EFFICIENCY MEASUREMENT

A METHODOLOGICAL REVIEW AND SYNTHESIS

PIERRE OUELLETTE
PATRICK PETIT

July 2010

Centre for Productivity
and Prosperity

HEC MONTRÉAL

The HEC Montréal Centre for Productivity and Prosperity, created in 2009, as a twofold mission.

First of all, it is devoted to research on productivity and prosperity, mainly in Quebec and in Canada as a whole.

The Centre also intends to transfer knowledge, make it widely accessible and, in the end, educate people about productivity and prosperity.

For more information on the Centre or for additional copies of this study, visit **www.hec.ca/cpp** or write us at **info.cpp@hec.ca**.

Address:
Centre for Productivity and Prosperity
HEC Montréal
3000 chemin de la Côte-Sainte-Catherine
Montreal, Quebec H3T 2A7 Canada

Telephone: 514-340-6449
Fax: 514-340-6469

This publication was produced with financial support from the ministère des Finances du Québec.

Acknowledgments

N.B.: Members of the *Productivity Analysis Research Network* (PARN) provided highly valuable input at the start of this project, following a call to all. In particular, V. Valdmanis, M.D. Rosko and B. Hollingsworth forwarded many of their own contributions and we would like to thank them for their help. We would also like to thank the Centre for Productivity and Prosperity - *HEC Montréal* for its financial support.

* The views expressed in this paper are those of the author and do not necessarily represent those of the IMF or IMF policy.

Efficiency Measurement
A Methodological Review and Synthesis

PIERRE OUELLETTE
Université du Québec à Montréal

PATRICK PETIT
International Monetary Fund (IMF)

Abstract

Measuring efficiency has been a major item on the health economics agenda over the past quarter century. A thorough review of the literature shows that almost all studies met the basic requirements proposed by Cowing and Stevenson in 1983, as they relied on the solid theoretical foundations of production economics. Many methods were nevertheless developed and used, with some grounded in statistics, others in operations research, or accounting. The objective of this paper is to show how these methods often fail to include *all* relevant theoretical considerations. For example, authors relying on economic theory have applied empirical methods with stochastic error terms that are sometimes at odds with certain properties of their models. In fact, almost all models can be approached as specific cases of a general model. We will show that each model implies specific assumptions on the nature of the data, and that in some cases, the models are incoherent.

Résumé

La mesure de l'efficience a constitué un objectif de recherche majeur en économie de la santé depuis les 25 dernières années. Presque toutes les études ont souscrit à l'exigence formulée par Cowing et Stevenson en 1983 de baser leur approche sur des fondements théoriques solides tirés de l'économie de la production. À partir de ces fondements, plusieurs méthodes ont été développées et utilisées, certaines puisant dans les méthodes statistiques, d'autres en recherche opérationnelle et certaines dans des méthodes comptables. Notre objectif est de montrer que ces méthodes ont souvent le défaut de ne pas avoir pris en compte *l'ensemble* des prescriptions de la théorie. Par exemple, les auteurs ayant basé leur approche sur la théorie économique ont superposé une structure stochastique de termes d'erreur qui est parfois incompatible avec certaines propriétés de la théorie. En fait, presque tous les modèles peuvent être vus comme des cas particuliers d'un modèle général. Nous montrerons qu'à chacun de ces modèles correspond un ensemble d'hypothèses sur la nature des données et que dans certains cas, les modèles sont incohérents.

Table of Contents

Introduction

Introduction

Research on the efficiency of health care systems has been motivated in large part by the increasing weight health budgets have exerted on public finances. Contributions on the efficiency of health institutions have by and large attempted to answer one or both of two main questions:

1. How to determine a hospital budget (or other health institutions)?
2. How to determine the price of a medical procedure?

In fact, these questions are two sides of the same coin, as they are both linked to the optimal output cost of a given hospital in a given environment. They are different, however, in the level of aggregation: in one case, it is the institution's cost that matters, and in the other, the cost of the specific procedures.

The answer to the first question is provided by the cost function $C(w, y)$,[1] whereas the answer to the second one by the properties of this cost function. In a first rank optimum, the price of output y_i, noted p_i, is the marginal cost of this output $p_i = \frac{\partial C}{\partial y_i}$, whereas in a second rank optimum (if for example firms are subject to a balanced budget constraint – also called the *Boîteux-Ramsay* optimum), it is the marginal cost corrected by the demand elasticities.[2] Regardless, knowledge of the cost function is essential and as it is unobserved, it must be inferred from available data. The goal of what follows is therefore precisely to show how one can recover the cost function from available price and quantity data.

[1] The following notation is used throughout the paper: C is the total cost, w is the input price vector, (with input quantity noted x), and y is the vector of outputs. For the sake of simplicity and to better integrate the various framework covered in this methodological synthesis, we will use a simple form of the cost function. i.e., the non-regulated cost function. It is possible to considerably generalize the firm's environment by introducing quasi-fixed or fixed inputs (also called non-discretionary inputs), regulation, technological parameters, etc. The issue of input and output quality will be discussed at a later point.

[2] Although not derived from optimal pricing theory, average cost pricing can also be used by governmental organizations for simplicity.

Theoretical Cost Function and Observed Costs

I. Theoretical Cost Function and Observed Costs

This section describes theoretical links between observed costs and minimal costs. Empirical methods to estimate minimal costs will be discussed later.

The cost minimization problem is described by:

$$C(w,y) \triangleq \min_{x} \{w'x : f(y,x) \leq 0\}.$$

The solution to this problem, if it exists (we will assume it does for what follows), is given by the vector of conditional factor demands:

$$x = x(w,y).$$

In fact, the value given by the cost function is the minimum budget required by an efficient hospital to provide the health service vector y, given input prices w and input quantity x, with a technology described by production function f.

However, the observed cost is:[3]

$$C^{obs} \triangleq \sum_{i=1}^{n} w_i^{obs} x_i^{obs} = w^{obs}{}' x^{obs}.$$

Comparing observed and theoretical costs amounts to checking whether the observed cost is the same as the minimum cost required for a given institution to provide a certain level of services. In other words, one should measure the gap between the minimum theoretical cost as determined by the cost function, and the observed cost.

Such gaps between cost variables stem from different types of errors:

1. Measurement errors;
2. Optimization errors.

[3] In general, the superscript *obs* refers to observed variables.

Much can be understood by the term "Optimization errors"; what exactly does it mean? Are these one-off random managerial errors? Or systematic sub-optimal decisions whose source lies in a deficient incentive system? In the first case, one can hardly think of ways to improve the situation and can probably only hope that such decisions do not occur too frequently. In the second case, we must also consider that managers behave optimally in the face of deficient incentives, hence a difference between observed and theoretical costs. Ma (1994) showed that it can be optimal for hospitals (but not for society at large) not to manage resources efficiently within the budget allocation mechanism they are subject to, in order to extract an economic rent.[4] The relationship between inefficiency on one hand and the economic and regulatory (e.g., budget allocation process) environment on the other hand is so close that one can hardly consider them separately. The economic literature relies on two main approaches to understand such errors: the additive and multiplicative models.

I.1. The Additive Model

The additive model is based on three types of variables: input quantity x, their price w, and output y. Observational errors on prices ε_w entirely explain the difference between the observed and effective prices faced by organizations:

$$w^{obs} = w + \varepsilon_w .$$

However, the difference between input quantities can stem from observational errors ε_x and from optimization errors ν_x:

$$x^{obs} = x + \varepsilon_x + \nu_x .$$

While the statistical distribution of error terms will be discussed later, at this stage, it is already obvious that observational and optimization errors entirely explain the gap between the observed costs C^{obs} and the theoretical (minimal) cost C. The relationship is as follows:

[4] In Ma's terminology, inefficiency is caused by a suboptimal cost-reducing effort that depends on the hardship of the managerial team's work.

$$
\begin{aligned}
C &= C(w,y) \\
&= w'x(w,y) \\
&= \left(w+\left(w^{obs}-w^{obs}\right)\right)'\left(x(w,y)+\left(x^{obs}-x^{obs}\right)\right) \\
&= \left(w^{obs}-\left(w^{obs}-w\right)\right)'\left(x^{obs}-\left(x^{obs}-x(w,y)\right)\right) \\
&= w^{obs}{}'x^{obs}-w^{obs}{}'\left(x^{obs}-x(w,y)\right)-x^{obs}{}'\left(w^{obs}-w\right)+\left(w^{obs}-w\right)'\left(x^{obs}-x(w,y)\right) \\
&= C^{obs}-w^{obs}{}'\left(\varepsilon_x+\nu_x\right)-x^{obs}{}'\varepsilon_w+\varepsilon_w{}'\left(\varepsilon_x+\nu_x\right).
\end{aligned}
$$

Hence:

$$
\begin{aligned}
C^{obs} &= C(w,y)+w^{obs}{}'\left(\varepsilon_x+\nu_x\right)+x^{obs}{}'\varepsilon_w-\varepsilon_w{}'\left(\varepsilon_x+\nu_x\right) \\
&= C(w,y)+\underbrace{\left[w^{obs}{}'\varepsilon_x+x^{obs}{}'\varepsilon_w-\varepsilon_w{}'\varepsilon_x\right]}_{\varepsilon_C}+\underbrace{\left(w^{obs}-\varepsilon_w\right)'\nu_x}_{\nu_C}
\end{aligned}
$$

This implies that the difference between observed and minimum cost (noted μ) is:

$$
\begin{aligned}
\mu &= \underbrace{\left[w^{obs}{}'\varepsilon_x+x^{obs}{}'\varepsilon_w-\varepsilon_w{}'\varepsilon_x\right]}_{\varepsilon_C}+\underbrace{\left(w^{obs}-\varepsilon_w\right)'\nu_x}_{\nu_C} \\
&= \varepsilon_C+\nu_C.
\end{aligned}
$$

Output-related errors are for now restricted to measurement issues:

$$
y^{obs} = y+\varepsilon_y \leftrightarrow y^{obs}-\varepsilon_y = y
$$

And thus:

$$
C^{obs} = C\left(w^{obs}-\varepsilon_x, y^{obs}-\varepsilon_y\right)+\underbrace{\left[w^{obs}{}'\varepsilon_x+x^{obs}{}'\varepsilon_w-\varepsilon_w{}'\varepsilon_x\right]}_{\varepsilon_C}+\underbrace{\left(w^{obs}-\varepsilon_w\right)'\nu_x}_{\nu_C}
$$

and

$$
x_i^{obs} = x_i\left(w^{obs}-\varepsilon_x, y^{obs}-\varepsilon_y\right)+\left(\varepsilon_{x_i}+\nu_{x_i}\right), \forall i = 1,\ldots,n.
$$

It can be noted that Shephard's relation holds both for the minimum cost (as an application of the envelope theorem) and for the observed cost:

$$\frac{\partial C(w,y)}{\partial w} = x(w,y) \text{ and } \frac{\partial C^{obs}}{\partial w^{obs}} = x^{obs}.$$

More important, the inefficiency term necessarily depends on input prices. This characteristic has been almost systematically ignored in econometric work.

