Discourses

EFFICIENCY AND PRODUCTIVITY ANALYSIS IN HEALTH SERVICES

Martin Dlouhý

Abstract:
Efficiency and productivity analysis is widely used in health services research. In the paper, two major modelling approaches to efficiency and productivity analysis in health services are evaluated. The first group comprises non-parametric methods based on the theory of mathematical programming, such as data envelopment analysis. The second group originates from econometric analysis of production function. Although the application of quantitative economic analysis in health services is not without problems, the methods discussed are able to help us with identifying best and worst practices within the health systems.

Keywords: data envelopment analysis, frontier analysis, health services

JEL classification: C50, C60, D24

1. Introduction

The purpose of this study is to demonstrate how the quantitative economic analysis can be successfully applied to efficiency and productivity measurement in health services. I argue that quantitative economic analysis is an important tool for understanding the health services at both the microeconomic as well as macroeconomic levels. On the other hand, it is also necessary that a researcher is aware of the limits. One health services researcher once stated: each country’s health system is a complex adaptive human system; it is not a table or a point on a regression line.

To produce health services, we need physicians, nurses, buildings, health technology, and other medical and non-medical inputs. This technological relationship is described by a production function, which is an abstract model of a production process. The model of production enables researchers to study key characteristics of a production process without going into technical details. The production frontier represents the maximum amount of output that can be produced by given amount of input (in the output-
maximization model) or, alternatively, the minimum amount of input required to produce the given amount of output (in the input-minimization model). The typical inputs and outputs used in health care studies are summarised in Table 1.

Table 1: Health Inputs and Outputs

<table>
<thead>
<tr>
<th>INPUTS</th>
<th>OUTPUTS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>National Health System</strong></td>
<td></td>
</tr>
<tr>
<td>Total Number of Physicians</td>
<td>Population Served</td>
</tr>
<tr>
<td>Total Number of Beds</td>
<td>Life Expectancy</td>
</tr>
<tr>
<td>Health Expenditures</td>
<td></td>
</tr>
<tr>
<td><strong>Hospital</strong></td>
<td></td>
</tr>
<tr>
<td>Number of Physicians</td>
<td>Number of Discharges (Case-Mix Adjusted)</td>
</tr>
<tr>
<td>Number of Nurses</td>
<td>Number of Inpatient Days</td>
</tr>
<tr>
<td>Drug Expenses</td>
<td>Number of Outpatient Visits</td>
</tr>
<tr>
<td>Capital Expenses</td>
<td>Number of Surgeries</td>
</tr>
<tr>
<td>Number of Beds (Acute, Long-tem, ICU)</td>
<td>Number of Residents Trained</td>
</tr>
</tbody>
</table>

I am distinguishing two major groups of methods of efficiency analysis: non-parametric and parametric methods. The first group comprises non-parametric methods based on the theory of mathematical programming, such as Data Envelopment Analysis (DEA) and Free Disposable Hull (FDH); the second group originates from econometric analysis of production function (Dlouhý, 2007). Both groups of methods are widely used in health care (Hollingsworth, 2003; O’Neill, Rauner, Heidenberger, Kraus 2008). Hollingsworth (2003) reviewed 188 papers on the frontier efficiency measurement in health care. The review mainly included DEA-based methods that dominated the literature, but there were an increasing number of parametric techniques, such as stochastic frontier analysis. DEA alone was used in 50% of studies; a quarter of studies used regression analysis in two stage analysis, typically to regress factors on the efficiency scores in an attempt to determine influences on efficiency. Stochastic frontier analysis and other parametric frontier techniques were used in 12% of studies; Malmquist techniques were used in 9% of studies. O’Neill, Rauner, Heidenberger and Kraus (2008) reviewed 79 DEA hospital efficiency studies published from 1984–2004 that represent 12 countries. Their cross-national comparison reveals significant differences with respect to important study characteristics such as type of DEA model selected and choice of input and output categories. Compared with the U.S. studies, European efforts are more likely to measure allocative rather than technical efficiency, use longitudinal data, and use fewer observations.

