desirable, to do so. A \((1-\lambda)\)100% confidence interval \((L_i, U_i)\) for the point estimate \( CE_i \) is provided by:

\[
L_i = \exp\{-\mu_i - q_{1-\lambda} \sigma_*\}, \quad U_i = \exp\{-\mu_i - q_{\lambda} \sigma_*\}, \tag{8}
\]
where
\[ Pr(Q > q_L) = \left( \frac{\lambda}{2} \right) \left[ 1 - \Phi \left( -\frac{\mu_i}{\sigma} \right) \right] \] (9)
\[ Pr(Q > q_u) = \left( 1 - \frac{\lambda}{2} \right) \left[ 1 - \Phi \left( -\frac{\mu_i}{\sigma} \right) \right] \]
and \( Q \) is distributed as \( N(0,1) \). Consequently
\[ q_L = \Phi^{-1} \left[ \left( 1 - \frac{\lambda}{2} \right) \left[ 1 - \Phi \left( -\frac{\mu_i}{\sigma} \right) \right] \right] \] (10)
\[ q_U = \Phi^{-1} \left[ \left( 1 - \left( 1 - \frac{\lambda}{2} \right) \left[ 1 - \Phi \left( -\frac{\mu_i}{\sigma} \right) \right] \right) \right] \]

With this information at hand, reimbursement might be guided by \( CE_i \), which provides a point estimate of the ability of each individual hospital to contain cost relative to best practice observed in the sample, and tempered by the confidence interval \((L_i, U_i)\). This strategy provides hospitals with incentives to contain cost, although this guidance is tempered by the relative importance of statistical noise in each hospital’s error term. Limited experience suggests that confidence intervals frequently overlap, and taking them into account makes it difficult to confidently distinguish hospitals on the basis of their cost efficiency, except at the tails of the distribution. The reason is that it is difficult to estimate \( \sigma_v^2 \) and \( \sigma_u^2 \) with the sort of precision required to generate narrow confidence intervals. An illuminating illustration, based on a sample of English public hospitals and using somewhat less sophisticated techniques, appears in Street (2003).

Two final points warrant mention. First, in the presence of panel data the assumption of unchanging technology is untenable, especially in healthcare. Improvements in medical and organisational technology can be proxied by introducing a time indicator into the cost frontier, which becomes \( c(y,w,z,t;\beta) \), where \( t \) is either a time trend or a set of time dummy variables. Second, the question arises as to whether the characteristics of the operating environment should be introduced into the cost frontier, as it has been done here, or into the half-normal error component to create \( u = u(z) \). In the former case, it is assumed that \( z \) influences the cost of transforming resources into services; in the latter case, it is assumed that \( z \) influences the cost efficiency with which resources are transformed into services. Of course \( z \) can be introduced into both the cost frontier and the half-normal error component, after which likelihood ratio tests of their placement can be conducted. An advantage of placing at least some elements of \( z \) into the half-normal error component is that this procedure may mitigate the impacts of heteroskedasticity, a violation of the assumption that both \( \sigma_v^2 \) and \( \sigma_u^2 \) are constant. This is important because heteroskedasticity is a common problem in the estimation of cost functions, and it creates even more serious problems in the estimation of cost frontiers because it leads to biased estimates of \( CE_i \), which is the whole point of the exercise. The second point also applies to the placement of the time indicator. Writing \( c(y,w,z,t;\beta) \) enables \( t \) to capture the impact on cost of changes in technology, while writing \( u = u(z,t) \) enables \( t \) to capture the impact on cost of changes in cost efficiency. Both points illustrate the importance of model specification alluded to above.

### 4 Data envelopment analysis

Data envelopment analysis (DEA) was developed as a performance measurement technique, primarily for use in the public, service, and not-for-profit sectors in which
price information is either missing or unreliable and managerial objectives are unclear. DEA is actually a family of dual linear programming problems, one for each type of performance under evaluation. This paper considers the DEA formulation of a cost minimisation problem because it is consistent with the SFA cost minimisation problem analysed in Section 3. DEA formulations of other optimisation problems, and refinements of the basic primal and dual models, are available in (Cooper, Seiford and Tone, 2000).