I.2. The Multiplicative Model

The multiplicative model is similar to the additive model, but errors multiply the main variables instead of being added to them:

$$\begin{aligned} w^{obs} &= we^{\varepsilon_w} \leftrightarrow w^{obs}e^{-\varepsilon_w} = w \\ x^{obs} &= xe^{\varepsilon_x+\nu_x} \leftrightarrow x^{obs}e^{-(\varepsilon_x+\nu_x)} = x \\ y^{obs} &= ye^{\varepsilon_y} \leftrightarrow y^{obs}e^{-\varepsilon_y} = y. \end{aligned}$$

We proceed as before to depict the relationship between minimum and observed costs:

$$\begin{aligned} C &= C(w,y) \\ &= w^{obs}{}'x^{obs} - w^{obs}{}'\left(x^{obs} - x(w,y)\right) - x^{obs}{}'\left(w^{obs} - w\right) + \left(w^{obs} - w\right)'\left(x^{obs} - x(w,y)\right) \\ &= w^{obs}{}'x^{obs} - w^{obs}{}'\left(x^{obs} - x^{obs}e^{-(\varepsilon_x+\nu_x)}\right) - x^{obs}{}'\left(w^{obs} - w^{obs}e^{-\varepsilon_w}\right) + \left(w^{obs} - w^{obs}e^{-\varepsilon_w}\right)'\left(x^{obs} - x^{obs}e^{-(\varepsilon_x+\nu_x)}\right) \\ &= w^{obs}{}'x^{obs} - w^{obs}{}'x^{obs}\left(1 - e^{-(\varepsilon_x+\nu_x)}\right) - x^{obs}{}'w^{obs}\left(1 - e^{-\varepsilon_w}\right) + \left(w^{obs} - w^{obs}e^{-\varepsilon_w}\right)'\left(x^{obs} - x^{obs}e^{-(\varepsilon_x+\nu_x)}\right) \\ &= w^{obs}e^{-\varepsilon_w}{}'x^{obs}e^{-(\varepsilon_x+\nu_x)} = \sum_i^n w_i^{obs}x_i^{obs}e^{-\left(\varepsilon_{w_i}+\varepsilon_{x_i}+\nu_{x_i}\right)}. \end{aligned}$$

This last result can be formulated as:

$$C^{obs} = \frac{C\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right)}{\sum_i^n \frac{w_i^{obs}x_i^{obs}}{C^{obs}} e^{-\left(\varepsilon_{w_i}+\varepsilon_{x_i}+v_{x_i}\right)}},$$

Or, following logarithmic transformation:

$$\ln C^{obs} = \ln C\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right) - \ln \sum_i \frac{w_i^{obs}x_i^{obs}}{C^{obs}} e^{-\left(\varepsilon_{w_i}+\varepsilon_{x_i}+v_{x_i}\right)}.$$

This entails that in a single input case (i = 1), the error term would be additive after logarithmic transformation:

$$\ln C^{obs} = \ln C\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right) + \underbrace{\left(\varepsilon_w + \varepsilon_x\right)}_{\varepsilon_C^m} + \underbrace{v_x}_{v_C^m}.$$

However, with many inputs, the double-log model with additive error terms would not be valid anymore. This is particularly damaging for stochastic frontier cost models where error terms are additive regardless of the composition discussed earlier, as the model consistency that ensures a relationship between cost and factor demand would be lost.[5]

The factor demand share system, noted S_i, is as follows:

[5] This lack of consistency was mentioned by McElroy (1987), but without a refering to inneficiency terms and measurement errors on w and y.

$$S_i = \frac{w_i x_i}{C} = \frac{w_i^{obs} x_i^{obs} e^{-\left(\varepsilon_{w_i}+\varepsilon_{x_i}+\nu_{x_i}\right)}}{C} = \frac{w_i^{obs} x_i^{obs}}{C^{obs}} \times e^{-\left(\varepsilon_{w_i}+\varepsilon_{x_i}+\nu_{x_i}\right)} \times \frac{C^{obs}}{C} = S_i^{obs} \times e^{-\left(\varepsilon_{w_i}+\varepsilon_{x_i}+\nu_{x_i}\right)} \times \frac{C^{obs}}{C}$$

$$\leftrightarrow$$

$$\begin{aligned} S_i^{obs} &= S_i\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right) \times e^{\left(\varepsilon_{w_i}+\varepsilon_{x_i}+\nu_{x_i}\right)} \times \frac{C}{C^{obs}} \\ &= S_i\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right) \times e^{\left(\varepsilon_{w_i}+\varepsilon_{x_i}+\nu_{x_i}\right)} \times \sum_j \frac{w_j^{obs} x_j^{obs}}{C^{obs}} e^{-\left(\varepsilon_{w_j}+\varepsilon_{x_j}+\nu_{x_j}\right)} \\ &= S_i\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right) \times e^{\left(\varepsilon_{w_i}+\varepsilon_{x_i}+\nu_{x_i}\right)} \times \sum_j S_j^{obs} e^{-\left(\varepsilon_{w_j}+\varepsilon_{x_j}+\nu_{x_j}\right)} \\ &= S_i\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right) \times \sum_j \left(S_j^{obs} \times e^{\left(\varepsilon_{w_i}+\varepsilon_{x_i}+\nu_{x_i}\right)-\left(\varepsilon_{w_j}+\varepsilon_{x_j}+\nu_{x_j}\right)} \right). \end{aligned}$$

But as $S_i = \frac{\partial \ln C\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right)}{\partial w_i} w_i = \frac{\partial \ln C\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right)}{\partial w_i^{obs}} w_i^{obs}$, we have:

$$S_i^{obs} = \frac{\partial \ln C\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right)}{\partial \ln w_i^{obs}} \times \sum_j \left(S_j^{obs} \times e^{\left(\varepsilon_{w_i}+\varepsilon_{x_i}+\nu_{x_i}\right)-\left(\varepsilon_{w_j}+\varepsilon_{x_j}+\nu_{x_j}\right)} \right).$$

In fact, the factor demand share system is a system of linear equations in S_i^{obs} that can be solved for observed shares. This system is of rank $n-1$, as the sum of shares must add to 1, by definition. In the simplest case where n = 2, we get:

$$S_1^{obs} = \frac{S_1 e^{\left(\varepsilon_{w_1}+\varepsilon_{x_1}+\nu_{x_1}\right)-\left(\varepsilon_{w_2}+\varepsilon_{x_2}+\nu_{x_2}\right)}}{1-S_1+S_1 e^{\left(\varepsilon_{w_1}+\varepsilon_{x_1}+\nu_{x_1}\right)-\left(\varepsilon_{w_2}+\varepsilon_{x_2}+\nu_{x_2}\right)}}.$$

One notes the complexity of the equation even in this basic case. Finally, the overall cost/ share system is as follows:

$$\ln C^{obs} = \ln C\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right) - \ln \sum_j S_j^{obs} e^{-\left(\varepsilon_{w_j}+\varepsilon_{x_j}+\nu_{x_j}\right)}$$

and

$$S_i^{obs} = \frac{\partial \ln C\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right)}{\partial \ln w_i^{obs}} \times \sum_j \left(S_j^{obs} \times e^{\left(\varepsilon_{w_i}+\varepsilon_{x_i}+\nu_{x_i}\right)-\left(\varepsilon_{w_j}+\varepsilon_{x_j}+\nu_{x_j}\right)} \right)$$

$$\leftrightarrow \ln S_i^{obs} = \ln S_i\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right) + \left(\varepsilon_{w_i}+\varepsilon_{x_i}+\nu_{x_i}\right) + \ln \sum_j S_j^{obs} e^{-\left(\varepsilon_{w_j}+\varepsilon_{x_j}+\nu_{x_j}\right)}.$$

This formulation is extremely complex as the observed cost and the shares (dependent variables) are on both sides of the equation. However, it is possible to simplify the share system by taking the ratio of shares:

$$\frac{S_i^{obs}}{S_n^{obs}} = \frac{S_i\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right)}{S_n\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right)} e^{\left(\varepsilon_{w_i}+\varepsilon_{x_i}+\nu_{x_i}\right)-\left(\varepsilon_{w_n}+\varepsilon_{x_n}+\nu_{x_n}\right)}, \forall i = 1,\ldots,n$$

$$\leftrightarrow \ln \frac{S_i^{obs}}{S_n^{obs}} = \ln \frac{S_i\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right)}{S_n\left(w^{obs}e^{-\varepsilon_w}, y^{obs}e^{-\varepsilon_y}\right)} + \left(\varepsilon_{w_i}+\varepsilon_{x_i}+\nu_{x_i}\right) - \left(\varepsilon_{w_n}+\varepsilon_{x_n}+\nu_{x_n}\right), \forall i = 1,\ldots,n-1.$$

Nevertheless, the cost equation remains complex and estimating the system using shares only would not allow recovering the information required to calculate absolute inefficiency. Indeed, unless we assume that one of the inefficiency terms is nil, we can only recover (n-1) inefficiency terms $\left(\nu_{x_i} - \nu_{x_n}\right), \forall i = 1,\ldots,n-1.$ To get around this problem, we can make a few assumptions. For example, by eliminating errors on input prices and quantity ($\varepsilon_{w_i} = \varepsilon_{x_i} = 0, \forall i,$) and if we assume that all inputs are equally inefficient ($\nu_{x_i} = \nu_C, \forall i,$), then:

$$\ln C^{obs} = \ln C\left(w^{obs}, y^{obs}e^{-\varepsilon_y}\right) + \nu_C$$

$$S_i^{obs} = \frac{\partial \ln C\left(w^{obs}, y^{obs}e^{-\varepsilon_y}\right)}{\partial \ln w_i}.$$

Although the absence of error terms on shares (except for the output) is rather restrictive and appears to completely rule out this approach, it is possible to avoid this shortcoming by assuming that $\varepsilon_{w_i} + \varepsilon_{x_i} = \varepsilon_C, \forall i,$ while keeping $\nu_{x_i} = \nu_C, \forall i,$ and thus:

$$\ln C^{obs} = \ln C\left(w^{obs} e^{-\varepsilon_w}, y^{obs} e^{-\varepsilon_y}\right) + \varepsilon_C + \nu_C$$
$$S_i^{obs} = \frac{\partial \ln C\left(w^{obs} e^{-\varepsilon_w}, y^{obs} e^{-\varepsilon_y}\right)}{\partial \ln w_i^{obs}}.$$

As a specific example, the Cobb-Douglas case would yield the following:

$$\begin{aligned}\ln C^{obs} &= a + b_1 \ln\left(w_1^{obs} e^{-\varepsilon_{w_1}}\right) + b_2 \ln\left(w_2^{obs} e^{-\varepsilon_{w_2}}\right) + b_y \ln\left(y^{obs} e^{-\varepsilon_y}\right) + \varepsilon_C + \nu_C \\ &= a + b_1 \ln w_1^{obs} + b_2 \ln w_2^{obs} + b_y \ln y^{obs} - \left(b_1 \varepsilon_{w_1} + b_2 \varepsilon_{w_2} + b_y \varepsilon_y\right) + \varepsilon_C + \nu_C \\ S_i^{obs} &= b_i.\end{aligned}$$

We note that the share system (which for this functional form does not include error terms and yields constant shares by definition) immediately provides coefficients $\left(b_1, b_2\right)$, and the cost equation becomes:

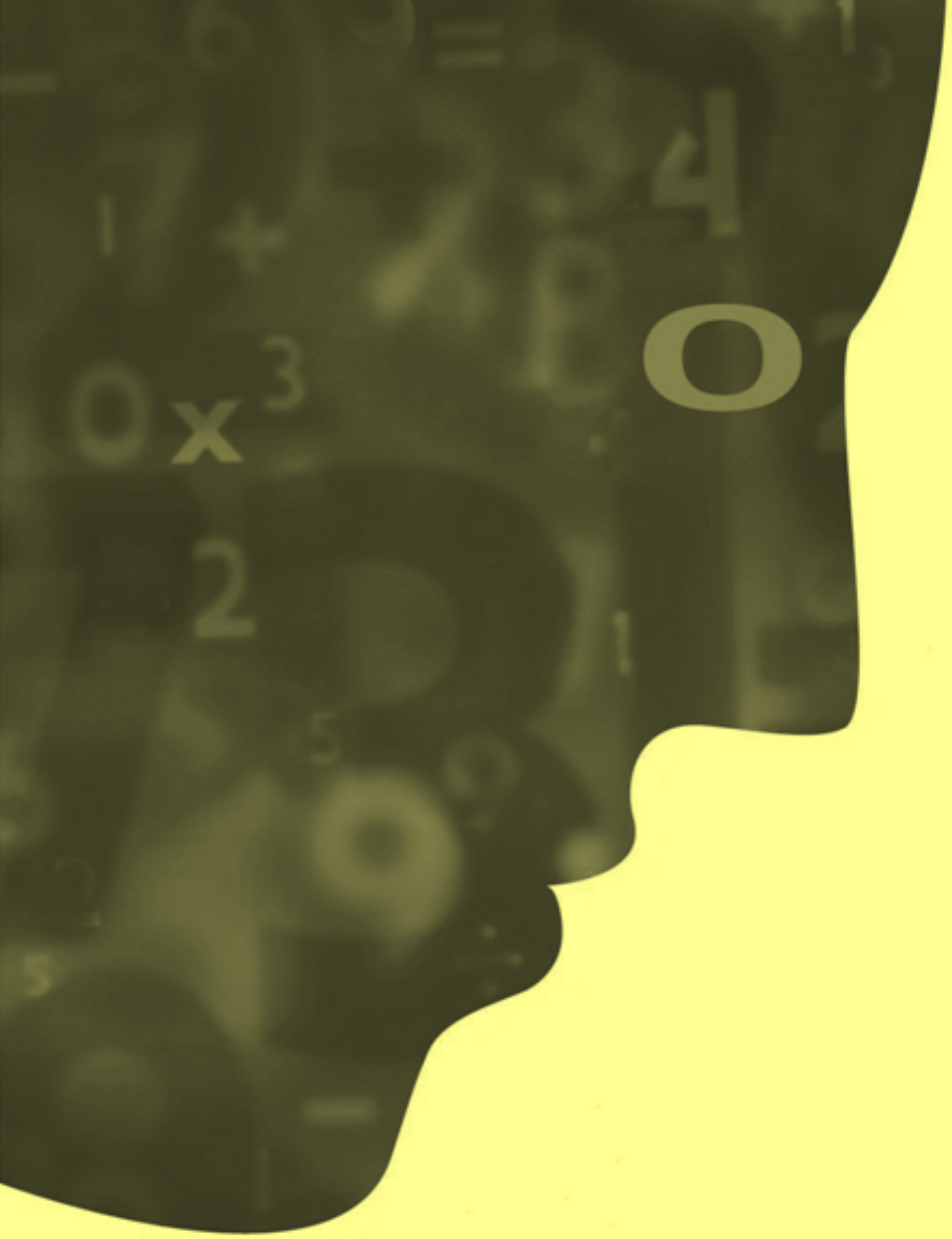
$$\ln C^{obs} - S_1^{obs} \ln w_1^{obs} - S_2^{obs} \ln w_2^{obs} = a + b_y \ln y^{obs} - \left(S_1^{obs} \varepsilon_{w_1} + S_2^{obs} \varepsilon_{w_2} + b_y \varepsilon_y\right) + \varepsilon_C + \nu_C.$$

We can then estimates $\left(a, b_y\right)$ taking into account the dependence of the error term to shares and to the output coefficient.

By replacing the assumption on input price and quantity measurement errors $\varepsilon_{w_i} + \varepsilon_{x_i} = \varepsilon_C, \forall i,$ by $\varepsilon_{w_i} = \varepsilon_w, \varepsilon_{x_i} = \varepsilon_x, \forall i,$ we get:

$$\begin{aligned}\ln C^{obs} - S_1^{obs} \ln w_1^{obs} - S_2^{obs} \ln w_2^{obs} &= a + b_y \ln y^{obs} - \left(S_1^{obs} \varepsilon_w + S_2^{obs} \varepsilon_w + b_y \varepsilon_y\right) + \varepsilon_w + \varepsilon_x + \nu_C \\ &= a + b_y \ln y^{obs} + \left(\varepsilon_x - b_y \varepsilon_y\right) + \nu_C.\end{aligned}$$

An important conclusion of the additive and multiplicative models is that assumptions on error terms of the model variables have an immediate impact on the way error terms enter estimated equations. This, in turn, has important implications for the overall consistency (including Shephard's lemma) of the model, and therefore for the empirical methods used to measure the cost function. These methods are described in the next three sections.



Statistics Models

II. Statistics Models

II.1. Stochastics Frontiers

The stochastic frontier approach has traditionally been used only for cost functions. The additive model is:

$$C^{obs} = C\left(w^{obs} - \varepsilon_w, y^{obs} - \varepsilon_y\right) + \underbrace{\left[w^{obs}{}'\varepsilon_x + x^{obs}{}'\varepsilon_w - \varepsilon_w{}'\varepsilon_x\right]}_{\varepsilon_C} + \underbrace{\left(w^{obs} - \varepsilon_w\right)'\nu_x}_{\nu_C}.$$

And the usual assumptions are:

$$\begin{aligned}
&\varepsilon_y = 0,\\
&\left[w^{obs}{}'\varepsilon_x + x^{obs}{}'\varepsilon_w - \varepsilon_w{}'\varepsilon_x\right] = \mu_C \sim \aleph\left(0, \sigma_\mu^2\right),\\
&0 \le \left(w^{obs} - \varepsilon_w\right)'\nu_x = \nu_C \sim ?\left(\overline{\nu}_C, \sigma_\nu^2\right).
\end{aligned}$$

Leaving aside for now the distribution of the inefficiency term ("?" in the equation above), we note that the treatment of measurement and optimization errors is a delicate issue. Indeed, the assumption that error terms are such that the last assumption above would be independent of observed prices and quantities is rather far-fetched. Yet, this issue has been ignored by researchers.

Further assuming that there is no measurement error on prices ($w = w^{obs} \leftrightarrow \varepsilon_w = 0$), the measurement error on cost becomes $w^{obs}{}'\varepsilon_x = \mu_C \sim \aleph\left(0, \sigma_\mu^2\right)$ and the inefficiency term is given by $0 \le w^{obs}{}'\nu_x = \nu_C \sim ?\left(\overline{\nu}_C, \sigma_\nu^2\right)$. The cost function simplifies to:

$$C^{obs} = C\left(w^{obs}, y^{obs}\right) + \underbrace{w^{obs}{}'\varepsilon_x}_{\varepsilon_C} + \underbrace{w^{obs}{}'\nu_x}_{\nu_C}.$$

This assumption however, does not eliminate the dependence of error and inefficiency terms to input prices. Consequently, the choice of a functional form for the cost function

$C\left(w^{obs}, y^{obs}; \beta\right)$, (where β is the vector of estimated parameters) will have a direct impact on results. Indeed, the choice of a functional form is far from trivial, as shown by Gagné and Ouellette (1998 and 2002). Similarly, the choice of a distribution for the inefficiency term "?" is also problematic and might impact estimated results.

This method requires that the estimated function meets the characteristics of a cost function and therefore be monotonous in (w, y) (and satisfies Shephard's lemma), concave in w, and homogeneous of degree one in w. While some of these properties can be imposed (equality-type constraints such as homogeneity, for example), others can only be tested (inequality-type constraints such as concavity in input prices).

A widespread practice in the field of efficiency measurement for health systems institutions is to make the inefficiency terms endogenous. Inevitably, as this entails prior identification of inefficiency variables, this approach will have to rely on an explicit vector of such variables:

$$\nu_C = \nu_C(Z),$$

where Z is the vector of inefficiency-related variables. It can include, for example the level of competition (Gini coefficients, Herfindhal index, geographic proximity of competitors), the regulation, the type of budget allocation mechanisms, institutional idiosyncrasies (e.g., research hospital or not), the type of patients (share of Medicare/ Medicaid patients, free care), etc.

The measurement of inefficiency proceeds in one or two steps. The two-step approach first measures inefficiency, which is then linked to its determinants through regression analysis. In the single-step approach, the impact of inefficiency determinants is estimated simultaneously with technological coefficients. In all cases, we obtain a measure of the inefficiency term ν_C, defined as the expected distance between the observed cost and the cost frontier, conditional on the error term on measurement μ_C.

Although making the inefficiency term endogenous can be seen as a step forward, we put forth that it can, in fact, represent a specification error. For example, pointing to regulation as a source of inefficiency is in itself an acknowledgement that a model without it is incomplete and should be modified to include it: the cost function itself should take into account that regulation

limits organizational choices, hence recourse to a regulated cost function, as opposed to a non-regulated cost function that subsequently corrects the inefficiency term to endogenize it with regulation. In other words, regulation modifies the entire relationship between inputs and outputs and not only the breakdown of the inefficiency term.[6]

Obviously, this discussion on cost function estimation should not overshadow that joint estimation of cost functions and factor demand (or corresponding shares) is preferable and econometrically more efficient.

II.2. Corrected Least Square Method

A rather simple method to measure and compare efficiency levels is to assume that compared units have the same technology, except for an additive term that represents efficiency difference. Firm i's technology is:

$$C_i^{obs} = \eta_i + C\left(w_i^{obs}, y_i^{obs}\right) + \varepsilon_{C_i}.$$

This implies the following assumptions:

$$\varepsilon_{C_i} \equiv \left[w_i^{obs} \,'\varepsilon_{x_i} + x_i^{obs} \,'\varepsilon_{w_i} - \varepsilon_{w_i} \,'\varepsilon_{x_i} \right]$$
$$\nu_{C_i} \equiv \left(w_i^{obs} - \varepsilon_{w_i} \right)' \nu_{x_i} \triangleq \eta_i.$$

The dependence of ε_{C_i} and η_i to input price and quantity is not taken into account. If the functional form includes a constant, say η_0, the efficiency parameter should be normalized by setting one of them to zero:[7]

$$C_i^{obs} = \eta_i + \left(\eta_0 + C^*\left(w_i^{obs}, y_i^{obs}\right) \right) + \varepsilon_{C_i}.$$

[6] On regulated cost functions, one can read Färe and Logan (1983), Ouellette and Vigeant (2001a and b and 2010).

[7] As for stochastic frontier models, panel data methods can obviously be used in this case too.

The estimation is performed by including a binary variable for each firm (except one) to recover efficiency terms. If the smallest one is positive, the firm without the binary variable is the most efficient (others have higher cost structures). Otherwise, we proceed as follows:

$$\eta_i^{corr} = \eta_i - \min\{\eta_1, \ldots, \eta_n\}.$$

The firm with the smallest η_i will be given a value of zero and all others will have a positive value, hence η_i representing the cross-firm efficiency gap (in cost terms).

We can also use a multiplicative model (or log-additive):