2. Nonparametric Methods

Farrell (1957) promoted the ideas to specify the production frontier as the most pessimistic piecewise linear envelopment of the data and to construct efficiency
measures based on radial uniform contractions or expansions from inefficient observations to the frontier. The success story of DEA began with the publication of the highly influential paper by Charnes, Cooper and Rhodes (1978). In DEA, the relative technical efficiency of a production unit is defined as the non-monetary ratio of its total weighted output to its total weighted input or, vice versa, as the ratio of its total weighted input to its total weighted output. The method allows each unit to choose its own weights of inputs and outputs in order to maximize its efficiency score. For each production unit, DEA (a) calculates the efficiency score; (b) determines the relative weights of inputs and outputs; and (c) identifies peers for each unit that is not technically efficient. The peers of an inefficient unit are technically efficient units with similar combinations of inputs and outputs. The peers serve as benchmarks, which show potential improvements that the inefficient unit can attain. For the computation, it is useful to consider a dual DEA model. The dual model involves fewer constraints than the primal model, and therefore, the dual model is the preferred model to solve. The formulation of the dual model in the matrix form is:

\[
\begin{align*}
\text{minimize} & \quad \theta_q - \varepsilon (e^T s^* + e^T s^-), \\
\text{subject to} & \quad X\lambda + s^- = \theta_q x_q, \\
& \quad Y\lambda - s^+ = y_q, \\
& \quad \lambda \geq 0, s^+ \geq 0, s^- \geq 0,
\end{align*}
\]

where \( \theta_q \) is the efficiency score, \( \lambda \) is the vector of variables, \( s^+, s^- \) are vectors of slack variables, \( x_q \) and \( y_q \) are vectors of inputs and outputs of the evaluated unit, \( X \) is a matrix of \( m \) inputs and \( Y \) is a matrix of \( r \) outputs, \( e^T = (1, 1, \ldots, 1) \), and \( \varepsilon \) is infinitesimal constant. The production unit \( q \) is technically efficient if the optimal value of variable \( \theta^*_{q} \) is one and all slack variables \( s^*_i \) and \( s^-_i \) equal zero (the symbol “\( * \)” denotes the optimal value). The output-oriented model version of the model is written as

\[
\begin{align*}
\text{maximize} & \quad \phi_q - \varepsilon (e^T s^* + e^T s^-), \\
\text{subject to} & \quad X\lambda + s^- = x_q, \\
& \quad Y\lambda - s^+ = \phi_q y_q, \\
& \quad \lambda \geq 0, s^+ \geq 0, s^- \geq 0,
\end{align*}
\]

where \( \phi_q \) is the efficiency score. The meaning of all other parameters and variables is the same as in the input-oriented model. A production unit \( q \) is technically efficient if the optimal value of \( \phi^*_{q} \) is one and all slack variables \( s^*_i \) and \( s^-_i \) equal zero. According to the returns to scale, the theory distinguishes four alternative DEA models: the constant-returns-to-scale model, the variable-returns-to-scale model, the non-increasing-returns-to-scale-model, and the non-decreasing-returns-to-scale model. These models differ in restrictions for the vector \( \lambda \).

Free Disposable Hull (FDH) was formulated by Deprins, Simar, and Tulkens (1984). Let us have \( n \) units, which consumes \( m \) inputs and produce \( r \) outputs. A mathematical formulation of the input-oriented FDH model follows:
minimize \( \theta_q - \varepsilon (e^T s^+ + e^T s^-) \),

subject to

\[ \begin{align*}
X\lambda + s^- &= \theta x_q, \\
Y\lambda - s^+ &= y_q, \\
e^T \lambda &= 1, \quad \lambda_k \text{ - binary, } s^+ \geq 0, s^- \geq 0,
\end{align*} \]

where \( Y \) is the matrix of outputs \((r, n)\), \( X \) is the matrix of inputs \((m, n)\), \( \lambda \) is the vector of variables, \( s^+ \) a \( s^- \) are the vectors of slack variables, \( e^T = (1,1,...,1) \), and \( \varepsilon \) is infinitesimal constant. The value of objective function is the technical efficiency. The FDH efficiency scores are higher than the DEA scores, because a production unit cannot be dominated by a convex combination of other units in the FDH model. Notice that the FDH model differs from convex DEA models with the constraint for the vector \( \lambda \). Advocates of FDH argue that there are no a priori economic grounds to believe that the production frontier is convex or non-convex, and that this flexibility is the advantage of FDH, which does not require any a priori assumption about the returns to scale. The model closely related to FDH is the Free Replicability Hull (FRH) that allows only integer replications of observed production units. The input-oriented FRH model is formulated as