Extending the notation introduced in Section 2, let $x = (x_1, \ldots, x_N)$, $w = (w_1, \ldots, w_N)$, $y = (y_1, \ldots, y_M)$ and $z = (z_1, \ldots, z_K)$ be the resource vector used, the resource price vector faced, the service vector provided and a vector of characteristics of the operating environment faced, by a hospital under evaluation. Assuming that there are I hospitals in the sample, let $X$ be an $NxI$ sample matrix of column vectors of hospital resources, let $Y$ be an $MxI$ sample matrix of column vectors of hospital outputs, and let $Z$ be a $KxI$ sample matrix of column vectors of hospital operating environment characteristics. Then the DEA formulation of the cost minimisation problem is:

$$\min_{x^*} w^T x^*$$

subject to

$$x^* \geq X\lambda$$
$$Y\lambda \geq y$$
$$z \leq Z\lambda$$
$$\lambda \geq 0, \sum \lambda_i = 1.$$ 

The program asks the hospital under evaluation to select a resource vector $x^*$ that may differ from its actual resource vector $x$, and that minimises expenditure, subject to the constraints that the resulting resource-service-operating environment vector $(x^*, y, z)$ not violate best practice standards established in the sample. Best practice standards are established by existing hospitals whose data are contained in $(X, Y, Z)$ and convex combinations of them. Such convex combinations have resource use not greater than $x^*$, service provision not less than $y$ and operating environment no more favourable than $z$. The elements of the vector $\lambda$ are intensity variables that serve to form the minimum cost frontier from convex combinations of the $(X, Y, Z)$ data. The convexity constraint $\sum \lambda_i = 1$ allows for variable returns to scale in production. The minimum cost frontier is established by all hospitals having $x^* = x$, and all convex combinations of them. The program is solved I times, once for each hospital in the sample.

Using the same 12 hospitals as in Figures 1 and 2, Figure 3 illustrates the DEA formulation of the minimum cost frontier. This formulation constructs the tightest fitting piecewise linear surface that envelops the cost-output combinations from below, plus horizontal and vertical extensions. In Figure 3 there are four cost-efficient hospitals, and the cost efficiency of any other hospital is calculated as the ratio of the minimum cost of providing its service vector to its actual expenditure. Thus for any hospital, cost efficiency is calculated as:

$$CE(y, w, z, x) = w^T x^* / w^T x \leq 1.$$ 

(11)

Once again, benchmarking against best practice provides hospitals with incentives to contain cost, and reimbursement can be guided by values of $CE(y, w, z, x)$ for each hospital.
The DEA model can be refined in a number of ways, one of which is of particular relevance in a reimbursement context. Eliminating the hospital under evaluation from \((X,Y,Z)\) creates a minimum cost frontier that is formed from all other hospitals. This refinement has no impact for hospitals that are cost-inefficient under the original formulation. However, it allows hospitals that are cost-efficient under the original formulation to become ‘super-efficient’. The advantage of using this refinement to guide reimbursement is that it creates an incentive for hospitals not just to strive to reach best practice, but to improve on it. The super-efficiency DEA model is widely employed in incentive regulation of electricity distribution.

An important feature of both the DEA models is that they are ostensibly deterministic, since they do not account explicitly for statistical noise. Consequently, it is not immediately obvious that guidance provided by CE\((y,w,z,x)\) can be tempered by confidence intervals, as in SFA. However, DEA is not without statistical foundations, and in large samples these foundations permit hypothesis testing in much the same way as SFA does. However, in samples of modest size and most likely to be encountered, at present there seems to be no alternative but the bootstrap as a way of generating confidence intervals for point estimates of cost efficiency.

5 Conclusions

Frontier analysis is widely used in academic studies of hospital performance evaluation. It is also popular in the setting of revenue caps in incentive-based regulation of utilities. Therefore, it is surprising that it has not migrated from any one of the applications to assist in hospital reimbursement schemes, particularly since the setting of revenue caps for regulated utilities and the design of incentive-compatible reimbursement schemes for hospitals are similarly structured economic problems. In both problems, funding agencies are confronted with information asymmetries and agency problems, with the attendant conflicts of interest among the agents involved. In both problems the objective is, or should be, to benchmark against the best rather than against the mediocre. Yet frontier analysis has gained acceptance in utilities regulation, but not in hospital reimbursement.
In an effort to enhance the migration process, a brief description of two prominent empirical techniques of frontier analysis has been provided. Like all empirical techniques, both require reliable data and a proper model specification. Unlike other techniques, both are capable of uncovering best practice, rather than ‘average’ practice, relationships. Both support statistical inference on the nature of the best practice frontier and on the distance of individual hospitals from it. In terms of comparative advantage, SFA is more appropriate when uncertainty about the data dominates uncertainty about the technology, and DEA is more appropriate in the opposite circumstance.

In conclusion, in concurrence with the cautious optimism expressed by Hadley and Zuckerman (1994) and more recently by Agrell and Bogetoft (2003), it is stated that frontier analysis can play an important role in the design and implementation of hospital reimbursement schemes, provided that agents cooperate on data collection and model specification, and provided that confidence intervals are assigned the same priority as point estimates of cost efficiency. Given these provisos, information gleaned from a carefully conducted frontier analysis of hospital cost efficiency can provide useful insights, if not sharp guidance, to those responsible for designing and implementing reimbursement schemes. Most significantly, the use of frontier analysis puts in place incentives for cost containment.

References