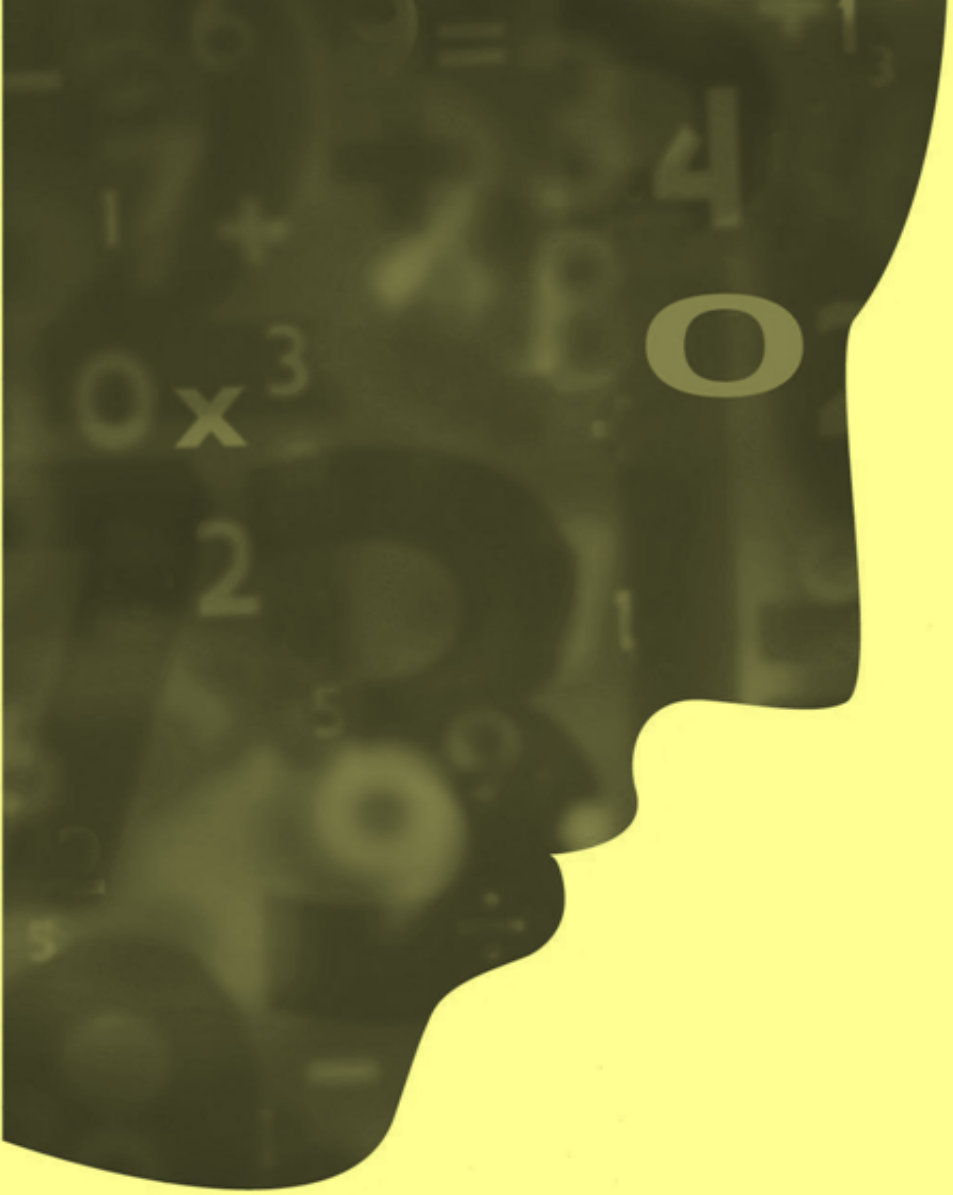
$$C_i^{obs} = e^{\eta_i} \times C\left(w_i^{obs}, y_i^{obs}\right) \times e^{\varepsilon_{C_i}} \leftrightarrow \ln C_i^{obs} = \eta_i + \ln C\left(w_i^{obs}, y_i^{obs}\right) + \varepsilon_{C_i},$$

And the difference in efficiency is (in cost terms):

$$C\left(w, y\right) \times \left(e^{\eta_i} - 1\right).$$

Obviously, efficiency gaps can be adjusted once more so that they are set to zero for the most efficient firm.

Nevertheless, in light of earlier comments, this model suffers from many disadvantages, notably the lack of consistency between cost and factor demand, and the imposition of a stochastic frontier that does not take into account the relationship between error terms and input price and quantity.



Operational Research Models

III. Operational Research Models

III.1. Data Envelopment Analysis (Farrell, 1957)

As any econometric method, stochastic frontiers draw a curve in a cloud of observation points, relying on statistical methods to divide the observations on each side of the curve and thus positioning it to reflect the overall pattern of observations. Farrell's method, however, consists in covering the production set[8] (observations) with many subsets whose definition is based on a few economic assumptions, of which the most common are:

- Free disposal of inputs;
- Free disposal of inputs;
- Convexity of the production set.

The FDH (*Free disposal hull*) model relies on the first two assumptions only. Adding the third assumption leads to the DEA model (*Data envelopment analysis*).

But before pursuing in this vein, let's go back to basic concepts on the relationship between observed and optimal values.

III.1.1. The additive model[9]

By definition, we have $F\left(y, x\left(w, y\right)\right) \equiv 0$. Substituting for observed variables yields: $F\left(y^{obs} - \varepsilon_y, x^{obs} - \varepsilon_x - \nu_x\right) \equiv 0$. By definition also, we have: $F\left(y^{obs} - \varepsilon_y, x^{obs} - \varepsilon_x\right) \leq 0$.

These expressions will be equal in the absence of inefficiency ($\nu_x = 0$), and otherwise unequal

[8] Or any other representation of technology, such as a cost, profit or distance function, isoquants, etc.

[9] Sections III.1.1 and III.1.2 follow from our choice regarding exogenous variables (i.e., input price and the output quantity) and endogenous variables (input quantity). We can change the nature (orientation) of the measure (which here is input-oriented) and obtain an output-oriented measure in the case of revenue maximization, or even a mixed input and output orientation for profit maximization. In the case of revenue maximization, output quantity is endogenous, and output price and input quantity are exogenous. For profit maximization, input and output quantity are endogenous and their prices exogenous.

$(\nu_x > 0)$. Assuming that the error term is the same for all inputs (in which case ν_x is a scalar), the technology can be recovered in the following fashion:

$$\max_{\nu_x} \left\{ \nu_x : F\left(y^{obs} - \varepsilon_y, x^{obs} - \varepsilon_x - \nu_x \right) \right\}.$$

Assuming otherwise however, we must define them with an aggregate variable. For example, we can choose to minimize the real financial loss related to the inefficiency:

$$\max_{\nu_x} \left\{ \left(w^{obs} - \varepsilon_w \right)' \nu_x : F\left(y^{obs} - \varepsilon_y, x^{obs} - \varepsilon_x - \nu_x \right) \right\}.$$

We can also select other criteria, such as minimizing the observed loss (which converges to the previous criteria if there is no observation error on prices):

$$\max_{\nu_x} \left\{ w^{obs} \,' \nu_x : F\left(y^{obs} - \varepsilon_y, x^{obs} - \varepsilon_x - \nu_x \right) \right\}.$$

III.1.2. The Multiplicative Model

Once again, by definition, we have $F\left(y, x\left(w, y\right)\right) \equiv 0$. Substituting for observed values yields $F\left(y^{obs} e^{-\varepsilon_y}, x^{obs} e^{-\varepsilon_x - \nu_x} \right) \equiv 0$. It follows by definition again that $F\left(y^{obs} e^{-\varepsilon_y}, x^{obs} e^{-\varepsilon_x} \right) \leq 0.$ These expressions will be equal in the absence of inefficiency ($\nu_x = 0$), and strictly unequal with inefficiencies $(\nu_x > 0)$. Assuming equal inefficiency terms, one way to recover technology will be:

$$\max_{\nu_x} \left\{ \nu_x : F\left(y^{obs} e^{-\varepsilon_y}, x^{obs} e^{-\varepsilon_x} e^{-\nu_x} \right) \right\} \leftrightarrow \min_{\nu_x} \left\{ e^{-\nu_x} : F\left(y^{obs} e^{-\varepsilon_y}, x^{obs} e^{-\varepsilon_x} e^{-\nu_x} \right) \right\}.$$

After defining $\theta_x = e^{-\nu_x}$, we can rewrite the problem as:

$$\min_{\theta_x} \left\{ \theta_x : F\left(y^{obs} e^{-\varepsilon_y}, \theta_x x^{obs} e^{-\varepsilon_x} \right) \leq 0 \right\}.$$

Not assuming that inefficiency terms are equal, we must here too define them with an aggregate variable. We can again choose to minimize the real financial loss related to the inefficiency:

$$\min_{\theta_x}\left\{\left(w^{obs}e^{-\varepsilon_w}\right)'\theta_x : F\left(y^{obs}e^{-\varepsilon_y}, \theta_x x^{obs}e^{-\varepsilon_x}\right)\leq 0\right\}$$

or the observed financial loss (again, these two measures converge if there are no errors on observed prices):

$$\min_{\theta_x}\left\{w^{obs}{}'\theta_x : F\left(y^{obs}e^{-\varepsilon_y}, \theta_x x^{obs}e^{-\varepsilon_x}\right)\leq 0\right\}.$$

Given this formulation, we are very close to Farrell's model and of Shephard's distance function (noted D and defined hereafter). In fact, if we assume that ν_{x_i} are equal ($\nu_{x_i}=\nu \leftrightarrow \theta_{x_i}=\theta$) and that there is no measurement error on variables, we obtain Shephard's distance function:

$$D\left(y^{obs},x^{obs}\right)^{-1} \triangleq \min_{\theta}\left\{\theta : F\left(y^{obs},\theta x^{obs}\right)\leq 0\right\} \leftrightarrow D\left(y^{obs},x^{obs}\right) \triangleq \max_{\phi}\left\{\phi : F\left(y^{obs},\frac{x^{obs}}{\phi}\right)\leq 0\right\}.$$

III.1.3. Data Envelopment

Free disposal assumptions create a set of feasible possibilities for each observation and the union of these sets constitutes an interior approximation of the overall set of possible outcomes with a given technology, whether the outcome is defined in terms of production or cost. This method typically provides a stair-like technology, assuming a convex possibility set results in a convex polyhedron, closer to standard representations of technology found in microeconomic textbooks.

The assumption of convex possibility set is first and foremost a matter of personal choice and is often the result of other considerations regarding input substitution. In general, higher levels of aggregation of the studied institutions or health systems make an easier case for convexity. At low levels of aggregation however, the possibility to substitute might be absent due to *putty clay* effects. For example, fixed coefficient technology at low aggregation levels might make it

impossible to replace a surgeon working in an operating room by more scalpels or suture thread. At a more aggregate level (the hospital, for example), substitution possibilities appear: we can replace some surgical operations (and thus the surgeon) by drugs and medical follow up. These substitutions can occur within a short time frame, or take much longer, depending on technical and organizational constraints.

The DEA method has been very popular in the past 20 years, in large parts because of the few assumptions it requires. Contrary to econometric methods, it is not necessary to impose a functional form or a particular distribution for the error terms. Furthermore, testing the theory ceases to be a concern. Nevertheless, DEA is very sensitive to outliers, but more important is the lack of confidence intervals, which has been its greatest drawback for long. This in fact boils down to assuming that the terms $\left(\varepsilon_x, \varepsilon_w, \varepsilon_y\right)$ are nil (or not significant).[10] In this case, the cost function becomes:

Additive model: $\max_{\nu_x} \left\{ w^{obs}{}'\nu_x : F\left(y^{obs}, x^{obs} - \nu_x\right)\right\}$;[11]

Multiplicative model: $\min_{\theta_x} \left\{ w^{obs}{}'\theta_x : F\left(y^{obs}, \theta_x x^{obs}\right) \le 0\right\}$.

And the goal of DEA is to calculate the value of $w'\nu_x$ or $w'\theta_x$ for each decisions making unit (firm, departments, divisions, entire health systems, etc.). Further assuming that inefficiency terms are equal generates:

[10] This implies $w^{obs} = w, x^{obs} = x + \nu_x, y^{obs} = y$ and $\nu_C = w'\nu_x$.

[11] Since $x^{obs} - \nu_x = x$, we can, for the additive model, rewrite this as:

$$\max_{\nu_x} \left\{ w^{obs}{}'\left(x^{obs} - x\right) : F\left(y^{obs}, x^{obs} - \nu_x\right)\right\} = C^{obs} - \min_x \left\{ w^{obs}{}'x : F\left(y^{obs}, x\right)\right\}$$

And resolve :

$$\min_x \left\{ w^{obs}{}'x : F\left(y^{obs}, x\right)\right\}.$$

This is the standard cost minimization by choice of inputs. The relationship is not as simple for the multiplicative model.

Additive model: $\max_{\nu_x} \left\{ \nu_x : F\left(y^{obs}, x^{obs} - \nu_x \right) \right\}$;

Multiplicative model: $\min_{\theta_x} \left\{ \theta_x : F\left(y^{obs}, \theta_x x^{obs} \right) \leq 0 \right\}$.

While this last model has become the standard for most applications, it is also obvious that nothing justifies equality of the error terms.

The fact the DEA *calculates* inefficiency as opposed to *estimating* it means that the dependence on price becomes irrelevant. This is a great advantage of this method.

If we do not assume that the terms $\left(\varepsilon_x, \varepsilon_w, \varepsilon_y \right)$ are nil or negligible, the inefficiency measure then depends on measurement errors and it becomes necessary to calculate confidence intervals. Recourse to bootstrap methods (Simar and Wilson, 1998) allowed estimating such intervals, but their validity is still not well established; in other words, we cannot be confident in the confidence measures themselves.