minimize \( \theta_q - \varepsilon (e^T s^+ + e^T s^-) \),

subject to

\[ \begin{align*}
X\lambda + s^- &= \theta x_q, \\
Y\lambda - s^+ &= y_q, \\
\lambda_k &\in \mathbb{Z}^+ \text{ (non-negative integer)}, s^+ \geq 0, s^- \geq 0.
\end{align*} \]

The output-oriented FRH model can be formulated analogically. The FDH and FRH models can be used for evaluation of technical efficiency in the health sector as DEA, however, applications of DEA dominate the literature. For example, Novosádová and Dlouhý (2007) analysed technical efficiency of 119 Czech hospitals in 2005 by data envelopment analysis. The inputs were the number of physicians, the number of nurses and the number of beds. The outputs were the number of discharges and the number of inpatient days. They analysed the sample with the input-oriented constant-returns-to-scale model and the input-oriented variable-returns-to-scale model. The authors found no significant correlation between technical efficiency scores and wages of physicians and nurses.

### 3. Econometric Analysis

Econometric analysis is a scientific discipline that is based on an amalgam of economic theory, mathematical economics, economic statistics, and mathematical statistics. A main tool of econometrics is the regression analysis that studies the dependence of one variable on one or more other variables. An econometric model is formulated in the form of one or more equations. The Cobb-Douglas production function, the most popular production function, can be expressed as

\[ \text{Econometric analysis} = \text{model + regression + statistics} \]
\[ y = a X_1^\alpha_1 X_2^\alpha_2 \cdots X_k^\alpha_k e^u, \]  
(1)

where \( y \) is the output, \( x_j \) are the inputs, \( u \) is the stochastic error term, and \( e \) is the base of natural logarithm. The interpretation of parameters in the Cobb-Douglas production function is as follows: parameter \( a \) is interpreted as a technical level of production process; parameter \( \alpha_j \) is the partial elasticity of output to input \( j \); the sum of parameters \( \alpha_j \) determines the returns to scale. The shortcoming of classical regression approach is that it concentrates on the estimation of average behaviour, not on the best practice. In theory, a production function is defined as a maximum output obtainable from given resources, not as an average or usual level of output. The classical regression analysis of production function was, therefore, extended to the frontier analysis (Kumbhakar, Lovell, 2000). Let us assume that production model can be written as

\[ y_i = f(x_i; \beta) \cdot TE_i, \]  
(2)

where \( y_i \) is the scalar output of \( i \)th unit (producer, hospital), \( x_i \) is the vector of \( n \) inputs used by \( i \)th unit, \( f(x_i; \beta) \) is the production frontier, \( \beta \) is the vector of unknown technology parameters to be estimated, and \( TE_i \) is the output-oriented technical efficiency of unit \( i \). The output-oriented technical efficiency \( TE_i \) is defined as the ratio of observed output \( y_i \) to the maximum feasible output, which is determined by the production frontier \( f(x_i; \beta) \). We have

\[ TE_i = \frac{y_i}{f(x_i; \beta)}. \]  
(3)

If the observed unit is technically efficient, then technical efficiency \( TE_i = 1 \); if the unit is not technically efficient, then \( TE_i < 1 \). This type of the frontier model is called the deterministic frontier analysis (DFA) because the difference between the observed output \( y_i \) and maximum feasible output \( f(x_i; \beta) \) is explained by the technical inefficiency. There is no stochastic term, which could represent random shocks in production. In comparison, the stochastic frontier analysis (SFA) assumes the existence of random shocks that affect efficiency of the production process. The deterministic model is modified to a stochastic model as

\[ y_i = f(x_i; \beta) \cdot \exp\{v_i\} \cdot TE_i, \]  
(4)