Finally, we might want to question that the implicit and necessary reliance on the theory is an advantage or not. The answer depends on whether theory is used as a work tool or as a support whose solidity should be tested. On one hand, the validity of DEA rests on that of economic theory or at least on the free disposal and convexity assumptions[12]. On the other hand, econometric methods allow testing the theory but at the price of assumptions on the functional form and on the distribution of error terms.

DEA does not allow calculating easily technological change (see Diewert and Parkan, 1983). For this reason, linking distance functions used in DEA and Malmquist indices has allowed different types of decomposition of inefficiency that have become increasingly popular since the work of Färe *et al*. (1992) and Caves *et al.* (1982).

III.2. Malmquist

Malmquist indices are ratios of distance functions. Caves *et al.* (1982) showed that it was possible to define productivity indices from Malmquist indices by assuming institutions were

[12] With added behavioral assumptions in the case of DEA with cost minimization.

fully efficient. This paved the way for Färe *et al.* (1992) to generalize their contribution without this assumption and to show how to calculate this index with DEA-related non-parametric techniques.

Färe *et al*. also showed that it was possible to decompose this index into various types of breakdowns. For example, it is now standard practice to break down productivity change into its efficiency and technological change components. Since then, many other possibilities have been implemented to take into account returns to scale, the presence of quasi-fixed (or non-discretionary) inputs, regulation, effects related to the composition of inputs and outputs, etc., in a movement that mirrors the 1970s work on the decomposition of Solow's residual.

Malmquist indices rely on the same non-parametric methods as the DEA and therefore have the same advantages (no functional form, no error term), but also the same shortcomings (absent or problematic confidence intervals, sensitivity to outliers, imposition of theoretical relationship and thus incapacity to test it).

III.3. Aigner and Chu (1968)'s Model

Aigner and Chu (1968)'s model did not generate much work. They use linear programming to calibrate technology, under constraints imposed by economic theory. Although they introduced many models, each with specific details, all are similar in nature and rely on a production function and a specific definition of efficiency, and we will thus limit this discussion to only one of their models.

In this model, the production and cost function can be introduced as follows:[13]

$$F\left(x^{obs}-\varepsilon_x, y^{obs}-\varepsilon_y\right)\geq 0$$

and

$$C^{obs}-C\left(w^{obs}-\varepsilon_w, y^{obs}-\varepsilon_y\right)-\underbrace{\left[w^{obs}{}'\varepsilon_x+x^{obs}{}'\varepsilon_w-\varepsilon_w{}'\varepsilon_x\right]}_{\varepsilon_C}-\underbrace{\left(w^{obs}-\varepsilon_w\right)'\nu_x}_{\nu_C}\geq 0.$$

[13] We will not introduce additive and multiplicative models here, for reasons that will become obvious shortly. Indeed, as technology needs to be linear (production and cost functions), variables of the multiplicative models will be in a logarithmic forms and the treatment of the additive and multiplicative models will therefore be identical, if not for the fact that variables are in level for the additive case and in logs for the multiplicative model.

We first choose a functional form for technology:

$$F\left(x^{obs}-\varepsilon_x, y^{obs}-\varepsilon_y;\alpha\right)\geq 0$$

and

$$C^{obs}-C\left(w^{obs}-\varepsilon_w, y^{obs}-\varepsilon_y;\beta\right)-\underbrace{\left[w^{obs}{}'\varepsilon_x+x^{obs}{}'\varepsilon_w-\varepsilon_w{}'\varepsilon_x\right]}_{\varepsilon_C}\geq 0,$$

where α and β are the coefficients of the functional form. Assuming that technology is linear in parameters (possibly after a log transformation, as in the Cobb-Douglas case used by Aigner and Chu) :

$$\left(x^{obs}-\varepsilon_x\right)'\alpha_x-\left(y^{obs}-\varepsilon_y\right)'\alpha_y\geq 0$$

and

$$C^{obs}-\left(w^{obs}-\varepsilon_w\right)'\beta_w-\left(y^{obs}-\varepsilon_y\right)'\beta_y-\underbrace{\left[w^{obs}{}'\varepsilon_x+x^{obs}{}'\varepsilon_w-\varepsilon_w{}'\varepsilon_x\right]}_{\varepsilon_C}\geq 0$$

.

Inefficiency ν is introduced as follows:

$$\left(x^{obs}-\varepsilon_x-\nu_x\right)'\alpha_x-\left(y^{obs}-\varepsilon_y\right)'\alpha_y=0$$

and

$$C^{obs}-\left(w^{obs}-\varepsilon_w\right)'\beta_w-\left(y^{obs}-\varepsilon_y\right)'\beta_y-\underbrace{\left[w^{obs}{}'\varepsilon_x+x^{obs}{}'\varepsilon_w-\varepsilon_w{}'\varepsilon_x\right]}_{\varepsilon_C}-\underbrace{\left(w^{obs}-\varepsilon_w\right)'\nu_x}_{\nu_C}=0.$$

We solve for error and inefficiency terms:

$$x^{obs}{}'\alpha_x-y^{obs}{}'\alpha_y=\underbrace{\left[\varepsilon_x{}'\alpha_x+\varepsilon_y{}'\alpha_y\right]}_{\varepsilon_\alpha}+\underbrace{\nu_x{}'\alpha_x}_{\nu_F}$$

and

$$C^{obs}-w^{obs}{}'\beta_w-y^{obs}{}'\beta_y=\underbrace{\left[\varepsilon_w{}'\beta_w+\varepsilon_y{}'\beta_y\right]}_{\varepsilon_\beta}+\underbrace{\left[w^{obs}{}'\varepsilon_x+x^{obs}{}'\varepsilon_w-\varepsilon_w{}'\varepsilon_x\right]}_{\varepsilon_C}+\underbrace{\left(w^{obs}-\varepsilon_w\right)'\nu_x}_{\nu_C}.$$

For firm i, we can write:

$$x_i^{obs}{}'\alpha_x - y_i^{obs}{}'\alpha_y = \underbrace{\left[\varepsilon_{x_i}{}'\alpha_x + \varepsilon_{y_i}{}'\alpha_y\right]}_{\varepsilon_{\alpha_i}} + \underbrace{\nu_{x_i}{}'\alpha_x}_{\nu_{F_i}}, \forall i = 1,\ldots,n$$

and

$$C_i^{obs} - w_i^{obs}{}'\beta_w - y_i^{obs}{}'\beta_y = \underbrace{\left[\varepsilon_{w_i}{}'\beta_w + \varepsilon_{y_i}{}'\beta_y\right]}_{\varepsilon_{\beta_i}} + \underbrace{\left[w_i^{obs}{}'\varepsilon_{x_i} + x_i^{obs}{}'\varepsilon_{w_i} - \varepsilon_{w_i}{}'\varepsilon_{x_i}\right]}_{\varepsilon_{C_i}} + \underbrace{\left(w_i^{obs} - \varepsilon_{w_i}\right)'\nu_{x_i}}_{\nu_{C_i}}, \forall i = 1,\ldots,n.$$

In order to obtain positive error terms, Aigner and Chu assume that measurement errors are nil or negligible, and thus:

$$x_i^{obs}{}'\alpha_x - y_i^{obs}{}'\alpha_y = \nu_{x_i}{}'\alpha_x = \nu_{F_i} \geq 0, \forall i = 1,\ldots,n$$

and

$$C_i^{obs} - w_i^{obs}{}'\beta_w - y_i^{obs}{}'\beta_y = w_i^{obs}{}'\nu_{x_i} = \nu_{C_i} \geq 0, \forall i = 1,\ldots,n.$$

They treat ν_{F_i} and ν_{C_i} as any other error term, except that they must be non-negative, $\nu_{F_i} \geq 0$ and $\nu_{C_i} \geq 0$.

Solving this problem requires a decision criterion, which can be, for example, the minimization of the squared error terms, as for OLS:

$$\min_{\alpha \geq 0} \sum_{i=1}^{n} \nu_{F_i}{}^2 = \sum_{i=1}^{n} \left(x_i^{obs}{}'\alpha_x - y_i^{obs}{}'\alpha_y\right)^2$$

and

$$\min_{\beta_w \geq 0, \beta_y \geq 0} \sum_{i=1}^{n} \nu_{C_i}{}^2 = \sum_{i=1}^{n} \left(C_i^{obs} - w_i^{obs}{}'\beta_w - y_i^{obs}{}'\beta_y\right)^2.$$

Subject to:

$$\nu_{F_i} = x_i^{obs}{}'\alpha_x - y_i^{obs}{}'\alpha_y \geq 0, \forall i = 1,\ldots,n$$

and

$$\nu_{C_i} = C_i^{obs} - w_i^{obs}{}'\beta_w - y_i^{obs}{}'\beta_y \geq 0, \forall i = 1,\ldots,n.$$

Although recourse to quadratic programming methods is necessary to solve this problem, these methods being sensitive to outliers, Aigner and Chu propose to minimize the sum of errors:[14]

$$\min_{\alpha \geq 0} \sum_{i=1}^{n} v_{F_i} = \sum_{i=1}^{n} \left(x_i^{obs} \, ' \alpha_x - y_i^{obs} \, ' \alpha_y \right)$$

and

$$\min_{\beta_w \geq 0, \beta_y \geq 0} \sum_{i=1}^{n} v_{C_i} = \sum_{i=1}^{n} \left(C_i^{obs} - w_i^{obs} \, ' \beta_w - y_i^{obs} \, ' \beta_y \right)$$

subject to:

$$v_{F_i} = x_i^{obs} \, ' \alpha_x - y_i^{obs} \, ' \alpha_y \geq 0, \forall i = 1, ..., n$$

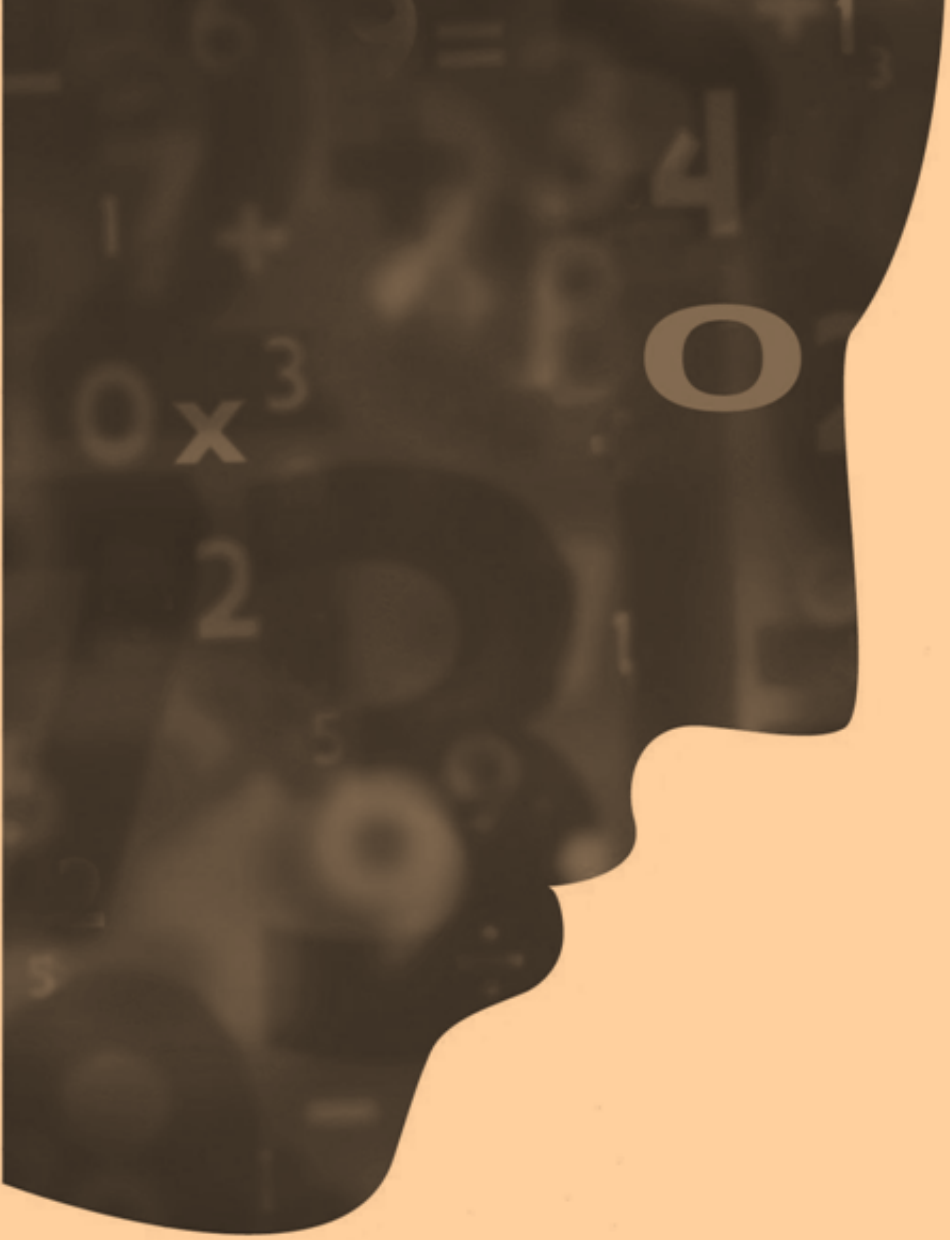
and

$$v_{C_i} = C_i^{obs} - w_i^{obs} \, ' \beta_w - y_i^{obs} \, ' \beta_y \geq 0, \forall i = 1, ..., n.$$

Simple linear programming methods are enough to solve this problem. Additional constraints can be imposed too. For example, if we impose constant returns to scale and transform in logs, then $\alpha_x \, ' \overline{1}_x = \alpha_y \, ' \overline{1}_y$ and $\beta_y \, ' \overline{1}_y = 1$, where $\overline{1}_x$ (resp. $\overline{1}_y$) is a vector of 1 of same dimension as vector x (resp. y). We can also impose homogeneity of degree 1 in price either by taking prices and costs in relative prices or by restricting the price coefficients ($1 = \beta_w \, ' \overline{1}_w$).

This approach is therefore halfway between DEA and econometric methods. The technique used are those of DEA (linear programming) and do not rely on a distribution of the error term, but are defined in a similar fashion as for econometrics (minimizing the sum of errors – squared or not – and choice of a functional form). Of course, a major shortcoming of this method is the recourse to linear production functions and the absence of confidence intervals.

[14] We can fully appreciate here why assuming no measurement error is so important. It implies that $v_{F_i} \geq 0$ and $v_{C_i} \geq 0$ hence the sum of errors is an adequate measure of inefficiency. This would not be the case if these terms were allowed to be negative and therefore compensate for positive terms.



Back of the Envelope Methods

IV. Back of the Envelope Methods

IV.1. The Accounting Method

The accounting method brings the calculation of reference unit costs to its simplest expression. Going back to the relationship between observed and minimum cost, we have for hospital h at time t:

$$C_{ht}^{obs} = C\left(\boldsymbol{w}_{ht}^{obs} - \varepsilon_{w_{ht}}, \boldsymbol{y}_{ht}^{obs} - \varepsilon_{q_{ht}}, t\right) + \underbrace{\left[w_{ht}^{obs}{}'\varepsilon_{x_{ht}} + x_{ht}^{obs}{}'\varepsilon_{w_{ht}} - \varepsilon_{w_{ht}}{}'\varepsilon_{x_{ht}}\right]}_{\varepsilon_{C_{ht}}} + \underbrace{\left(w_{ht}^{obs} - \varepsilon_{w_{ht}}\right)'\nu_{x_{ht}}}_{\nu_{C_{ht}}}.$$

Two assumptions are necessary: the hospital (or department) produces a single output (i.e., y is a scalar) and the production shows constant returns to scale. In this case, the observed cost becomes:

$$C_{ht}^{obs} = c\left(w_{ht}^{obs} - \varepsilon_{w_{ht}}, t\right) \times \left(y_{ht}^{obs} - \varepsilon_{y_{ht}}\right) + \varepsilon_{C_{ht}} + \nu_{C_{ht}}.$$

It is also possible to work straight from unit costs, in which case we only need to divide the cost by the output. Unit cost c^{obs} is:

$$c_{ht}^{obs} \equiv \frac{C_{ht}^{obs}}{y_{ht}^{obs}} = c\left(w_{ht}^{obs} - \varepsilon_{w_{ht}}, t\right) \times \frac{\left(y_{ht}^{obs} - \varepsilon_{y_{ht}}\right)}{y_{ht}^{obs}} + \frac{\varepsilon_{C_{ht}} + \nu_{C_{ht}}}{y_{ht}^{obs}}.$$

If we are willing to assume no measurement errors on variables, the unit cost then becomes:

$$c_{ht}^{obs} = c\left(w_{ht}^{obs}\right) + \frac{\nu_{C_{ht}}}{y_{ht}^{obs}}.$$

If in addition we assume that price and quality indicators are the same for all establishments, it becomes possible to compare the unit costs across the different health institutions and time.

For example, with the rule that unit cost at time t is the average of unit costs at time t-1, we have:

$$c^{référence} = \text{mean}\left\{c\left(w^{obs}, t-1\right) + \frac{v_{C_{ht-1}}}{y^{obs}_{ht-1}}\right\}$$
$$= \text{mean}\left\{c\left(w^{obs}, t-1\right)\right\} + \text{mean}\left\{\frac{v_{C_{ht-1}}}{y^{obs}_{ht-1}}\right\}.$$

Of course, other rules can be adopted: taking the median instead of the average to avoid sensitivity to outliers yields:

$$c^{référence} = \text{median}\left\{c\left(w^{obs}, t-1\right) + \frac{v_{C_{ht-1}}}{y^{obs}_{ht-1}}\right\}.$$

And finally, it is also possible to take the minimum unit cost, in which case we get a formulation akin to that of the corrected least squares. Any difference in unit cost becomes a measure of inefficiency. Overall, the accounting method is simple, but relies on very stringent assumptions that are certainly unacceptable from a theoretical standpoint. Among the least credible ones are constant returns to scale and a single output, not to mention the absence of measurement errors.

IV.2. Variations on the Accounting Method

Let's go back to the observed cost equation, i.e., the efficient cost plus the inefficiency cost (maintaining single input assumption and omitting measurement errors):

$$C^{obs}_{ht} = C\left(\boldsymbol{w}^{obs}_{ht}, y^{obs}_{ht}, t\right) + v^{obs}_{ht}.$$

Invariably, differences will appear between the unit costs of the various h establishments, and the question is to know whether these differences stem from inefficiencies or specific factors that disadvantage a particular establishment. This altogether questions the accounting method as it implies that there cannot be a reference unit cost for all establishments. If certain factors, other than efficiency, outside the establishment's control also imply higher costs, it is necessary

to take them into account to not unduly penalize this establishment. This aspect of the problem is fully factored in by the stochastic frontier approach and DEA through additional variables like price and quantity, but not by the accounting method introduced in the previous section.

Reasons for unit cost differences are obvious in the above equation: differences in prices, building and equipment endowments (and age), production scales, quality, the presence of intangible assets, weather conditions (e.g., heating costs), etc. are as many factors that will affect unit costs.

In fact, with information on these factors, it is possible to simulate the cost of an establishment within the context of another one that is used as a reference. Taking a reference institution r whose explanatory variables are $\left(w_{rt}^{obs}, y_{rt}^{obs}, v_{C_{rt}}\right)$, and cost $C_{rt}^{obs} = C\left(w_{rt}^{obs}, y_{rt}^{obs}, t\right) + v_{C_{rt}}$, we can take a first order Taylor's expansion on the efficient cost of the reference institution $C\left(w_{rt}^{obs}, y_{rt}^{obs}, t\right)$ around institution h, i.e., $\left(w_{ht}^{obs}, y_{ht}^{obs}\right)$ to get (for a given year t):[15]

$$C\left(\mathbf{w}_{rt}^{obs}, y_{rt}^{obs}, t\right) - C\left(\mathbf{w}_{ht}^{obs}, y_{ht}^{obs}, t\right) = \sum_{i=1}^{I} x_{ht}^{i} \times \left(w_{rt}^{i,obs} - w_{ht}^{i,obs}\right) + \frac{\partial C}{\partial y} \times \left(y_{rt}^{obs} - y_{ht}^{obs}\right).$$

Each of the right-hand term is an explanatory factor that partly explains the cost difference between institution h and the reference institution r.

The first term indicates that costs will be different, *ceteris paribus*, if institution h faces different prices than those of the reference institution: higher prices will increase its costs. The other term refers to the scale of production (hence the partial derivative of cost with respect to output y).

At this stage, it is almost impossible to take into account the impact of other variables. For this reason, we shall put them aside for now, remaining aware that this will introduce a bias in the calculation of expected costs. The cost difference becomes:

$$C\left(\mathbf{w}_{rt}^{obs}, y_{rt}^{obs}, t\right) - C\left(\mathbf{w}_{ht}^{obs}, y_{ht}^{obs}, t\right) = \sum_{i=1}^{I} x_{ht}^{i} \times \left(w_{rt}^{i,obs} - w_{ht}^{i,obs}\right) + \frac{\partial C}{\partial y} \times \left(y_{rt}^{obs} - y_{ht}^{obs}\right).$$

[15] We use Shephard's lemma: $\partial C / \partial w^{i} = x^{i}$.

We can also write:

$$C\left(\boldsymbol{w}_{rt}^{obs}, y_{rt}^{obs}, t\right) = C\left(\boldsymbol{w}_{ht}^{obs}, y_{ht}^{obs}, t\right) + \sum_{i=1}^{I} x_{ht}^{i} \times \left(w_{rt}^{i,obs} - w_{ht}^{i,obs}\right) + \frac{\partial C}{\partial y} \times \left(y_{rt}^{obs} - y_{ht}^{obs}\right).$$

This last expression emphasizes that the minimum cost of institution h, corrected for a certain number of terms, is equal to the reference institution's cost. The presence of partial derivatives with regards to output requires an additional assumption. With constant returns to scale in variable inputs, we can replace the marginal cost $\partial C/\partial y$ by the efficient unit cost $C\left(w_{ht}^{obs}, y_{ht}^{obs}, t\right)\Big/ y_{ht}^{obs}$. After substitution, it follows that:

$$C\left(\boldsymbol{w}_{ht}^{obs}, y_{ht}^{obs}, t\right) = C\left(\boldsymbol{w}_{rt}^{obs}, y_{rt}^{obs}, t\right) + \sum_{i=1}^{I} x_{ht}^{i,obs} \times \left(w_{ht}^{i,obs} - w_{rt}^{i,obs}\right) + \frac{C\left(\boldsymbol{w}_{ht}^{obs}, y_{ht}^{obs}, t\right)}{y_{ht}^{obs}} \times \left(y_{ht}^{obs} - y_{rt}^{obs}\right),$$

$$0 = C\left(\boldsymbol{w}_{rt}^{obs}, y_{rt}^{obs}, t\right) + \sum_{i=1}^{I} x_{ht}^{i,obs} \times \left(w_{ht}^{i,obs} - w_{rt}^{i,obs}\right) - \frac{C\left(\boldsymbol{w}_{ht}^{obs}, y_{ht}^{obs}, t\right)}{y_{ht}^{obs}} \times y_{rt}^{obs}.$$

Dividing by y_{rt} on both sides yields:

$$\frac{C\left(\boldsymbol{w}_{ht}^{obs}, y_{ht}^{obs}, t\right)}{y_{ht}} = \frac{C\left(\boldsymbol{w}_{rt}^{obs}, y_{rt}^{obs}, t\right)}{y_{rt}} + \sum_{i=1}^{I} \frac{x_{ht}^{i,obs}}{y_{rt}} \times \left(w_{ht}^{i,obs} - w_{rt}^{i,obs}\right).$$