where \( f(x_i; \beta)\exp\{v_i\} \) is the stochastic production frontier, which consists of two parts: the deterministic part \( f(x_i; \beta) \) and the stochastic part \( \exp\{v_i\} \), where \( v_i \) is the two-sided random component with the normal distribution \( N(0, \sigma^2_v) \). The output-oriented technical efficiency is given by the ratio

\[ TE_i = \frac{y_i}{f(x_i; \beta) \exp\{v_i\}}. \]  
(5)
Similarly as in the deterministic frontier model, the production unit is technically efficient if \( TE = 1 \) and inefficient if \( TE_i < 1 \). Advocates of the stochastic frontier analysis argue that it is not possible to ignore random shocks as the deterministic frontier analysis proposes, and therefore, the stochastic frontier analysis should be preferred. One of the possible estimation methods of deterministic frontier analysis is known as the Corrected Ordinary Least Squares (COLS). This method estimates production frontier in two steps. In the first step, the ordinary least squares method is applied to obtain consistent and unbiased estimates of the slope parameters \( \beta_j \) and consistent but biased estimate of intercept \( \beta_0 \). In the second step, the intercept is “corrected” by adding a residual with maximal value:

\[
\hat{\beta}_0^* = \hat{\beta}_0 + m a x\{ \hat{u}_i \}.
\] (6)

The production unit with a maximum value of residual \( \hat{u}_i \) is from the definition efficient and lies on the production frontier. Residuals of OLS regression are corrected by:

\[
-\hat{u}_i^* = \hat{u}_i - m a x\{ \hat{u}_i \}.
\] (7)

The technical efficiency of the production unit is calculated by \( TE_i = \exp(-\hat{u}_i^*) \). The main advantage of COLS is its simplicity; the disadvantages of the method are ignoring the stochastic term and assuming that the efficient production frontier is parallel to the OLS regression (Kumbhakar, Lovell, 2000).

Methods of the SFA assume that a difference between the observed and the feasible output can be partitioned into two components, one representing technical inefficiency and one representing effects of random shocks (Meeusen, van den Broeck, 1977; Aigner et al., 1977). If we assume the case of the Cobb-Douglas production function then we can write the stochastic frontier analysis model as

\[
\ln y_i = \ln \beta_0 + \sum_{j=1}^{k} \beta_j \ln x_{ij} + v_i - u_i ,
\] (8)

where \( v_i \) is the two-sided random component with the normal distribution \( N(0, \sigma_v^2) \) and \( u_i \) is the non-negative technical efficiency component of error term with the non-negative half normal distribution \( N(0, \sigma_u^2) \). The error components \( u_i \) and \( v_i \) are independent of each other, and of the regressors. The composed error term \( \epsilon_i \) is defined as \( v_i - u_i \). The composed error term is asymmetric since \( u_i \geq 0 \). The distribution of \( u_i \) is usually assumed to be non-negative half normal, truncated normal, exponential or gamma (Kumbhakar, Lovell, 2000). Such restrictive assumptions about the distribution of \( u_i \) may be considered as a weakness of SFA.

If a set of observations of the same units over several time periods is available, methods of panel data analysis can be used. Naturally, the panel data must contain more information than the cross-sectional data because the panel data include a time dimension. The analysis of panel data is a special part of econometrics that contains a variety of advanced econometric models (e.g., Baltagi, 2001). Moreover, the panel data analysis can be combined with frontier analysis. I will present here only one
panel data model by which efficiency of production can be estimated: the fixed-effects model (FEM). Its main advantage is its relative simplicity. The term fixed effects is due to the fact that, although the intercepts may differ across individuals, each individual’s intercept does not vary over time; it is time invariant. The fixed-effects model differs from the frontier analysis in the way the efficiency is estimated. The frontier analysis estimates the efficiency from the error term, whereas the fixed-effects model introduces dummy variables that measure firm effects and time efficiency effects. The firm effect is interpreted as the level of efficiency. The approach of FEM has an advantage over the frontier analysis in the fact that no assumptions on the distribution of efficiency have to be made. Because FEM is not a frontier approach, production units are not necessarily evaluated in terms of best practice. Assuming the Cobb-Douglas production function, the FEM model is specified as

\[ \ln y_{it} = \ln \beta_0 + \sum_{j=1}^{k} \beta_j \ln x_{ijt} + \sum_{i=2}^{n} \alpha_i z_i + \sum_{t=2}^{T} \lambda_t w_t + u_{it}, \]  