And we can substitute the observed costs:

$$\frac{C_{ht}^{obs} - v_{ht}}{y_{ht}^{obs}} = \frac{C_{rt}^{obs} - v_{rt}}{y_{rt}^{obs}} + \sum_{i=1}^{I} \frac{x_{ht}^{i,obs}}{y_{rt}^{obs}} \times \left(w_{ht}^{i,obs} - w_{rt}^{i,obs}\right),$$

Hence,

$$\frac{C_{ht}^{obs}}{y_{ht}^{obs}} = \frac{C_{rt}^{obs}}{y_{rt}^{obs}} + \sum_{i=1}^{I} \frac{x_{ht}^{i,obs}}{y_{rt}^{obs}} \times \left(w_{ht}^{i,obs} - w_{rt}^{i,obs}\right) + \frac{v_{ht}}{y_{ht}^{obs}} - \frac{v_{rt}}{y_{rt}^{obs}}.$$

We note that choosing a reference institution (here r), implies that we are assuming that this establishment is efficient.[16] In other words, $\frac{v_{rt}}{y_{rt}^{obs}} = 0$ and thus the efficient cost of the reference institution is also its observed cost C_{rt}^{obs}. By substituting this definition in the previous equation, it follows that:

$$\frac{C_{ht}^{obs}}{y_{ht}^{obs}} = \left\{ \frac{C_{rt}^{obs}}{y_{rt}^{obs}} + \sum_{i=1}^{I} \left[\frac{x_{ht}^{i,obs}}{y_{rt}^{obs}} \times \left(w_{ht}^{i,obs} - w_{rt}^{i,obs} \right) \right] \right\} + \frac{v_{ht}}{y_{ht}^{obs}}.$$

The term in braces is the expected unit cost function for establishment h. The term $\frac{v_{ht}}{y_{ht}^{obs}}$ represents the unit cost difference caused by institution h's inefficiency and that should not be taken into account to determine this establishment's budget.

Obviously, this correction assumes that input prices and quantities can be observed. If it is not the case, we would need to restrict the work to inputs for which these variables are observed.

Finally, we should note that although the above expression correcting institution h's cost is based on an approximation of the cost of efficient institution r around the observed cost of institution h, we can proceed in the other direction, approximating the cost of a reference institution h around the observed cost of r. We then obtain a slightly different correction factor:

$$\frac{C_{ht}^{obs}}{y_{ht}^{obs}} = \left\{ \frac{C_{rt}^{obs}}{y_{rt}^{obs}} + \sum_{i=1}^{I} \left[\frac{x_{rt}^{i,obs}}{y_{ht}^{obs}} \times \left(w_{ht}^{i,obs} - w_{rt}^{i,obs} \right) \right] \right\} + \frac{v_{ht}}{y_{ht}^{obs}}.$$

IV.3. The Use of Econometric Methods to Supplement Accounting Methods

At this stage, we have a measure of the expected cost that is corrected for differences in variable input prices (at least for those with available data). The previous section also showed that non-constant returns to scale could explain unit cost differences. We will now see in this

[16] It might be more accurate to call it relative efficiency.

section that other factors can also be factored in, which implies that a part of the additional cost $\frac{v_{ht}}{y_{ht}^{obs}}$ can in fact be explained by one or many factors. Although we can easily adapt what follows to include other factors, we shall for now limit ourselves to the case of non-constant returns to scale.

From the relationship between h's efficient cost and r:

$$C\left(\boldsymbol{w}_{rt}^{obs}, y_{rt}^{obs}, t\right) - C\left(\boldsymbol{w}_{ht}^{obs}, y_{ht}^{obs}, t\right) = \sum_{i=1}^{I} x_{ht}^{i,obs} \times \left(w_{rt}^{i,obs} - w_{ht}^{i,obs}\right) + \frac{\partial C}{\partial y} \times \left(y_{rt}^{obs} - y_{ht}^{obs}\right),$$

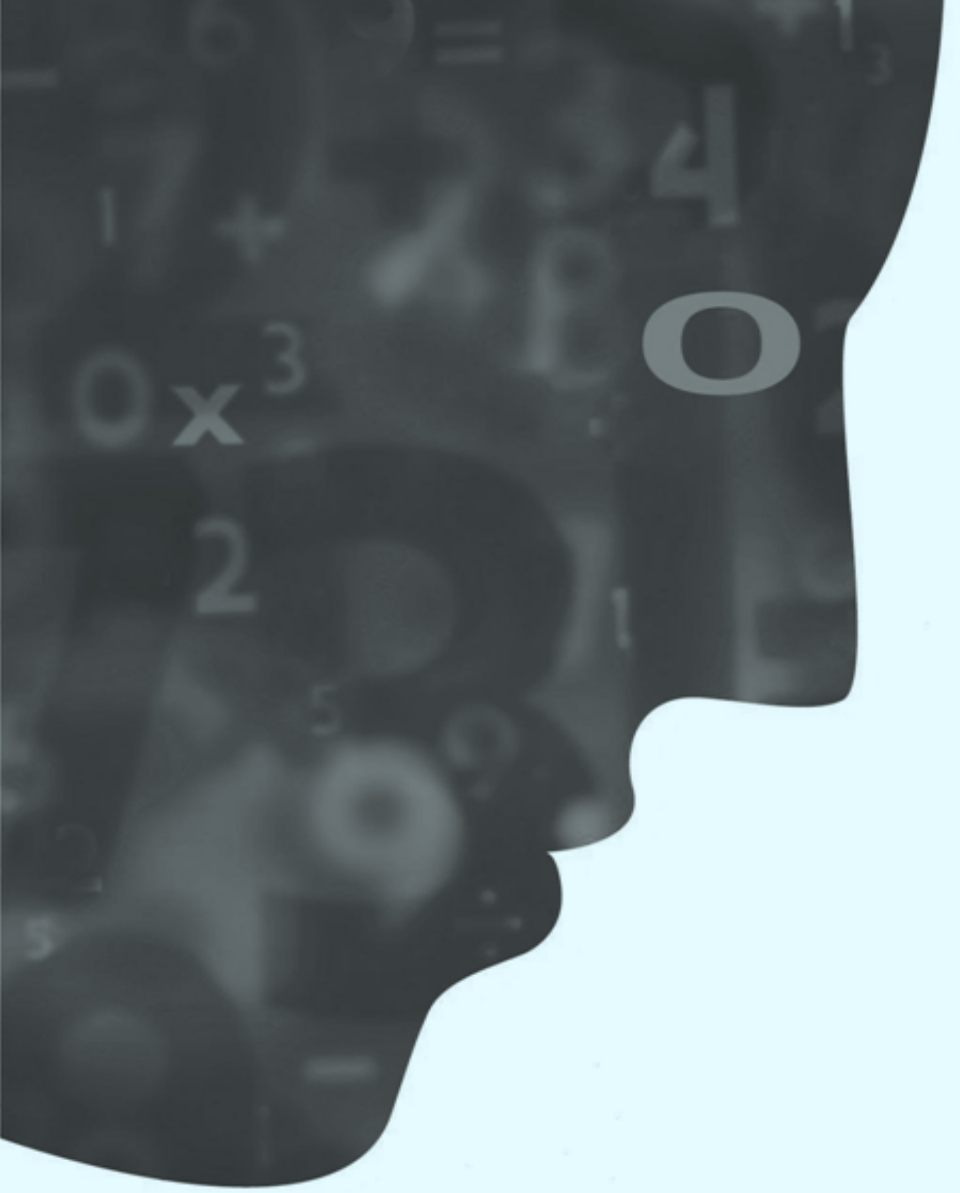
We can show that:

$$\frac{C_{ht}^{obs}}{y_{ht}^{obs}} = \left\{\frac{C_{rt}^{obs}}{y_{rt}^{obs}} + \sum_{i=1}^{I} x_{ht}^{i,obs} \times \left(w_{ht}^{i,obs} - w_{ht}^{i,obs}\right)\right\} + \frac{1}{y_{rt}^{obs}}\left\{\left(\frac{\partial C}{\partial y} - \frac{C\left(\boldsymbol{w}_{ht}^{obs}, y_{ht}^{obs}, t\right)}{y_{ht}^{obs}}\right) \times \left(y_{ht}^{obs} - y_{rt}^{obs}\right)\right\} + \frac{v_{ht}}{y_{ht}^{obs}}.$$

The first right-hand term is the expected cost with the corrected accounting method. The third term measures inefficiency. The second term is assumed to be nil or negligible in the corrected accounting method; it represents the impact of non-constant returns (the gap between the marginal and average cost). We will have the following relationship:

$$\frac{C_{ht}^{obs}}{y_{ht}^{obs}} - \left\{\frac{C_{rt}^{obs}}{y_{rt}^{obs}} + \sum_{i=1}^{I} x_{ht}^{i,obs} \times \left(w_{ht}^{i,obs} - w_{ht}^{i,obs}\right)\right\} = \frac{1}{y_{rt}^{obs}}\left\{\left(\frac{\partial C}{\partial y} - \frac{C\left(\boldsymbol{w}_{ht}^{obs}, y_{ht}^{obs}, t\right)}{y_{ht}^{obs}}\right) \times \left(y_{ht}^{obs} - y_{ht}^{obs}\right)\right\} + \frac{v_{ht}}{y_{ht}^{obs}}.$$

The right-hand term includes all measurable terms. This equation implies that the difference between the unit cost of establishment h and that of the corrected accounting method, i.e., $\frac{C_{ht}^{obs}}{y_{ht}^{obs}} - \left\{\frac{C_{rt}^{obs}}{y_{rt}^{obs}} + \sum_{i=1}^{I} x_{ht}^{i,obs} \times \left(w_{ht}^{i,obs} - w_{ht}^{i,obs}\right)\right\}$, should itself be modified to include the impact of these variables. Unfortunately, the presence of partial derivatives prevents calculation of a correction term from these observations alone. However, with a reliable dataset, it is possible to regress this additional cost on explanatory variables. This entire approach can also be made more general by including measurement errors on exogenous variables.



From Theory to Practice

V. From Theory to Practice

Good data, one would think, should bring similar results for any of the above methods. However, comparative assessments studies do not reach this conclusion. Why?

V.1. Number of Inputs : Aggregation and Quality

The large number of inputs used by any hospital and the wide spectrum of outputs are a source of significant problems. For example, in the case of a *translog* cost function with m inputs, n outputs, and a technical change time trend, there would be $(1+(m+n)+(m+n)\times(m+n+1)/2$ parameters.[17] In previous work, we have put together a dataset on Quebec hospitals; before aggregation, it featured over 100 inputs and thousands of output[18] (wide array of medical procedures, lab and other types of tests, room keeping services, laundry, etc.). For example, taking 100 inputs and 1000 outputs yields over 600,000 parameters; with the hundred or so hospitals in Quebec, empirical work would require 6,000 years of annual data. Omitting variables is obviously not an option, and reducing the number of variables is therefore necessary… but how to proceed?

This is a delicate question that is often dealt with by maximizing the informational content of the dataset, i.e., it is the data at hand that determines what procedure is used in the end. In this context, omitting variables is a recurrent issue, along with all that it implies for the credibility of results, including potential biases. Leaving aside omission of variables, how should we aggregate available data? The aggregation process essentially consists in finding aggregative functions W, X and Y such that:

$$\begin{aligned} F(x_1,...,x_m;y_1,...,y_n) &= F\left(X_1(x_1,...,x_a),...,X_\delta(x_{c+1},...,x_m);Y_1(y_1,...,y_d),...,Y_\gamma(y_{f+1},...,y_n)\right) \\ &= F\left(X_1,X_2,...,X_\delta;Y_1,Y_2,...,Y_\gamma\right) \\ &\text{and} \\ C(x_1,...,x_m;y_1,...,y_n) &= C\left(W_1(w_1,...,w_a),...,W_\delta(w_{c+1},...,w_m);Y_1(y_1,...,y_d),...,Y_\gamma(y_{f+1},...,y_n)\right) \\ &= C\left(W_1,W_2,...,W_\delta;Y_1,Y_2,...,Y_\gamma\right). \end{aligned}$$

[17] Taking the homogeneity property into account.

[18] The case of doctors is a telling example, with roughly thirty specialties and five types of employment status; adding to this nurse categories, nursing aides, administrative personnel, etc. we quickly get to over 100 inputs.

This obviously makes sense only if $\delta < m$ and $\gamma < n$.

The issue of aggregation, exact or not, is of course fundamental, but it rapidly leads to the even wider issue of data quality.[19] Indeed, certain properties of the data that are essential for aggregation are not immediately and obviously respected and relying on them to justify calculated aggregates can sometimes be a bit of a stretch. For example, assuming that the marginal substitution rate of the two output components of an output aggregate are independent of other outputs might not be economically straightforward and *a priori* verifiable; stating that the growth rate of this aggregate should be the sum of the components' growth rates weighted by their marginal cost leads nowhere if, as often, we do not know the marginal costs and there are no competitive market prices to be used as proxy. Such obstacles explain the important *ad hoc* aspect of aggregates retained by researchers in health economics. Nevertheless, even not taking these difficulties in consideration, it remains that aggregation should be minimally consistent with the characteristics of studied institutions. In fact, the important questions are really as follows: what do we lose by aggregating and how should we mitigate that loss?

If the aggregation is exact, we will have (we take the case of output aggregation, the case of other variables is identical):

$$Y_i = Y_i\left(y_1,\ldots,y_d\right).$$

If the aggregation is not exact, an additional error term is introduced:

$$\tilde{Y}_i = \tilde{Y}_i\left(y_1,\ldots,y_d\right) + \tilde{\varepsilon}_y.$$

A new error term $\tilde{\varepsilon}_y$ is added to the other measurement and optimization error terms, which further complicates the use of stochastic frontiers, as it remains impossible to conclude that the new error term is independent of other variables. In the case of a non-exact aggregation, we can use vector value functions and replace the non-observed output vector by a set of smaller dimension variables. This is best illustrated by a concrete example. If we assume, in the case of patients having undergone appendectomy, that appendicitis are treated differently whether the

[19] Blackorby *et al.* (1978) is a useful reference for aggregation and includes an extensive review.

patient is young, adult, or old and that there are only two outcomes, success or failure followed by death, there are six possible outputs:

- Number of successful appendectomies among young individuals;
- Number of failed appendectomies among young individuals;
- Number of successful appendectomies among adult individuals;
- Number of failed appendectomies among adult individuals;
- Number of successful appendectomies among old individuals;
- Number of failed appendectomies among old individuals;

As these outputs might be too many for the data at hand, we can consider the three following outputs:

- Number of appendectomies (sum of the six categories above);
- Percentage of successful procedures;
- Average age of patients.

This is only one of many possible aggregation examples. This type of aggregation is often used in health systems research as well as in other fields, such as transportation and is called "hedonic". Depending on available degrees of freedom, we might have to aggregate further by, for example, dropping average age. Of course, we aggregate only when forced to by lack of data, but this will always be the case.

At this juncture, it becomes necessary to turn our attention to notational issues. In the previous example, the number of outputs is reduced from six to three. In the terminology used by researchers in health, these variables are called differently. In fact, some are called output, while others are quality variables. In our example, the number of appendectomies would be an output, whereas the percentage of failed procedures and average age are quality variables. This is purely a rhetorical matter and has no bearing on the issue at hand. Indeed, the three variables are qualitative in nature and have meaning only when taken together to quantify a six-output aggregate.

Bending to common practice by regrouping quantitative measures of service (such as the number of appendectomies) in the aggregate vector of output (noted Y) and other variables in a quality vector (noted q_o), the cost function becomes:

$$C\left(w,Y,q_o\right)\triangleq\min_x\left\{w'x:f\left(x,Y,q_o\right)\leq 0\right\}.$$

The same reasoning can be used on the input side. For example, in the presence of different nurse categories (noted $\left(x_1,...,x_a\right)$ where $i=1$, ..., and *a* years of experience), we have many options:

We can simply add the number of hours worked assuming that more experienced and less experiences nurses can be substituted, i.e. $X\left(x_1,...,x_a\right)=\sum_{i=1}^{a}x_i$ (which is current practice), or;

We can create for each hospital a seniority index to take into account experience-related productivity gains: $X\left(x_1,...,x_a\right)=\left(\sum_{i=1}^{a}x_i,\sum_{i=1}^{a}i\frac{x_i}{\sum_{j=1}^{a}x_j}\right)$,[20] in which case we reduce the number of categories from a to two.

Proceeding as before, we can define quality and quantity input vectors (respectively noted q_i and X), along with the corresponding input price vector (noted W) and write the cost function using input aggregates:

$$C\left(W,q_i,Y,q_o\right)\triangleq\min_x\left\{W'X:f\left(X,q_i,Y,q_o\right)\leq 0\right\}.$$

The need to aggregate grows when using data from entire institutions, because of the large number of inputs and outputs: the more inputs and outputs, the greater the need to aggregate. In this context, the choice of quality indices (such as success/ mortality rates, seniority) becomes more and more important. Unfortunately, quality indices are not plenty and it is often difficult to find an adequate overall indicator. This is why so many studies delve into this issue without finding answers that are satisfactory to all. For example, many studies attacked this problem by using Diagnostic Related Groups (DRG), which are in fact aggregates based on a set of priors regarding the relative intensity of resources used to treat patients. Although DRGs allow

[20] This seniority index, as well as any other input quality index, is included in the vector q_i.

reduction of the number of variables, there is no indication that they do not also introduce significant aggregation bias.

V.2. Endogenous Production

Up to now, we have assumed that output was exogenous and thus that the demand for hospital services was rather rigid, even if patients do not have to pay. However, patients can be attracted by a hospital's good reputation, the short waiting time, health care quality (at least as perceived by the patients), the presence of other institutions nearby, the constraints imposed by insurers (obligation to use a specific hospital or not), etc.

One way to approach this issue is to introduce the demand as an additional condition through the equality condition between the marginal revenue and marginal cost. Marginal revenue of institution i is given by the demand for its own services. Obviously, the formulation of this demand also depends on market conditions, hence the need for competition proxies (Herfindhal indices, for example).

In this context, we should jointly estimate the cost function (and the factor demand shares) and the output optimality condition:

$$
\begin{aligned}
C^{obs} &= C(w, y) + \text{ error term for the cost function} \\
S_i^{obs} &= S_i(w, y) + \text{ error term for the } i\text{th share equations} \\
Rm_i(y, \phi_i) &= \frac{\partial C}{\partial y_i} + \text{ error term for the optimal output condition.}
\end{aligned}
$$

Where ϕ_i represents the characteristics of the demand and market for output i. Alternatively, we can impose the constraint that the firm's supply is equal to the demand:

$$
\begin{aligned}
C^{obs} &= C(w, y) + \sigma_C \\
S_i^{obs} &= S_i(w, y) + \sigma_i \\
y_i &= D_i(p, \phi_i) + \sigma_y .
\end{aligned}
$$

Where p is the vector of prices consumers base their decisions on, and ϕ_i the vector of other information and characteristics of the demand for health service i. We can substitute this condition in the cost function:

$$C^{obs} = C\left(w, D\left(p, \phi\right) + \sigma_y\right) + \sigma_C$$
$$S_i^{obs} = S_i\left(w, D\left(p, \phi\right) + \sigma_y\right) + \sigma_i.$$

This formulation is very close to the one introduced earlier, and the same remarks apply. Obviously, additional considerations on measurement errors and demand characteristics can be factored in as well.

Conclusion

Conclusion

Our goal in this paper was to provide an overview of methods used to measure the efficiency of health institutions. In an accompanying document (Ouellette and Petit, 2010), we also present a review of 600 articles published since the 1950s, which apply methods discussed in this paper, to the health sector. Interested readers may refer to this paper for a comprehensive list of innovative, though at times inconsistent, examples and contributions. Indeed the current paper is broadly relevant to various industrial sectors, whereas the accompanying document provides examples referring specifically to the health care sector.

The approach used in this paper runs somewhat counter to that used by most researchers: instead of formulating an economic model first and then building in the error structure, we started by defining error terms and subsequently included variables relevant for the economic model. This seemingly banal inversion allowed the formulation of intrinsically consistent models from the standpoint of economic theory and methods used to measure efficiency. In light of our review, it appears that the most commonly used methods do not respect the main theoretical tenets, and that models are either incomplete or inconsistent. Our approach also helps highlight underlying assumptions in currently used models.

Our efforts, however, went beyond commenting on past contributions. We explained how models should be specified and introduced intrinsically consistent models for the usual empirical approaches (statistics, operational research, or accounting methods). These models all share the same usability and do not require any further innovation in terms of econometric methods or mathematical programming.

References

Blackorby, C., D. Primont and R. Russell (1978), "*Duality, Separability, and Functional Structure: Theory and Economic Applications*", North-Holland, Amsterdam.

Caves, D.W., L.R. Christensen and W.E. Diewert (1982), "*The Economic Theory of Index Numbers and the Measurement of Input, Output, and Productivity*", Econometrica, Vol. 50, No. 6, p. 1393-1414.

Diewert, W.E., and C. Parkan (1983), "*Linear Programming Tests of Regularity Conditions for Production Functions*", p. 131-158, in W. Eichhorn, R. Henn, K. Neumann and R.W. Shephard (Eds.) Quantitative Studies on Production and Prices, Wien: Physica Verlag.

Färe, R., S. Grosskopf, B. Lindgren and P. Roos (1992), "*Productivity Changes in Swedish Pharamacies 1980–1989: A Non-parametric Malmquist Approach*" Journal of Productivity Analysis, Vol. 3, No. 1-2, p. 85-101.

Färe, R., and J. Logan. (1983), "*The Rate-of-Return Regulated Firm: Cost and Production Duality*" The Bell Journal of Economics, Vol. 14, No. 2, p. 405–414.

Gagné, R., and P. Ouellette (1998), "*On the Choice of Functional Forms: Summary of a Monte Carlo Experiment*", Journal of Business and Economic Statistics, Vol. 16, No. 1, p. 118-124.

Gagné, R., and P. Ouellette (2002), "*The Effect of Technological Change and Technical Inefficiencies on the Performance of Functional Forms*", Journal of Productivity Analysis, Vol. 17, No. 3, p. 233-247.

Ma, C.T.A. (1994), "*Health Care Payment Systems: Cost and Quality Incentives*", Journal of Economics and Management Strategy, Vol. 3, p. 93-112.

McElroy, M.B. (1987), "*Additive General Error Models for Production, Cost, and Derived Demand or Factor Share Systems*", Journal of Political Economy, Vol. 95, No. 4, p. 737-755.

Ouellette, P., and P. Petit (2010), « Efficience budgétaire des institutions de santé : une revue de littérature », Centre sur la productivité et la prospérité, HEC Montréal.

Ouellette, P., P. Petit, L.-P. Tessier-Parent and S. Vigeant (2010), "*Introducing Regulation in the Measurement of Efficiency, with an Application to the Canadian Air Carriers Industry*", European Journal of Operational Research, Vol. 200, p. 216-226.

Ouellette, P., and S. Vigeant (2001a), "C*ost and Production Duality: The Case of the Regulated Firm*", Journal of Productivity Analysis, Vol. 16, No. 3, p. 203-224.

Ouellette, P., and S. Vigeant (2001b), "*On the Existence of a Regulated Production Function*", Journal of Economics, Vol. 73, No. 2, p. 193–200.

Simar, L., and P.W. Wilson (1998), "*Sensitivity Analysis of Efficiency Scores: How to Bootstrap in Nonparametric Frontier Models*", Management Science, Vol. 44, No. 1, p. 49-61.