(9)

where the variables \( x_{ijt} \) are the inputs (production unit \( i \), input \( j \), time \( t \)), the binary variables \( z_i \) and \( w_t \) identify individual production units and time periods, \( n \) is the number of production units (firms), \( T \) is the number of time periods, coefficients \( \alpha_i \) are the firm effects and coefficients \( \lambda_t \) are the time effects, and \( u_{it} \) is the two-sided error term. The firm effects, i.e. the relative efficiencies of production units, are constant over the period; the time effects capture the changes in overall efficiency in the industry. Bryce, Engberg and Wholey (2000) applied the fixed-effects model for the efficiency evaluation of health maintenance organizations (HMOs). They analysed a data set containing 585 HMOs operating from 1985 through 1994. Because of new entry, mergers, and failures, the panel was unbalanced. The output was defined as total member-years of coverage provided by an HMO during the year. The inputs used by HMOs consist of hospital days, ambulatory visits, administrative expenditures, and other expenditures. In their study, Bryce et al. compared three methods of efficiency analysis: data envelopment analysis, stochastic frontier analysis, and the fixed-effects model. They computed three estimates of technical efficiency for each observation in the sample. All three methods identified similar trends for the HMO industry as a whole; however, the methods differed in assessing relative technical efficiency of individual health maintenance organizations. Thus, these techniques are limited for either benchmarking or setting rates because the units identified as technically efficient may be a consequence of model selection rather than actual performance.

4. Discussion

Efficiency evaluation in the public sector (including health services) is a theoretical as well as a practical problem. Health care is an application field with specific context and characteristics: final output (health improvement) is hard to measure and hard to express in monetary values; the causality between input and output is not always certain; apart from the efficiency of the system, the equity and moral principles have
to be borne in mind; there is usually a high variation in demand for services (Dlouhý, 2007). But this complexity is the reason why health care as an application area is so challenging for researchers.

I have described two major approaches to efficiency and productivity analysis in health care: non-parametric methods, such as DEA, FDH and FRH, and parametric methods (econometric analysis of production function). In my view, the two main strengths of non-parametric methods are the ability to deal with multiple inputs and outputs and to identify (for technically inefficient units) the peers that are real production units. The weakness is a deterministic nature of these methods because they do not include an error term in the model, so there is a possibility of a measurement error. Hypothesis testing is much less developed in comparison to econometric analysis, although it was shown recently that the statistical properties of DEA estimators can be established (Cooper, Seiford, Zhu, 2004). The strength of FDH and FRH, in contrast to DEA, is that no assumptions about the nature of returns to scale are necessary. On the other hand, FDH and FRH are less powerful in discriminating between the production units, especially for smaller samples.

The main advantage of parametric methods over non-parametric ones is a highly developed methodology of hypothesis testing. The selection of explanatory variables and of functional form can be tested and serve as a feedback to a researcher. On the other hand, it is more complicated to deal with multiple inputs and outputs in econometric analysis than in non-parametric methods. It is also necessary to cope with many problems related to the estimation method and the nature of data such as the multicollinearity, autocorrelation, heteroscedasticity, etc. The applications of DEA easily (unfortunately) avoid to deal with such issues, so the choice of model and variables is not based on sound arguments. SFA is able to incorporate random shocks in efficiency evaluation, but its weakness is that assumptions about the distribution of technical efficiency have to be made. This weakness may be solved if panel data are available because no assumptions about the distribution of technical efficiency have to be made in the fixed-effects model.

Chilingerian and Sherman (2004) suggest that frontier analysis may be helpful to understand the behaviour of the entire population of health care providers, and DEA might be used when research centres more on individual health care providers. However, the current key challenge is that researches must shift from illustrations of methods to advancing the field of real-world performance improvement in health care. An important issue is a development of conceptual model of how to deal with outputs and health outcomes. Although the application of quantitative economic analysis in health care is not without its problems, we believe that methods we discussed are able to help us with identifying best and worst practices within the more and more expensive health systems.
References


DOI: 10.18267/j.pep.348